frontiers in Genetics

METHODS
published: 04 November 2020
doi: 10.3389/fgene.2020.569227

RKDOSCNV: A Local Kernel Density-Based Approach to the Detection of Copy Number Variations by Using Next-Generation Sequencing Data

Guojun Liu, Junying Zhang, Xiguo Yuan* and Chao Wei*

School of Computer Science and Technology, Xidian University, Xi'an, China

Copy number variations (CNVs) are significant causes of many human cancers and genetic diseases. The detection of CNVs has become a common method by which to analyze human diseases using next-generation sequencing (NGS) data. However, effective detection of insignificant CNVs is still a challenging task. In this study, we propose a new detection method, RKDOSCNV, to meet the need. RKDOSCNV uses kernel density estimation method to evaluate the local kernel density distribution of each read depth segment (RDS) based on an expanded nearest neighbor (k-nearest neighbors, reverse nearest neighbors, and shared nearest neighbors of each RDS) data set, and assigns a relative kernel density outlier score (RKDOS) for each RDS. According to the RKDOS profile, RKDOSCNV predicts the candidate CNVs by choosing a reasonable threshold, which it uses split read approach to correct the boundaries of candidate CNVs. The performance of RKDOSCNV is assessed by comparing it with several current popular methods via experiments with simulated and real data at different tumor purity levels. The experimental results verify that the performance of RKDOSCNV is superior to that of several other methods. In summary, RKDOSCNV is a simple and effective method for the detection of CNVs from whole genome sequencing (WGS) data, especially for samples with low tumor purity.

Keywords: copy number variation, next-generation sequencing, kernel density estimation, split read, biological meanings

OPEN ACCESS

Edited by:
Alfredo Pulvirenti,
University of Catania, Italy

Reviewed by:
Sheida Nabavi,
University of Connecticut,
United States
Weichen Zhou,
University of Michigan, United States

****Correspondence:***
Junying Zhang
jyzhang_xidian@qq.com
Xiguo Yuan
xiguoyuan@mail.xidian.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *03 June 2020*
Accepted: *04 September 2020*
Published: *04 November 2020*

Citation:
Liu G, Zhang J, Yuan X and Wei C (2020) RKDOSCNV: A Local Kernel Density-Based Approach to the Detection of Copy Number Variations by Using Next-Generation Sequencing Data. Front. Genet. 11:569227. doi: 10.3389/fgene.2020.569227

INTRODUCTION

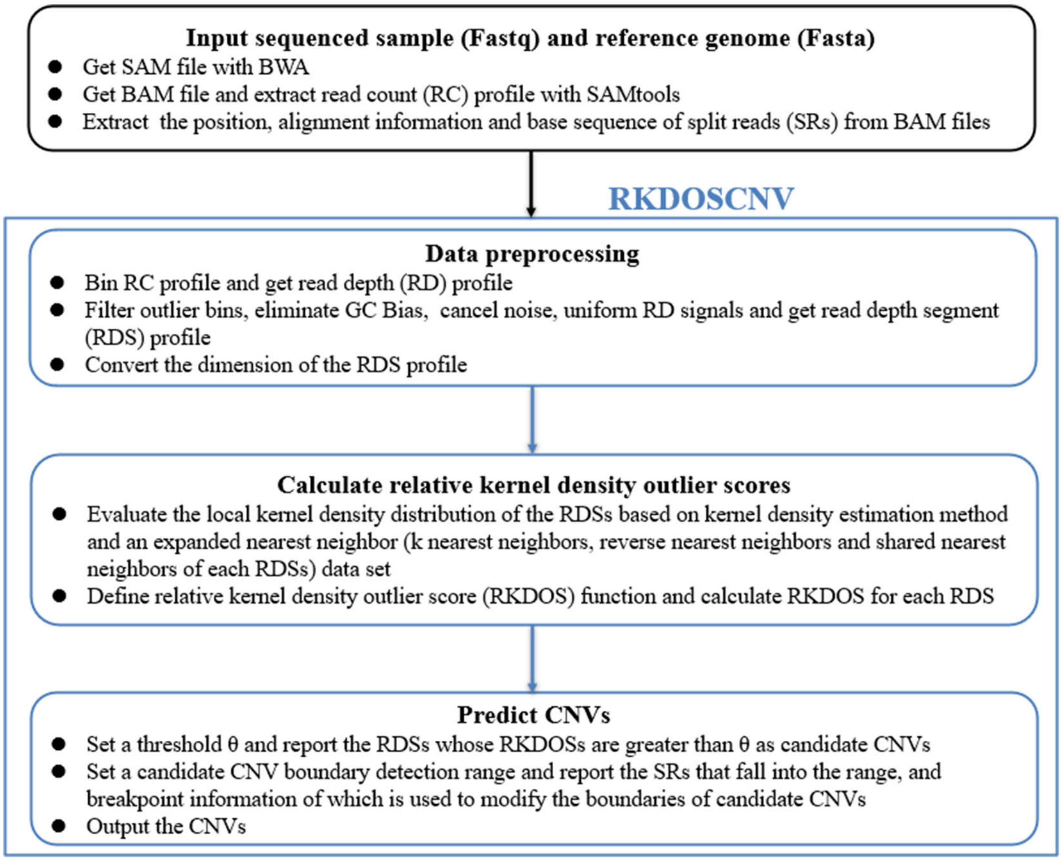
With the rapid development of next-generation sequencing (NGS) technology, many sequencing data sets that are used to detect and characterize human genome variation have been produced (Medvedev et al., 2009). Copy number variation (CNV) is one of the important forms of genome structural variation (Freeman et al., 2006). It has been reported that many human cancers and diseases are caused directly or indirectly by CNVs (Zhao et al., 2013). It is therefore necessary for humans to accurately detect CNVs using NGS data to effectively discover disease-causing genes and develop targeted drugs (Yuan et al., 2018). The workflow of the general CNV detection method includes the following steps: (1) comparing reads to a reference genome (Metzker, 2010) and generating a SAM file with BWA (Li and Durbin, 2010);

(2) obtaining a read count (RC) profile using SAMtools (Li et al., 2009); (3) binning the RC profile (Chiang et al., 2009) and obtaining the read depth (RD) profile; (4) preprocessing the RD profile [eliminating GC bias, removing mappability bias (Dohm et al., 2008), and denoising the data (Cai et al., 2018)]; (5) modeling and forecasting CNVs. In recent years, many CNV detection methods have been developed based on this process. According to the number of entered samples, the CNV detection methods can be divided into three categories, namely those that use multiple samples, matched case-control samples, and a single sample, respectively.

Many detection methods have been developed based on multiple samples as inputs, such as cn.MOPS (Povysil et al., 2017), RDXplorer (Yoon et al., 2009), CoNIFER (Krumm et al., 2012), XHMM (Fromer et al., 2012), and CODEX (Jiang et al., 2015). The cn.MOPS method assumes that the RD at each location is subject to a hybrid Poisson model that is used to predict CNVs. RDXplorer detects CNVs using an event-wise testing method, namely significance testing based on preprocessed RDs of coverage. CoNIFER uses singular value decomposition to detect CNVs and discover genotype copy number polymorphic loci. XHMM is a statistical tool that uses principal component analysis to preprocess RD profiles, and establishes a hidden Markov model to announce CNVs. CODEX builds a Poisson latent factor model to normalize RDs, and uses a Poisson likelihood-based recursive segmentation procedure to declare CNVs. Multiple sample-based methods are suitable for the detection of the driver genes of cancer in the same diseased group, which is conducive to the development of some targeted drugs for clinical treatment. Via the preceding analysis, it is clear that most CNV detection methods based on multiple samples use a distribution model to fit the RD signals. However, in practical applications, RD signals do not obey the assumed distribution due to data noise, sequencing errors, and sample contamination, which leads to inaccurate test results. In daily life, doctors often encounter patients with different diseases who require different treatment plans. Therefore, the development of single sample-based methods is imperative.

Some popular methods based on matched case-control samples have been developed in recent years, and include XCAVATOR (Magi et al., 2017), SeqCNV (Chen et al., 2017), BIC-seq (Xi et al., 2010), CNAnorm (Gusnanto et al., 2012), and CNAseg (Ivakhno et al., 2010). XCAVATOR uses a two-step procedure in which RC biases are removed to identify the absolute copy number and a shifting-level model method is used to predict CNVs. SeqCNV comprises a maximum penalized likelihood estimation method to calculate the copy number ratio and detect CNVs. BIC-seq adopts the minimization of the Bayesian information criterion method to identify CNVs based on RD information. CNAnorm calibrates contamination with normal cells and assesses the ploidy to calculate the copy numbers of detected areas. CNAseg adopts flowcell-to-flowcell variability in case and control samples to reduce the false positive rate and identify CNVs. The matched case-control sample-based methods are suitable for detection in individual patients, and can be used to identify germline and somatic CNVs. These methods can detect disease-causing genes associated with cancer and identify the difference between the normal and abnormal tissues of a single patient, and can effectively identify somatic CNVs using targeted NGS data, which is very important especially in clinical examination and cancer research. However, these methods also have some limitations and a control sample of a patient is required, the collection and production of which are relatively expensive. Via the preceding analysis, it is evident that the paired sample-based methods are suitable for the detection of disease in niche people or for conducting experimental research that requires specific clinical needs.

