frontiers in Genetics

METHODS
published: 28 June 2021
doi: 10.3389/fgene.2021.699510

A Cluster-Based Approach for the Discovery of Copy Number Variations From Next-Generation Sequencing Data

*Guojun Liu and Junying Zhang**

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

The next-generation sequencing technology offers a wealth of data resources for the detection of copy number variations (CNVs) at a high resolution. However, it is still challenging to correctly detect CNVs of different lengths. It is necessary to develop new CNV detection tools to meet this demand. In this work, we propose a new CNV detection method, called CBCNV, for the detection of CNVs of different lengths from whole genome sequencing data. CBCNV uses a clustering algorithm to divide the read depth segment profile, and assigns an abnormal score to each read depth segment. Based on the abnormal score profile, Tukey's fences method is adopted in CBCNV to forecast CNVs. The performance of the proposed method is evaluated on simulated data sets, and is compared with those of several existing methods. The experimental results prove that the performance of CBCNV is better than those of several existing methods. The proposed method is further tested and verified on real data sets, and the experimental results are found to be consistent with the simulation results. Therefore, the proposed method can be expected to become a routine tool in the analysis of CNVs from tumor-normal matched samples.

Keywords: next-generation sequencing, copy number variation, clustering algorithm, abnormal score, Tukey's fences

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Ruifeng Hu,
Harvard Medical School, United States
Cuncong Zhong,
University of Kansas, United States

***Correspondence:**
Junying Zhang
jyzhang@mail.xidian.edu.cn

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

Received: *23 April 2021*
Accepted: *07 June 2021*
Published: *28 June 2021*

Citation:
Liu G and Zhang J (2021) A Cluster-Based Approach for the Discovery of Copy Number Variations From Next-Generation Sequencing Data. Front. Genet. 12:699510. doi: 10.3389/fgene.2021.699510

INTRODUCTION

The copy number variation (CNV) of DNA fragments has been widely recognized as a major type of structural variations, and can cause the amplification or deletion of DNA fragments, the lengths of which are greater than 1 kbp in the human genome (Freeman et al., 2006). Some CNVs, called germline CNVs, are also present in normal tissues of the human body; these generally originate from family inheritance, and can cause cancers and diseases (Kuiper et al., 2010; Krepischi et al., 2012). The CNVs in tumor tissue are generally called somatic CNVs, which are acquired CNVs, and cause tumor formation by oncogene and tumor suppressor gene mutations (Stratton et al., 2009; Beroukhim et al., 2010; Pei et al., 2020). Many experimental studies have proven that CNVs can change the doses of genes and lead to the reorganization of chromosome structure (Sharp et al., 2005; Magi et al., 2017; Pei et al., 2021b), and makes an important contribution to the occurrence and formation of tumors and various disorders (Pei et al., 2021a). For example, it can cause schizophrenia and autism disorders in humans

(Sebat et al., 2007; Cook and Scherer, 2008; Stone et al., 2008). Some studies have shown that CNVs are related to cancer, such as breast and ovarian cancer (Tchatchou and Burwinkel, 2008; Adam and David, 2009; Malek et al., 2011). In practical applications, there is a strong requirement to capture CNVs of various range lengths, which requires the developed tools to have higher resolution and better robustness than previously developed tools to reduce the false positive rate of test results. Therefore, it is still a difficult task to effectively detect CNVs of different lengths.

Compared with traditional (array-based) detection methods (Carter, 2007; Buysse et al., 2009), the detection cost has been greatly reduced and resolution has reached the base-pair level with the emergence of next-generation sequencing technology. In recent years, most related tools for CNV detection using next-generation sequencing data have been developed based on paired-end mapping (PEM) (Korbel et al., 2007) and depth of coverage (DOC) (Yoon et al., 2009) strategies. The basic concept of PEM-based methods is that the insertion size of aligned paired-end reads is significantly different from the insertion size preset by the laboratory (Medvedev et al., 2009). While PEM-based methods can detect amplification, deletion, insertion, translocation, etc., they can only identify those insertion variants whose lengths are less than the preset insertion length. The basic concept of DOC-based methods is that the number of reads aligned to each position of the reference genome is proportional to the number of copies corresponding to that position (Yoon et al., 2009). In principle, DOC-based methods can detect CNVs of various lengths. However, in practical applications, they are more suitable for the detection of long CNVs, and cannot accurately detect the boundaries of the CNVs.