Due to the needs of practical applications, single sample-based CNV detection methods emerged. In contrast to the two other types of methods, these methods only require a sample as the input, which reduces the cost of patient testing. For example, as compared with the paired sample-based methods, the cost of patient testing is reduced by about half because the provision of a control sample is not required. In recent years, many single sample-based methods have been developed, most of which use depth of coverage (DOC) information to establish a model and forecast CNVs; the basic principle is that the RC of each location of a reference genome is proportional to the copy number of each location (Yoon et al., 2009). In theory, DOC-based methods can detect CNVs of any size; thus, the vast majority of existing methods have been developed based on DOC technology, and include ReadDepth (Miller et al., 2011), CNVnator (Abyzov et al., 2011), GROM-RD (Smith et al., 2015), iCopyDAV (Dharanipragada et al., 2018), FREEC (Boeva et al., 2011), and CNV_IFTV (Yuan et al., 2019). ReadDepth employs preprocessed RD information to fit a negative binomial distribution and circular binary segmentation to forecast CNVs; it is suitable for the detection of high-purity tumor samples, though its accuracy is relatively low. CNVnator uses the mean-shift method with GC content correction and multiple-bandwidth partitioning to discover CNVs; it is not suitable for testing low-purity samples or the investigation of short CNVs, and, while it can achieve a high recall rate, it has low precision. GROM-RD corrects repeat deviations and adopts a two-pipeline masking approach to discover CNVs in duplicate and complex areas; however, its performance is limited to the detection of low-purity tumor samples, and the false positive rate of test results is relatively high. iCopyDAV is an integrated software platform composed of the detection, annotation, and visualization of CNVs, and is suitable for testing samples with high purity and medium coverage. FREEC preprocesses the RD with GC content and uses GC content profiles for segmentation, which is used to allocate copy numbers for each area. Its performance is more balanced, and it achieves high recall and precision. CNV_IFTV calculates the anomaly score for each area with the isolation forest method and employs a total variation model to smooth the anomaly score profile, based on which a gamma distribution is established to predict CNVs. It achieves a better tradeoff between recall and precision as compared to the other previously mentioned methods, but it has a higher time complexity. Via the analysis of these methods, it is evident that the development of some new methods with complete functions based on a single sample is necessary. The existing methods mainly have the following limitations. (1) The general



Input sequenced sample (Fastq) and reference genome (Fasta)

- Get SAM file with BWA
- Get BAM file and extract read count (RC) profile with SAMtools
- Extract the position, alignment information and base sequence of split reads (SRs) from BAM files

RKDOSCNV

Data preprocessing

- Bin RC profile and get read depth (RD) profile
- Filter outlier bins, eliminate GC Bias, cancel noise, uniform RD signals and get read depth segment (RDS) profile
- Convert the dimension of the RDS profile

Calculate relative kernel density outlier scores

- Evaluate the local kernel density distribution of the RDSs based on kernel density estimation method and an expanded nearest neighbor (k nearest neighbors, reverse nearest neighbors and shared nearest neighbors of each RDSs) data set
- Define relative kernel density outlier score (RKDOS) function and calculate RKDOS for each RDS

Predict CNVs

- Set a threshold θ and report the RDSs whose RKDOSs are greater than θ as candidate CNVs
- Set a candidate CNV boundary detection range and report the SRs that fall into the range, and breakpoint information of which is used to modify the boundaries of candidate CNVs
- Output the CNVs

FIGURE 1 | Overview of the workflow of RKDOSCNV. It is mainly composed of four parts, namely input file preparation, data preprocessing, relative kernel density outlier score calculation, and CNV prediction.

methods assume that the RD information follows a certain distribution, such as a negative binomial distribution or a mixed Poisson distribution. In actual experiments, it is found that the distribution of RD signals deviates from the assumption, which is caused by sequencing errors, GC-content bias, mappability bias, and experimental noise. (2) Most methods only use RD information to build models and predict CNVs. In theory, these methods can detect CNVs of any size and type, but cannot accurately detect the boundaries of variation regions. (3) It is very difficult for most methods to detect the hemizygous loss regions, the copy numbers of which are one. The regions can easily be regarded as normal areas with copy numbers of two. Therefore, it is necessary to adopt reasonable strategies and reduce the impact on the test results.

Based on the previously discussed considerations, a novel method, called RKDOSCNV (local kernel density estimation-based approach for CNV detection), is developed in this study to predict CNVs from single tumor samples via the use of whole genome sequencing (WGS) data. A kernel density estimation (KDE) method is adopted to evaluate the distribution of the local kernel density (LKD) for each read depth segment (RDS) based on the extended nearest-neighbor data set, which is composed of the k-nearest neighbors (KNNs), reverse nearest neighbors (RNNs), and shared nearest neighbors (SNNs) of each RDS (Tang and He, 2017). After the calculation of the LKD of RDSs, the relative kernel density outlier score (RKDOS) is evaluated for each RDS, which can indirectly reflect the degree of deviation of each RDS as compared to its three types of nearest neighbors. By choosing a suitable threshold θ, the RKDOS of each RDS is compared with θ. For example, if the RKDOS of an RDS is greater than θ, it is considered as a candidate CNV. Based on the prediction results, the boundaries of candidate CNVs are more precisely refined using split read (SR) approach. The performance of the proposed RKDOSCNV is estimated based on simulated data sets and compared with the performances of several popular methods. To further verify the validity of RKDOSCNV, it is used to detect real tumor samples, and it is found that some CNVs are associated with cancers and diseases, thus proving the effectiveness of the method.

METHOD AND MATERIALS

Overview of RKDOSCNV

RKDOSCNV is developed based on the DOC and SR methods, and is applied to detect single tumor samples without the provision of a matched control sample. **Figure 1** describes in detail the overall workflow of RKDOSCNV, which is composed of the following four steps. First, the sequenced sample (Fastq) and reference genome (Fasta) are offered as inputs. Many of the reads that come from the Fastq file are compared to the reference genome (e.g., HG19), which can generate a SAM file using BWA (Li and Durbin, 2010). The SAM file can be converted to a BAM file, and the RC profile is extracted from the SAM file with SAMtools (Li et al., 2009). The positions, alignment information, and base sequences of SRs are further collected from BAM files. The second step mainly includes defining the bins, filtering anomalous bins, calibrating the GC bias, noise reduction, balancing RD signals, and converting the dimensions of RD data. The third step, RKDOSCNV calculates the RKDOS for each RDS based on the extended nearest-neighbor data set (Tang and He, 2017). Finally, RKDOSCNV reports candidate CNV regions whose RKDOSs are $>$ θ, which is a set threshold, and applies the SR approach to further refine the boundaries of the candidate CNV regions. The RKDOSCNV software is developed in the R and Python languages. It can be downloaded from https://github.com/gj-123/RKDOSCNV/releases, and is easy to install and use after reading the user manual.

Data Preprocessing

The sequenced sample is compared to the reference genome by BWA, which obtains the SAM file from which the RC profile is extracted with SAMtools. The reference genome is composed of many "A," "T," "G," "C," and "N," of which "N" represents the positions of the reference genome that are missing in the sequencing process. When reads compares to positions of "N," which will cause the RCs to be equal to zero at these positions (Yuan et al., 2018). To obtain a reasonable RC profile, the following strategies are adopted to solve this problem. A bin is defined, which is a continuous and non-overlapping sliding window. In this study, the bin size was set to 2000 bp. If a bin contains positions of "N," it will be filtered as an abnormal bin. An RD is defined as a numerical value that can be determined by calculating the RC mean of a bin. Based on this processing, the GC bias is caused by the PCR amplification that is calibrated by Equation (1) (Yoon et al., 2009):

$$RD_i^{'} = \frac{\overline{RD} \cdot RD_i}{RD_{gc}}, \tag{1}$$

where $RD_i^{'}$ represents the RD of the i-th bin after correction, $\overline{RD}$ represents the mean of all RDs, RD_i represents the RD of the i-th bin, and RD_{gc} denotes the mean of RDs that have the same GC content.

The noise of RDs severely affects the accuracy of CNV detection; the noisy data is directly used to detect CNVs, which results in inaccurate detection results. Therefore, the elimination of noise is a critical step in the CNV detection process. In this study, a total variation method is adopted, which can reduce noise, segment, and smooth one-dimensional discrete signals based on the regularized total variation (Condat, 2013) and least-squares approaches (Duan et al., 2013). With the RDS profile, the RDS signals are asymmetric because the copy number gains and losses are asymmetric as compared with the normal copy number. If the model is based on asymmetric RD signals, some insignificant CNVs will be easily overlooked. Here, Equation (2) is used to adjust the symmetry of RD signals (Yuan et al., 2018):

$$RDS_i^{'} = \left| \frac{RDS_{\min}}{RDS_{\max}} \right| \cdot (RDS_i - RDS_M), \tag{2}$$

where $RDS_i^{'}$ denotes the RD of the i-th segment after equalization, $RDS_{\min}$ denotes the minimum value of the RDSs, $RDS_{\max}$ denotes the maximum value of the RDSs, RDS_i denotes the RD of the i-th segment, and RDS_M denotes the mode of the RDSs. Based on the balanced RDS profile, the one-dimensional RDS profile is converted into a two-dimensional profile RDS_s, which are described in detail by Equations (3–5), respectively.

$$RDS_x = \frac{RDS_i}{\overline{RDS}} \tag{3}$$

$$RDS_y = \begin{cases} \frac{\sum_{j=i+1}^{i+m} |RDS_x[i]-RDS_x[j]|}{m} & i = 1 \\ \frac{\sum_{j=1}^{i-1} |RDS_x[i]-RDS_x[j]| + \sum_{j=i+1}^{i+m} |RDS_x[i]-RDS_x[j]|}{i+m-1} & 1 < i \le m \\ \frac{\sum_{j=i-m}^{i-1} |RDS_x[i]-RDS_x[j]| + \sum_{j=i+1}^{i+m} |RDS_x[i]-RDS_x[j]|}{2m} & m < i \le |RDS_x| - m \\ \frac{\sum_{j=i-m}^{i-1} |RDS_x[i]-RDS_x[j]| + \sum_{j=i+1}^{|RDS_x|-1} |RDS_x[i]-RDS_x[j]|}{|RDS_x|-i-1+m} & |RDS_x| - m < i \le |RDS_x| - 1 \\ \frac{\sum_{j=i-m}^{i-1} |RDS_x[i]-RDS_x[j]|}{m} & i = |RDS_x| \end{cases} \tag{4}$$

$$RDS_s = \{RDS_x, RDS_y\} \tag{5}$$

In Equation (3), RDS_i represents the RD of the i-th segment, $\overline{RDS}$ represents the mean of all the RDSs, and RDS_x represents the ratio between RDS_i and $\overline{RDS}$, which indirectly reflects the distribution of the copy number at each location. In Equation (4), $|RDS_x|$ represents the number of elements in the RDS_x. RDS_y represents the difference between an RDS and the surrounding RDSs in a certain depth of exploration (m), which is helpful for the detection of insignificant CNVs in the local range. Here, the default value of m is set to 10. In Equation (5), RDS_s represents a two-dimensional data set that is composed of RDS_x and RDS_y, and each element of it is treated as an object (O). The main purpose of this step is conducive to the capture of local insignificant CNVs, and provides an effective data set for the calculation of the RKDOSs presented in the subsequent section.

Calculation of Relative Kernel Density Outlier Scores

Based on the RDS_s profile, RKDOSCNV uses the KDE approach to evaluate an LKD distribution, and assigns the RKDOS that can reflect the degree of isolation of an object as compared with its three types of nearest neighbors for each RDS (Tang and He, 2017). This is different from the traditional RD-based

CNV detection methods, which generally build statistical models according to the approximate distribution of RDSs and choose an appropriate significance level to predict CNVs. Here, the LKD distribution function is defined using Equations (6–9).

$$f(O) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^{d}}K(\frac{O-O_i}{\sigma}) \tag{6}$$

$$K(\frac{O-O_i}{\sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}}e^{(-\frac{||O-O_i||^2}{2\sigma^2})} \tag{7}$$

In Equation (6), O represents any element in the RDS_s, O_i represents the i-th element in the RDS_s, σ represents the bandwidth of the kernel function, d represents the dimension of RDS_s, and $K(\frac{O-O_i}{\sigma})$ represents a multivariate Gaussian kernel function and is defined in Equation (7). $||O - O_i||^2$ represents the Euclidean distance between and O_i.To accurately calculate the LKD of O, its three types of neighbor relations (KNNs, RNNs, and SNNs) are calculated, and are used as kernels in the Gaussian kernel function and represented with three sets (RDS_{s-knn}, RDS_{s-rnn}, and RDS_{s-snn}), which are described in detail by Equations (8–11).

$$RDS_{s-knn}(O) = (RDS_{s-knn}[1], RDS_{s-knn}[2], ..., RDS_{s-knn}[k]) \tag{8}$$

$$RDS_{s-rnn}(O) = (RDS_{s-rnn}[1], RDS_{s-rnn}[2], ..., RDS_{s-rnn}[i]) \; or \varnothing \tag{9}$$

$$RDS_{s-snn}(O) = (RDS_{s-snn}[1], RDS_{s-snn}[2], ..., RDS_{s-snn}[j]) or \varnothing \tag{10}$$

$$RDS_{s-u}(O) = RDS_{s-knn}(O) \cup RDS_{s-rnn}(O) \cup RDS_{s-snn}(O)$$
$$1 \le i \le k, 1 \le j \le k, i, j \in N \tag{11}$$

In Equation (8), O represents any object in the RDS_s, $RDS_{s-knn}(O)$ represents the set of the KNNs of O. Here, the default value of k is set to 60. In Equation (9), $RDS_{s-rnn}(O)$ is a set that is composed of the RNNs of O, which are defined as the objects, the KNNs of which include object O. In practical applications, the RNNs of O may or may not exist. In Equation (10), $RDS_{s-snn}(O)$ represents the SNNs set of O, which is defined as the objects that have the same nearest neighbors as O. Similarly, the SNNs of O may or may not exist. In Equation (11), $RDS_{s-u}(O)$ represents the union of three types of nearest neighbors, based on which the proposed method can detect not only an isolated anomalous individual, but also a cluster of local insignificant anomalous individuals. The Equation (12) is used to calculate the LKD of O:

$$f(O) = \frac{1}{|RDS_{s-u}(O)|}\sum_{O_i \in RDS_{s-u}(O)}\frac{1}{\sigma^{d}}K(\frac{O_i - O}{\sigma}), \tag{12}$$

where $|RDS_{s-u}(O)|$ represents the number of elements in the $RDS_{s-u}(O)$. Based on Equation (12), the calculation of the RKDOS for O is described by Equation (13):

$$RKDOS(O) = \frac{\sum_{O_i \in RDS_{s-u}(O)} f(O_i)}{|RDS_{s-u}(O)|f(O)}, \tag{13}$$

where $RKDOS(O)$ is defined as the ratio between the average of the LKD of the three types of nearest neighbors of O and the LKD of O.

Declaring CNVs

Based on the RKDOS of each object, the degree of anomalies is progressively analyzed, and a threshold θ is chosen as the cutoff for those abnormal objects. The threshold θ is a constant that is determined by users according to their own application scenarios. In this work, the θ is set to 1.1. The basic judgment principle is as follows. (1) If the RKDOS of an object is $>$ θ, it is considered as a candidate CNV area. (2) If the RKDOS of an object is $\le$θ, it is considered as a candidate normal area. Structural variants (SVs) is an important manifestation of human chromosome variation, which includes the tandem amplification, interspersed amplification, deletion, insertion, and rearrangements (translocation and inversion) of DNA fragments (Stankiewicz and Lupski, 2010; Guan and Sung, 2016; Yuan et al., 2020). The SR-based SV and CNV detection methods can accurately detect the boundaries of the mutation areas, and reduce false positives and false negatives in the test results (Ye et al., 2009; Abyzov and Gerstein, 2011; Rausch et al., 2012; Layer et al., 2014). In this study, the SR approach is used to determine the locations of the breakpoints of candidate CNVs. Three features (the POS, CIGAR, and sequence) of SRs are extracted from BAM files (Li et al., 2009). The POS indicates the position of the leftmost first base of an SR sequence aligned to the reference genome. The CIGAR indicates the state of an SR aligned to the reference genome. It is a string, which is composed of numbers, M and S. Here, the type of CIGAR that only includes M and S is extracted, and is described by Equation (14):

$$CIGAR = \begin{cases} mMnS & m, n \epsilon N, m + n = R_l, \\ nSmM & ditto \end{cases} \tag{14}$$

where M represents an exact match, S represents no match, m indicates that the m bases of an SR completely match the reference genome, n indicates that the n bases of an SR cannot be completely matched to the reference genome, and R_l represents the length of an SR, which is generally set to 100 bp. Here, a CIGAR with a value of n >10 is chosen. If the mismatch length of an SR is too small, it may be caused by sequencing errors. The sequence represents the base sequence of an SR. Most SRs are useless for boundary detection; therefore, conditions must be set to filter out effective SRs. Here, the search step size is set to the length of n bins (L). The starting point and endpoint indexes of a candidate CNV region that represent the leftmost position and rightmost position of a candidate CNV are respectively used as search centers according to which forward and backward searches are performed; This step forms two search ranges where those SRs are recorded. **Figure 2** depicts an example of exploration process of SRs. The n SRs (from SR_1 to SR_n) and

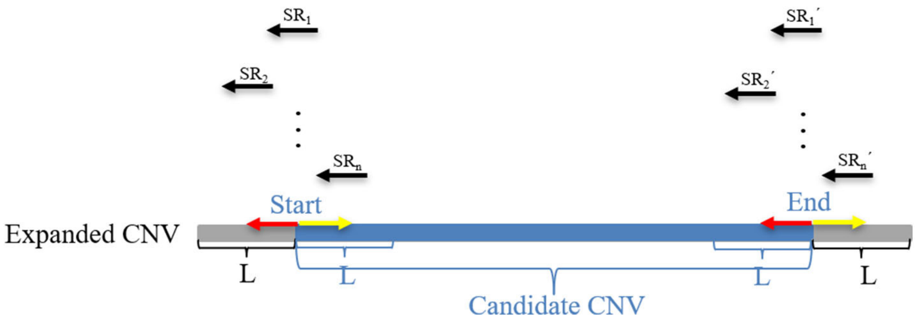


FIGURE 2 | An example of the search SRs method. The black arrow represents the split reads (SRs) that fall into the areas of length 2*L* and are extracted from BAM files. The red arrow represents the forward search direction, and the yellow arrow represents the backward search direction. *L* represents the depth of the search. The blue area represents the candidate CNV area, and the gray area represents the expanded parts of the candidate CNV area, which constitute an expanded CNV region.

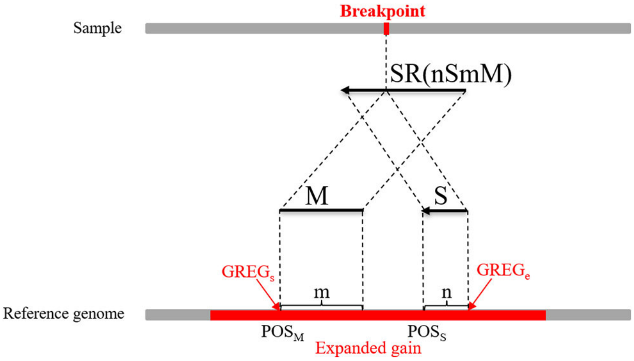


FIGURE 3 | An example of the boundary detection method in the gain area. The split reads (SRs) come from a sequenced sample and contain breakpoint information, such as SR (*mSnM*). *M* represents the matched part of the SR, while *S* represents the unmatched part. Moreover, *m* represents the length of *M*, and *n* represents the length of *S*. The *M* and *S* of the SR are mapped on the reference genome. POS_M and POS_S represent the alignment positions of the first base at the left ends of *M* and *S*, respectively. The red area represents an expanded gain, the boundary of which is represented by $GREG_s$ and $GREG_e$.

n SR's (from SR'_1 to SR'_n) fall into the detection range centered on start and end, length of which are 2L, respectively. Based on the extracted SRs and SR's, the boundary of the candidate CNV (blue area) is corrected in the expanded CNV (Blue area plus two gray areas). Next, the SRs that contain a substantial amount of breakpoint information are used to accurately define the boundaries of the mutated regions. According to the type of mutation zone, the correction procedure is composed of two parts. (1) The first part of the correction procedure is the correction of the candidate gain regions. The SM-type SRs are extracted in the gain region, which must meet the following conditions: a number of SRs greater than or equal to two, and a length of *S* that is the largest, which is perfectly matched in the expanded gain region. The SM-type SRs are used to calculate the starting point of the mutation region ($GREG_s$) and the endpoint of the mutation region ($GREG_e$), which are expressed by Equations (15, 16), respectively:

$$GREG_s = POS_M, \tag{15}$$

$$GREG_e = POS_S + n - 1, \tag{16}$$

where POS_M indicates the alignment position of the first base at the left end of *M*, POS_S indicates the alignment position of the first base at the left end of *S*, which can be calculated using the string-matching method, and *n* indicates the length of *S*. An example of the boundary correction process of the CNV gain area is described in detail in **Figure 3**. An SR is split into two parts (M and S) from the breakpoint. M and S are compared to the expanded gain (red area), and their length (m and n) and position (POS_M and POS_S) are used to correct the boundary of the expanded CNV ($GREG_s$ and $GREG_e$). (2) The second part of the correction procedure is the correction of the candidate loss regions. The MS-type SRs and SM-type SRs are extracted in the candidate loss regions. If the overlap between a MS-type SR and a SM-type SR exceeds 60%, this pair of SRs is considered to contain the same breakpoints. Moreover, the number of pairs of SRs is guaranteed to be more than two, and those SRs with the greatest length of *S* are chosen. A dynamic programming algorithm is adopted to calculate the overlapping ratio. MS-type SRs are used to calculate the starting point of the mutation region ($LREG_s$), and SM-type SRs are used to calculate the endpoint of the mutation region ($LREG_e$) (Wu et al., 2013), which are respectively described by Equations (17, 18):

$$LREG_s = POS + m, \tag{17}$$

$$LREG_e = POS' - 1, \tag{18}$$

where POS indicates the alignment position of the first base at the left end of *M* of the MS-type SR, *m* indicates the length of *M*, and $POS^{'}$ indicates the alignment position of the first base at the left end of *M* of the MS-type SR. An example of the boundary correction process of the CNV loss area is described in detail in **Figure 4**. SR and SR' are split into two parts (M and S) from the breakpoint, respectively. The Ms and Ss are compared to the expanded loss (blue area), and their length (m, m', n, and n') and

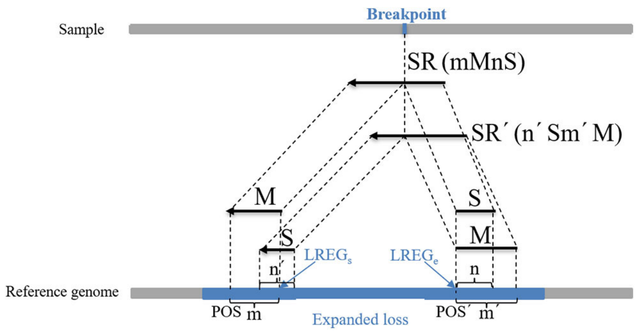


FIGURE 4 | An example of the boundary detection method in the loss area. The split reads (SRs) come from a sequenced sample, and contain breakpoint information, such as SR (*mMnS*) and SR'(*n'Sm'M*). *M* represents the matched part of SR and SR', and *S* represents the unmatched part. Moreover, *m* and *m'* represent the length of *M* of SR and SR', and *n* and *n'* represent the length of *S*. This pair of SRs contains the same breakpoint, and their *M* and *S* are mapped on the reference genome. *POS* and *POS'* represent the alignment positions of the first base at the left ends of *M* of SR and SR', respectively. The blue area represents an expanded loss, the boundary of which is represented by *LREG$_s$* and *LREG$_e$*.

position (POS and POS') are used to correct the boundary of the expanded CNV (LREG$_s$ and LREG$_e$).

RESULTS AND DISCUSSION

To evaluate and verify the performance of RKDOSCNV, it was tested using both simulated and real data. In the simulated data experiment, the proposed method was evaluated by comparing it with four published algorithms (CNV_IFTV, CNVnator, FREEC, and SeqCNV). Three performance indicators (recall, precision, and F1-score) were adopted to assess the performance of each method. In the real data experiment, three types of data (three real samples from the 1,000 Genomes Project, two ovarian cancer samples and one breast cancer sample) were used to test RKDOSCNV. The CNVs of three real samples from the 1000 Genomes Project were provided in the Database of Genomic Variants, which was used as the ground truth file to calculate the recall, precision, and F1-score of each method to evaluate their performances. Via the analysis of the real data samples, it was found that RKDOSCNV detected that some CNVs have important biological significance, which can provide powerful assistance for cancer prevention and targeted drug development.

Simulated Data Experiments

The simulated data sets were generated by IntSIM software (Yuan et al., 2017). Before using the software, the following settings were made for the reference genome as input, tumor purity (TP), and sequencing coverage (SC). Chromosome 21 of hg19 was chosen as the reference genome, the TP was set to 0.1 or 0.2, and the SC was set to 30×. The ground truth file was composed of six gains, four hemizygous losses (hemi-losses) and four homozygous losses (homo-losses). Under each set of settings, 50 simulation samples were generated.

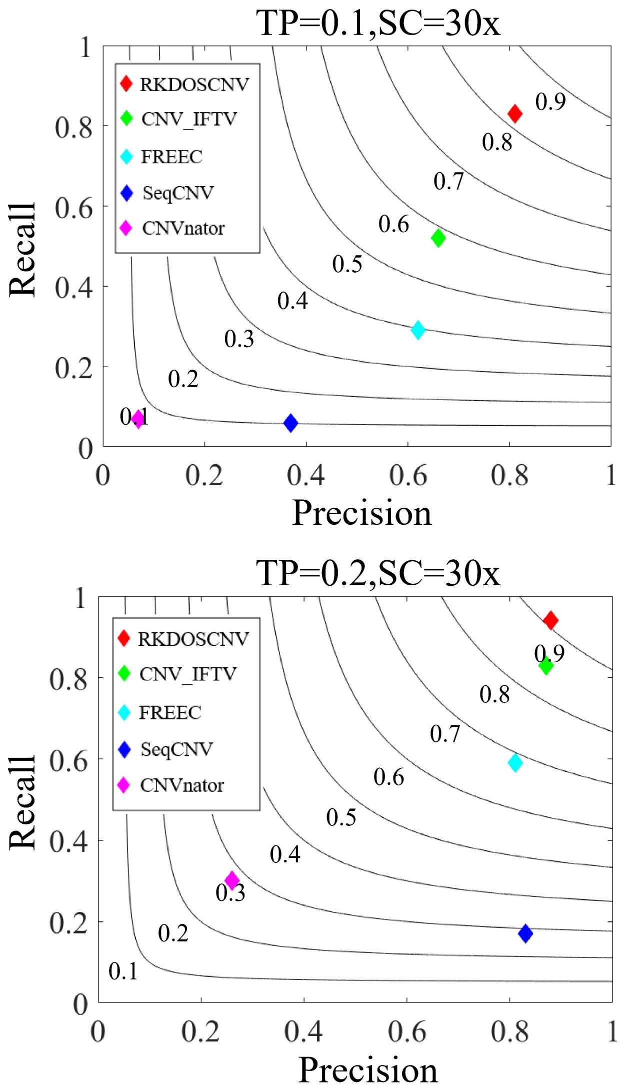


FIGURE 5 | The performance of five methods was evaluated on two sets of simulated data sets. The F1-score of each method was calculated as the evaluation indicator of performance. Black curves indicate F1-score levels that are harmonic means of recall and precision. The equations on the left and right sides of the comma represent the tumor purity (TP) and sequencing coverage (SC), respectively.

Based on these simulated data sets, the performances of RKDOSCNV and four published methods (CNV_IFTV, CNVnator, FREEC, and SeqCNV) were tested to calculate their recall, precision, and F1-score. Recall is defined as the ratio between the number of CNVs correctly detected by a method and the total number of CNVs in the ground truth file. Precision is defined as the ratio between the number of CNVs correctly detected by a method and the total number of CNVs detected by the method. The F1-score is defined as the harmonic mean of precision and recall. The comparison results are presented in detail in **Figure 5**, from which it is evident that the overall trend of performance changes of

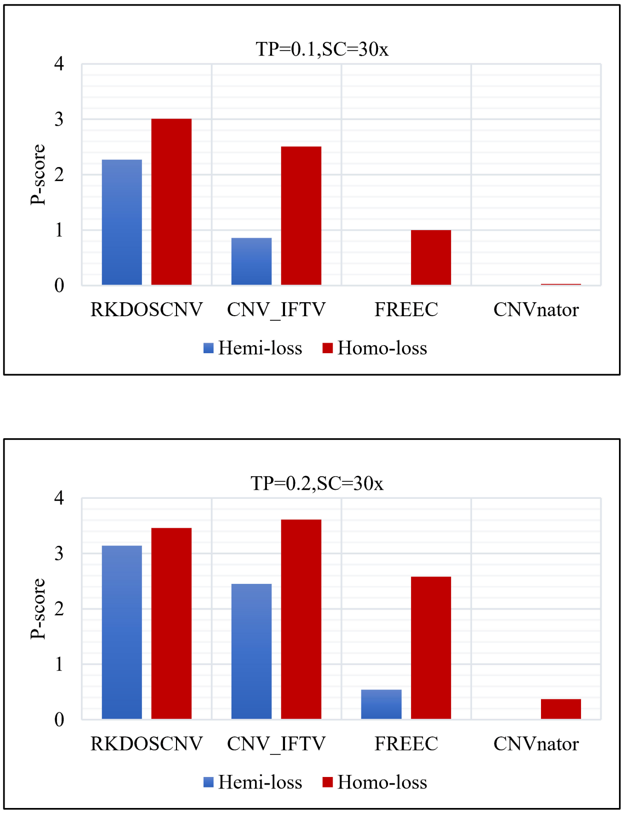


FIGURE 6 | Histogram of the number of hemizygous loss (hemi-loss) and homozygous loss (homo-loss) for each method under two different sets of conditions. The equations on the left and right sides of the comma represent the tumor purity (TP) and sequencing coverage (SC), respectively.

each method increased along with the increase in TP. For example, the smallest F1-score was close to 0.1 when the TP was set to 0.1, and the largest F1-score was >0.9 when the TP was set to 0.2. Under each set of conditions, RKDOSCNV achieved the best F1-score, followed by CNV_IFTV, FREEC, SeqCNV, and finally CNVnator. CNVnator achieved the lowest precision under each set of conditions; it detected many long CNVs, most of which were false positives. SeqCNV achieved the lowest recall but higher precision, which demonstrates that the detection of this method was more conservative than the other methods. The performance of FREEC was better than those of CNVnator and SeqCNV. Its performance improved by nearly 0.3 when the TP was increased from 0.1 to 0.2, which demonstrates that it is not sufficiently sensitive to detect samples with extremely low purity. CNV_IFTV achieved a better tradeoff between recall and precision than these three methods. Of all the tested methods, RKDOSCNV achieved both the highest recall and precision. Via the preceding analysis and discussion, it is proven that RKDOSCNV is an effective and reliable CNV detection method.

To verify this conclusion, the detection power score (P-score) of each method was detected in the regions of insignificant variation (hemi-loss regions) where the copy number is one, which can easily be mistakenly detected as normal areas. The P-score is expressed as follows:

$$P - score = L_n \cdot \frac{L_n}{L_n + L_{fp}} \tag{19}$$

where L_n represents the number of correctly detected loss areas (hemi-loss or homo-loss) and L_{fp} represents the number of false positive loss areas; the higher the P-score, the better the detection power. Moreover, the numbers of homo-loss regions were also counted. Four methods (RKDOSCNV, CNV_IFTV, CNVnator, and FREEC) were chosen for comparison, and SeqCNV was not included because it did not detect any loss areas. The P-scores of each method in both regions are described in detail in **Figure 6**, where it is evident that the P-score of each method presented an increasing trend with the increase of the TP from 0.1 to 0.2. The P-score of homo-loss of each method was higher than the P-score of hemi-loss of each method under all settings, which demonstrates that homo-loss can be more easily detected than hemi-loss. When the TP was equal to 0.1, the score of hemi-loss of RKDOSCNV was far superior to those of the other methods, which demonstrates that it is a reliable tool for the detection of insignificant CNVs. CNV_IFTV achieved higher P-scores and ranked second among all methods. FREEC achieved lower P-scores of hemi-losses than RKDOSCNV and CNV_IFTV, and did not detect any hemi-loss when the TP was equal to 0.1. CNVnator did not detect any hemi-loss under two sets of conditions. Overall, RKDOSCNV achieved the highest P-score, followed by CNV_IFTV, FREEC, and CNVnator, which is consistent with the conclusion of the simulation experiment.

As a supplement to these experiments, the correctness of the boundaries of correctly detected CNVs was further evaluated. The correctness of boundary (COB) is defined as the number of the boundary of correctly detected CNVs, the starting and ending points of which are consistent and compared with the ground truth file. As presented in **Figure 7**, that COB of each method is described using box plots under each set of simulation conditions, which includes 50 simulated samples. **Figure 7** shows that the COBs exhibited an increasing trend with the increase of TP. RKDOSCNV achieved the highest average COB value under each set of conditions. The COB of CNV_IFTV was better than that of FREEC when TP was equal to 0.1; on the contrary, the COB of FREEC was better than that of CNV_IFTV when TP was equal to 0.2. Their COB values were relatively close under two sets of conditions. Both CNV_IFTV and FREEC get a higher COB than the other methods, but their average COB is lower than RKDOSCNV. CNVnator detected fewer correct CNV boundaries as compared to the above three methods, and SeqCNV could not detect the correct CNV boundaries under each set of conditions. Via the analysis and discussion of these experiments, it is evident that the performance of RKDOSCNV was the best among all compared methods, which fully verifies that it is a reliable and effective CNV detection tool.

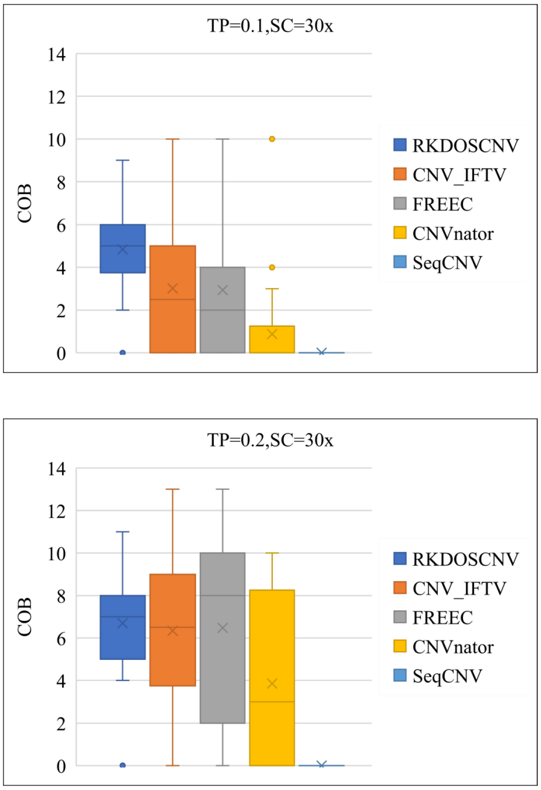


FIGURE 7 | Box plots of the correctness of boundary (COB) of each method under two different sets of conditions.

Detection of Real Samples From the 1,000 Genomes Project

To validate the proposed RKDOSCNV method, it was applied to the analysis of three real sequencing samples (NA12878, NA12891, and NA12892), which were downloaded from published article(Yuan et al., 2019). It was compared with three other single sample-based methods (CNV_IFTV, CNVnator, and FREEC). The test results of the three samples are recorded in the DGV database, by which the recall, precision, and F1-score of each method can be roughly calculated. The comparison results are described in detail in **Figure 8**. For the NA12878 sample, RKDOSCNV achieved the highest precision and a moderate recall. FREEC achieved a better recall and the lowest precision because it detected many false positives. CNV_IFTV achieved a better balance between recall and precision, and CNVnator achieved a lower precision and the highest recall. For the NA12878 sample, RKDOSCNV achieved the best F1-score, followed by CNV_IFTV, CNVnator, and FREEC. For the NA12891 sample, RKDOSCNV achieved the best F1-score, followed by CNV_IFTV, CNVnator, and FREEC. For the NA12892 sample, the F1-score of CNV_IFTV ranked first, followed by those of RKDOSCNV, CNVnator, and FREEC. Overall, from the results of the compared methods on the three samples, RKDOSCNV achieved a superior trade-off between recall and precision as compared to the other three methods. The preceding analysis and discussion demonstrate that RKDOSCNV is a relatively reliable CNV detection tool.

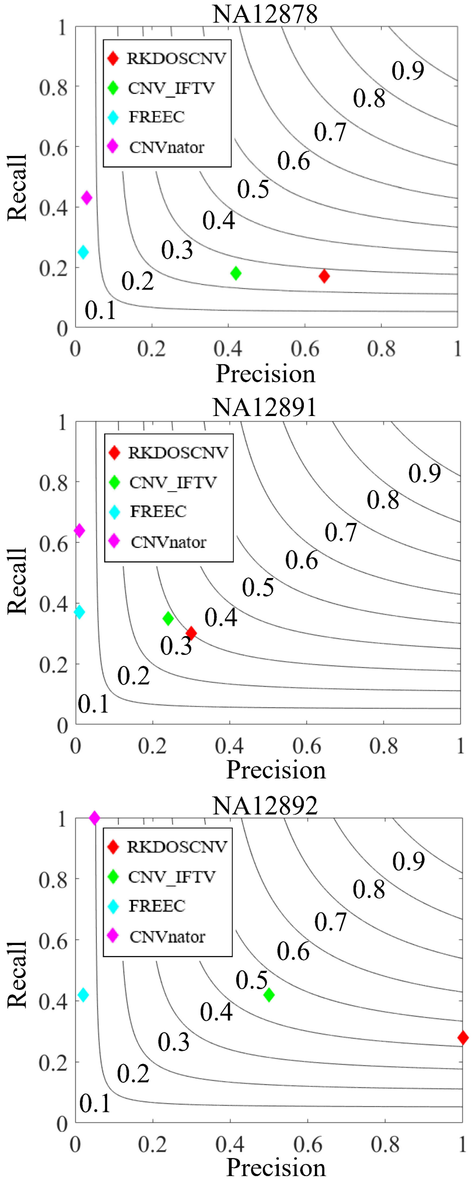


FIGURE 8 | The performances of four methods were evaluated on three real data samples (NA12878, NA12891, and NA12892). Black curves indicate F1-score levels that are harmonic means of recall and precision.

Detection of Two Ovarian Cancer Samples

We used RKDOSCNV to detect the genome-wide data of two ovarian cancer samples (EGAR00001004838 and EGAR00001004839), which can download at published article (Yuan et al., 2018). Here, 22 autosomes of each sample are used for analysis. RKDOSCNV was compared with the other two methods (CNV_IFTV and FREEC). Number of overlapping events and Number of predicted events of each method are recorded in detail in **Table 1**. FREEC gets the most overlapping

TABLE 1 | Number of overlapping events and predicted events for each method on three real datasets.

Sample ID	Type	RKDOSCNV	CNV_IFTV	FREEC
EGAR00001004838	Number of overlapping events	10,912	2,319	11,529
	Number of predicted events	56,110	15,985	619,312
EGAR00001004839	Number of overlapping events	10,537	2,200	11,422
	Number of predicted events	22,764	17,308	625,767
PD4192a	Number of overlapping events	4,128	3,376	5,230
	Number of predicted events	16,672	13,848	306,257

TABLE 2 | Summary of ODS for each method on three real datasets.

Sample ID	RKDOSCNV	CNV_IFTV	FREEC
EGAR00001004838	4,874	441	1,155
EGAR00001004839	7,177	432	1,982
PD4192a	1,739	1,279	725

events, but it detects the most non-overlapping events. The number of overlapping events detected by RKDOSCNV is slightly less than FREEC. It detects a moderate number of non-overlapping events. CNV_IFTV Obtain the least overlapping events and non-overlapping events in each sample. In order to further compare the performance of each method, we use Equation (20) to calculate the overlapping density score (ODS) (Yuan et al., 2019) of each method.

$$ODS = M_o \cdot M_p, \tag{20}$$

Where M_o represents the mean of the number of overlapping events (the mean of the intersection between one method and the other methods), M_p represents the ratio between M_o of the method and number of events predicted by it.

The corresponding comparison results are listed in Table 2. RKDOSCNV get the highest ODS in each sample. In Table 1, FREEC detected the most overlapping events, but its ODSs are lower than RKDOSCNV. This is because it detects a large number of non-overlapping events. CNV_IFTV gets the lowest ODSs in two samples. In conclusion, RKDOSCNV predicted a moderate amount of CNV events in these samples and showed a relatively high overlapping density compared to other methods.

Detection of a Breast Cancer Sample

To further validate RKDOSCNV, it was applied to the detection of CNVs in a breast cancer whole-genome sample (PD4192a), which was downloaded from published article (Li et al., 2019). The 22 autosomes of the breast cancer sample were extracted by SAMtools (Li et al., 2009). CNV_IFTV and FREEC were selected for comparison with RKDOSCNV. The experimental results were described in detail in **Tables 1**, **2**. FREEC predicts the most CNV events and overlapping events, but it gets lowest ODS. It detected a large number of CNV events, most of which were long CNVs and proved to be false positives in the previous experiments. CNV_IFTV detects the fewest CNV events and overlapping events, and gets ODS between FREEC and RKDOSCNV. RKDOSCNV predicts the moderate number of CNV events and overlapping events, and it gets the highest ODS in all methods. **Figure 9** presents an overview of the detected CNV distribution of the 22 autosomes, which is composed of four rings. From the outside to the inside, the first ring represents the distribution and length of the 22 autosomes, and the second, third, and fourth rings represent the detection of the gain and loss regions by FREEC, CNV_IFTV, and RKDOSCNV, respectively. The red and blue dots represent gain areas and loss areas, respectively. It was found that FREEC detects a large number of CNVs in each chromosome. CNV_IFTV did not detect any CNVs in chromosomes 6, 10, and 22, whereas the other two methods did detect CNVs. It shows that the detection of this method is more conservative than the other two methods. RKDOSCNV detected a moderate number of CNVs, and relatively more gain areas than loss areas.

Based on the preceding analysis, the biological meanings of detected CNVs were further investigated, many of which are associated with cancer or complex diseases. For example, the CNV gains at 1p11.2 (Jiang et al., 2011), 11q21 (Kazantseva et al., 2016), 16q11.2 (Savelyeva et al., 1994), and 17q11.1 are associated with breast cancer and are detected by each method, which is described in **Figure 9**. Those at 14q11.2 (Kawasaki et al., 2007) and 16p11.2 (Weiss et al., 2008) are respectively related to lung cancer and autism, that at 10q11.21 (Rees et al., 2016) is associated with schizophrenia, those at 1q21 (Grzasko et al., 2012) and 2q12.3 (Erickson et al., 2014) are associated with multiple myeloma, and those at 14q11.1 (Thean et al., 2018) and 18q21.1 (Druliner et al., 2018) are associated with colorectal cancer. CNV loss at 20q13.2 (Hidaka et al., 2000) is associated with colorectal cancer, that at 14q32.33 (Ledet et al., 2013) is associated with prostate cancer, that at 4q13.2 (Yang et al., 2008) is associated with osteoporosis, and that at 3q29 (Biamino et al., 2016) is related to autism. Via the preceding analysis, it is found that the four methods are very effective and can detect some valuable CNVs, and can therefore provide great assistance for clinical treatment and drug development.

DISCUSSION AND CONCLUSION

In this study, a new method called RKDOSCNV was presented for CNV detection via the use of NGS data. RKDOSCNV was developed based on DOC and SR methods and uses a local perspective to detect CNVs, which replaces general methods that use global modeling to predict CNVs. RKDOSCNV is a single sample-based CNV detection method, and does not require the provision of paired samples. Via the verification of experiments on simulated and real data, it was proven that RKDOSCNV can detect many meaningful CNVs, and can provide effective assistance for the development of targeted drugs

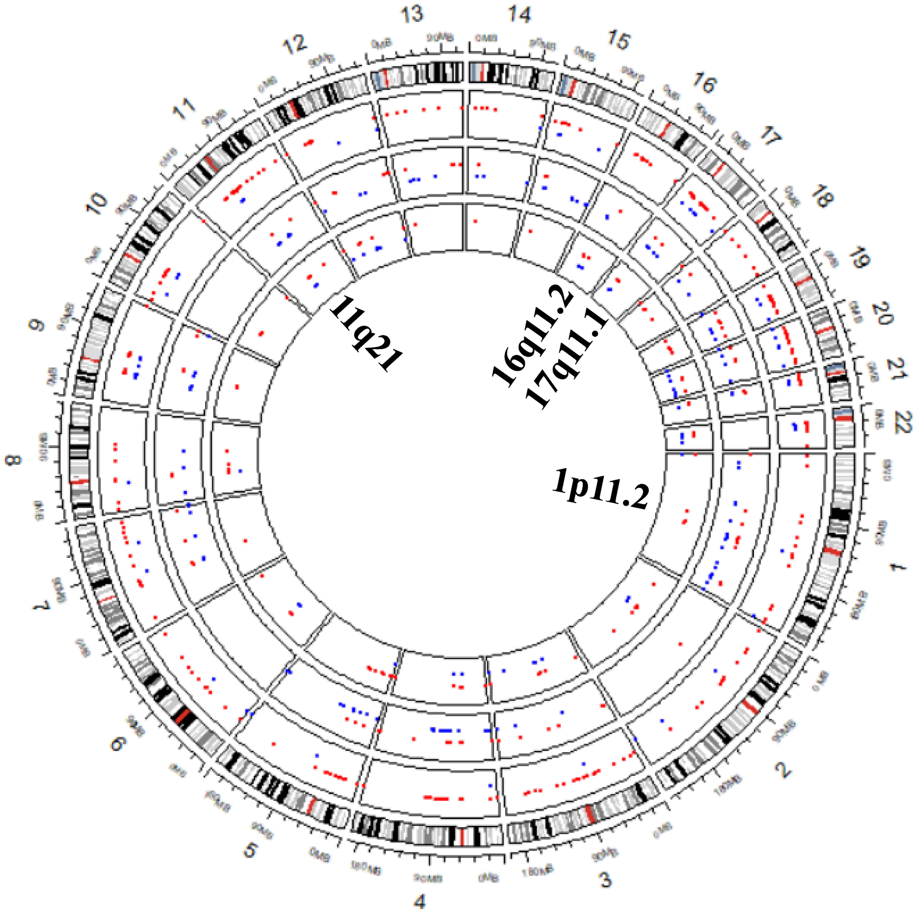


FIGURE 9 | The Circos plot shows that the detected CNVs of each method are distributed in the 22 autosomes of a breast cancer sample. The red and blue dots represent gain areas and loss areas, respectively. The outermost ring describes the distribution of 22 autosomes. From the outside to the inside, the second, third, and fourth rings present the test results of FREEC, CNV_IFTV, and RKDOSCNV, respectively. The innermost ring corresponds to the four cytobands (1p11.2, 11q21, 16q11.2, and 17q11.1) detected by each method, which are associated with breast cancer.

and cancer prediction. It is unique as compared to traditional detection methods that build statistical models based on global data and use hypothesis testing methods to predict CNVs. The three characteristics of RKDOSCNV are defined as follows. (1) Unlike traditional methods, RKDOSCNV does not need to make assumptions about the distribution of RD signals, and discerns the difference of RD signals from a local perspective. (2) By calculating the three types of neighbor relations of RDS signals, RKDOSCNV can successfully detect many insignificant CNVs. (3) Based on detected candidate CNVs, RKDOSCNV uses the SR approach to further determine the boundaries of candidate CNVs.

The effectiveness of RKDOSCNV was verified using both simulated and real data sets. In the simulated data experiment, RKDOSCNV was compared with four existing algorithms, and three performance indicators (recall, precision, and F1-score) of each method were analyzed to measure their performances. The ability of each method to detect insignificant CNVs and correctly identify the number of CNV boundaries was further evaluated. The experimental results demonstrate that RKDOSCNV achieved the best performance in terms of the F1-score, P-score, and COB. In the real data experiments, the performance of RKDOSCNV was evaluated using six real samples, and the biological significance of the detected CNVs was analyzed and discussed. Overall, RKDOSCNV is an effective and reliable CNV detection tool, especially for tumor samples of low purity.

During the experiment, some shortcomings were discovered. For example, the selection of the number of neighbors (*k*) is a critical step in the proposed method; it is set as an empirical value with reference to the traditional outlier detection methods (Breunig et al., 2000; Jin et al., 2006; Tang and He, 2017), which meets the needs of most situations, but may not be suitable in extreme cases. The threshold (θ) setting also has a great influence on the accuracy of the detection results. Based on applications in different scenarios (Tang and He, 2017), a moderate baseline was chosen to meet the application needs. The proposed method does not support detection of interspersed amplification, which is an important form of mutation. We will expand the functionality of the method so that it can detect multiple types of mutations. The CNVs detected by RKDOSCNV includes germline and somatic CNVs, but it cannot distinguish between the two variant types. If a control sample is input, our method can identify germline and somatic CNVs. Currently, there is no matched control sample, so the detection result is a mixture of two types of variants. In this study, the methods compared by RKDOSCNV detect all CNVs (germline and somatic CNVs). In future work, these problems will be addressed to improve the performance of RKDOSCNV, and reasonable methods will be developed to automatically select the optimal parameters and effectively identify other types of mutations. Based on the existing methods, other distance measurement methods and density evaluation methods will be chosen to further improve the performance of RKDOSCNV in the accurate and effective identification of CNVs.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

GL participated in the design of the algorithms and experiments. JZ and XY participated in the design of the entire framework of CNV detection and directed the whole work. CW participated in the analysis of the performance of the proposed method. JZ and XY conceived of the study and helped revise the manuscript. All authors read the final manuscript and agreed on its contents for submission.

FUNDING

This work was supported by the National Natural Science Foundation of China under (Grant No. 61571341) and the Fundamental Research Funds for the Central Universities and Innovation Fund of Xidian University (Grant No. CS2020-16).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.569227/full#supplementary-material

REFERENCES

Abyzov, A., and Gerstein, M. (2011). AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* 27, 595–603. doi: 10.1093/bioinformatics/btq713

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110

Biamino, E., Di Gregorio, E., Belligni, E. F., Keller, R., Keller, R., Gandione, M., et al. (2016). A novel 3q29 deletion associated with autism, intellectual disability, psychiatric disorders, and obesity. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* 171, 290–299. doi: 10.1002/ajmg.b.32406

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J. P., Janoueix-Lerosey, I., Delattre, O., et al. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* 27, 268–269. doi: 10.1093/bioinformatics/btq635

Breunig, M. M., Kriegel, H. P., Ng, R. T., and Sander, J. R. (2000). "LOF: Identifying density-based local outliers," in *Proceeding ACM SIGMOD 2000 International Conference on Mangement of Data* (Dallas, TX).

Cai, H., Chen, P., Chen, J., Cai, J., Song, Y., and Han, G. (2018). WaveDec: a wavelet approach to identify both shared and individual patterns of copy-number variations. *IEEE Trans. Biomed. Eng.* 65, 353–364 doi: 10.1109/TBME.2017.2769677

Chen, Y., Zhao, L., Wang, Y., Cao, M., Gelowani, V., Xu, M., et al. (2017). SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data. *BMC Bioinformatics* 18:147. doi: 10.1186/s12859-017-1566-3

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., et al. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 6, 99–103. doi: 10.1038/nmeth.1276

Condat, L. (2013). A direct algorithm for 1-D total variation denoising. *IEEE Signal Process. Lett*. 20, 1054–1057 doi: 10.1109/LSP.2013.2278339

Dharanipragada, P., Vogeti, S., and Parekh, N. (2018). iCopyDAV: Integrated platform for copy number variations-Detection, annotation and visualization. *PLoS ONE* 13:e0195334. doi: 10.1371/journal.pone.0195334

Dohm, J. C., Claudio, L., Tatiana, B., and Heinz, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*. 36:e105. doi: 10.1093/nar/gkn425

Druliner, B. R., Ruan, X., Sicotte, H., O'Brien, D., Liu, H., Kocher, J.-P. A., et al. (2018). Early genetic aberrations in patients with sporadic colorectal cancer. *Mol. Carcinog.* 57, 114–124. doi: 10.1002/mc.22738

Duan, J., Zhang, J. G., Deng, H. W., and Wang, Y. P. (2013). CNV-TV: A robust method to discover copy number variation from short sequencing reads. *BMC Bioinformatics* 14:150. doi: 10.1186/1471-2105-14-150

Erickson, S., Raj, V., Stephens, O., Dhakal, I., Chavan, S., Sanathkumar, N., et al. (2014). Genome-wide scan identifies variant in 2q12.3 associated with risk for multiple myeloma. *Blood* 124, 2001–2003. doi: 10.1182/blood-2014-07-586701

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., Mccarroll, S. A., Altshuler, D. M., et al. (2006). Copy number variation: new insights in genome diversity. *Genome Res*. 16, 949–961. doi: 10.1101/gr.3677206

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet*. 91, 597–607. doi: 10.1016/j.ajhg.2012.08.005

Grzasko, N., Hus, M., Pluta, A., Jurczyszyn, A., Walter-Croneck, A., Morawska, M., et al. (2012). Additional genetic abnormalities significantly worsen poor prognosis associated with 1q21 amplification in multiple myeloma patients. *Hematol. Oncol*. 33, 41–48. doi: 10.1002/hon.2018

Guan, P., and Sung, W. (2016). Structural variation detection using next-generation sequencing data a comparative technical review. *Methods* 102, 36–49. doi: 10.1016/j.ymeth.2016.01.020

Gusnanto, A., Wood, H. M., Pawitan, Y., Rabbitts, P., and Berri, S. (2012). Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* 28, 40–47. doi: 10.1093/bioinformatics/btr593

Hidaka, S., Yasutake, T., Takeshita, H., Kondo, M., Tsuji, T., Nanashima, A., et al. (2000). Differences in 20q13.2 copy number between colorectal cancers with and without liver metastasis. *Clin. Cancer Res.* 6, 2712–2717.

Ivakhno, S., Royce, T., Cox, A. J., Evers, D. J., Cheetham, R. K., and Tavaré, S. (2010). CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* 26, 3051–3058. doi: 10.1093/bioinformatics/btq587

Jiang, Y., Oldridge, D. A., Diskin, S. J., and Zhang, N. R. (2015). CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.* 43:e39. doi: 10.1093/nar/gku1363

Jiang, Y., Shen, H., Liu, X., Dai, J., Jin, G., Qin, Z., et al. (2011). Genetic variants at 1p11.2 and breast cancer risk: a two-stage study in Chinese women. *PLoS ONE* 6:e21563. doi: 10.1371/journal.pone.0021563

Jin, W., Tung, A. K. H., Han, J., and Wang, W. (2006). Ranking outliers using symmetric neighborhood relationship. *Adv. Knowl. Discov. Data Mining* 3918, 577–593. doi: 10.1007/11731139_68

Kawasaki, T., Yokoi, S., Tsuda, H., Izumi, H., Kozaki, K., Aida, S., et al. (2007). BCL2L2 is a probable target for novel 14q11.2 amplification detected in a non-small cell lung cancer cell line. *Cancer Sci.* 98, 1070–1077. doi: 10.1111/j.1349-7006.2007.00491.x

Kazantseva, P., Slonimskaya, E., Litviakov, N., and Tsyganov, M. (2016). Individual approach to the planning of neoadjuvant chemotherapy (NAC) in patients with luminal B breast cancer. *Ann. Oncol.* 27:V161. doi: 10.1093/annonc/mdw364.61

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., et al. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res.* 22, 1525–1532. doi: 10.1101/gr.138115.112

Layer, R., Chiang, C., Quinlan, A., and Hall, I. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 15:R84. doi: 10.1186/gb-2014-15-6-r84

Ledet, E. M., Hu, X., Sartor, O., Rayford, W., Li, M., and Mandal, D. (2013). Characterization of germline copy number variation in high-risk African American families with prostate cancer. *Prostate* 73, 614–623. doi: 10.1002/pros.22602

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, Y., Zhang, J., and Yuan, X. (2019). BagGMM: Calling copy number variation by bagging multiple Gaussian mixture models from tumor and matched normal next-generation sequencing data. *Digit. Signal Process* 88, 90–100. doi: 10.1016/j.dsp.2019.01.025

Magi, A., Pippucci, T., and Sidore, C. (2017). XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC Genomics* 18:747. doi: 10.1186/s12864-017-4137-0

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* 6, S13–20. doi: 10.1038/nmeth.1374

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/nrg2626

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:e16327. doi: 10.1371/journal.pone.0016327

Povysil, G., Tzika, A., Vogt, J., Haunschmid, V., Messiaen, L., Zschocke, J., et al. (2017). panelcn.MOPS: copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum. Mutat.* 38, 889–897. doi: 10.1002/humu.23237

Rausch, T., Zichner, T., Schlattl, A., Stutz, A., Benes, V., and Korbel, J. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28, I333–I339. doi: 10.1093/bioinformatics/bts378

Rees, E., Kendall, K., Pardinas, A. F., Legge, S. E., Pocklington, A., Escott-Price, V., et al. (2016). Analysis of intellectual disability copy number variants for association with schizophrenia. *JAMA Psychiatry* 73, 963–969. doi: 10.1001/jamapsychiatry.2016.1831

Savelyeva, L., Schneider, B., Finke, L., Schlag, P., and Schwab, M. (1994). Amplification of satellite DNA at 16q11.2 in the germ-line of a patient with breast-cancer. *Int. J. Oncol.* 4, 347–351 doi: 10.3892/ijo.4.2.347

Smith, S. D., Kawash, J. K., and Grigoriev, A. (2015). GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. *PeerJ.* 3:e836. doi: 10.7717/peerj.836

Stankiewicz, P., and Lupski, J. (2010). Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* 61, 437–455. doi: 10.1146/annurev-med-100708-204735

Tang, B., and He, H. (2017). A local density-based approach for outlier detection. *Neurocomputing* 241, 171–180. doi: 10.1016/j.neucom.2017.02.039

Thean, L. F., Low, Y. S., Lo, M., Teo, Y.-Y., Koh, W.-P., Yuan, J.-M., et al. (2018). Genome-wide association study identified copy number variants associated with sporadic colorectal cancer risk. *J. Med. Genet.* 55, 181–188. doi: 10.1136/jmedgenet-2017-104913

Weiss, L. A., Yiping, S., Korn, J. M., Dan, E., Arking, M. D. T., Ragnheidur, F., et al. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* 358, 667–675. doi: 10.1056/NEJMoa075974

Wu, Y., Tian, L., Pirastu, M., Stambolian, D., and Li, H. (2013). MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads. *Front. Genet.* 4:157. doi: 10.3389/fgene.2013.00157

Xi, R., Luquette, J., Hadjipanayis, A., Kim, T.-M., and Park, P. J. (2010). BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. *Genome Biol.* 11, 1128–1136. doi: 10.1186/1465-6906-11-S1-O10

Yang, T.-L., Chen, X.-D., Guo, Y., Lei, S.-F., Wang, J.-T., Zhou, Q., et al. (2008). Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *Am. J. Hum. Genet.* 83, 663–674. doi: 10.1016/j.ajhg.2008.10.006

Ye, K., Schulz, M., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865–2871. doi: 10.1093/bioinformatics/btp394

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Yuan, X., Bai, J., Zhang, J., Yang, L., Duan, J., Li, Y., et al. (2018). CONDEL: detecting copy number variation and genotyping deletion zygosity from single tumor samples using sequence data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1147–1153. doi: 10.1109/TCBB.2018.2883333

Yuan, X., Gao, M., Bai, J., and Duan, J. (2020). SVSR: a program to simulate structural variations and generate sequencing reads for multiple platforms. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 17, 1082–1091. doi: 10.1109/TCBB.2018.2876527

Yuan, X., Yu, J., Xi, J., Yang, L., Shang, J., Li, Z., et al. (2019). CNV_IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2920889. [Epub ahead of print].

Yuan, X., Zhang, J., and Yang, L. (2017). IntSIM: an integrated simulator of next-generation sequencing data. *IEEE Trans. Biomed. Eng.* 64, 441–451. doi: 10.1109/TBME.2016.2560939

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 14:S1. doi: 10.1186/1471-2105-14-S11-S1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.