Generally, DOC-based methods require the input of tumor-normal matched samples to detect the tumor genome and effectively capture CNVs. The workflow of this type of method is: (1) input tumor-normal matched samples; (2) obtain read count profiles with SAMtools (Li et al., 2009); (3) bin read count profiles (Chiang et al., 2009) and generate read depth profiles; (4) use the joint read depth information of the tumor-normal matched samples to build a statistical model; (5) choose a suitable threshold to predict CNVs. It is generally believed that the deviation caused by sequencing is consistent in the same areas of the two samples. Therefore, DOC-based methods use the read depth ratio information to eliminate these deviations (GC content and mappability biases) (Bentley et al., 2008; Chiang et al., 2009). Some well-known methods have been developed to detect CNVs from tumor-normal matched samples, including BIC-seq2 (Xi et al., 2016), SeqCNV (Chen et al., 2017), and CNVkit (Talevich et al., 2016). BIC-seq2 preprocesses the sequenced reads, including by calibrating the GC content bias, removing mappability bias, and normalizing reads at the nucleic acid level. Based on the preprocessed data, the segmentation procedure is executed using the bayesian information criterion, by which CNVs are forecasted. It is not sensitive to the detection of short CNVs. SeqCNV extracts the read depth information of the tumor-normal matched samples to build a maximum penalized likelihood estimation model to predict CNVs. It detects a small number of CNVs, most of which are the gain areas and true positives, and its detection is more conservative than that of BIC-seq2. It has a long running time and is not suitable for testing samples with long CNVs. CNVkit is a software toolkit that extracts the information of on- and off-target sequenced reads. It adopts a rolling median method to normalize the GC content bias, mappability bias, and target density bias, and to reduce the impact on the true copy number status. CNVkit detects the CNVs, many of which are deletion regions. However, it is not sensitive to the detection of short CNVs.

In consideration of the limitations of the existing methods, in this study, a new tumor-normal matched sample-based CNV detection method, called CBCNV (cluster-based approach for CNV detection), is proposed for the prediction of CNVs using whole-genome sequencing data. CBCNV extracts the read count profiles of tumor-normal matched samples with SAMtools (Li et al., 2009). The preprocessing program is executed on the read count profiles, which can yield the read depth segment profiles, the dimensions of which are transformed into two-dimensional space. CBCNV adopts the k-means algorithm to cluster the preprocessed read depth segment profiles (Hartigan and Wong, 1979), which can yield clusters of different sizes. The clusters are sorted from largest to smallest according to the number of elements in each cluster. Then, by setting a boundary threshold, these clusters are divided into large and small clusters. Based on the above definition, CBCNV assigns a cluster-based abnormal score for each read depth segment. Using the cluster-based abnormal score profiles, Tukey's fences method is employed to announce candidate CNVs (Zijlstra et al., 2007). The performance of the proposed method is verified using simulated and real data sets, and is compared with several existing CNV detection methods. The experimental results show that the performance of CBCNV is better than several other comparison methods, especially for low-purity samples. In addition, CBCNV is also found to detect some biologically meaningful CNVs, which can provide some valuable reference information for assistance with clinical diagnosis and targeted drug research.

The remainder of this article is organized as follows. Section "Materials and Methods" includes the workflow of CBCNV, data preprocessing, the calculation of cluster-based abnormal score, and the prediction of CNVs. In section "Results," simulation and real experiments are designed, and the experimental results are analyzed and discussed. Section "Discussion and Conclusion" summarizes this research and puts forward ideas for future work.

MATERIALS AND METHODS

Overview of CBCNV

CBCNV is a DOC-based approach that is suitable for the detection of tumor-normal matched samples, and can identify somatic CNVs and germline CNVs from whole-genome sequencing data. The pipeline of CBCNV is described in detail in **Figure 1**. The sequenced tumor-normal matched samples that are composed of a large number of sequenced reads are compared to the reference genome using the BWA tool (Li and Durbin, 2010). Then, the read count profiles of the tumor-normal matched samples are generated with SAMtools (Li et al., 2009).

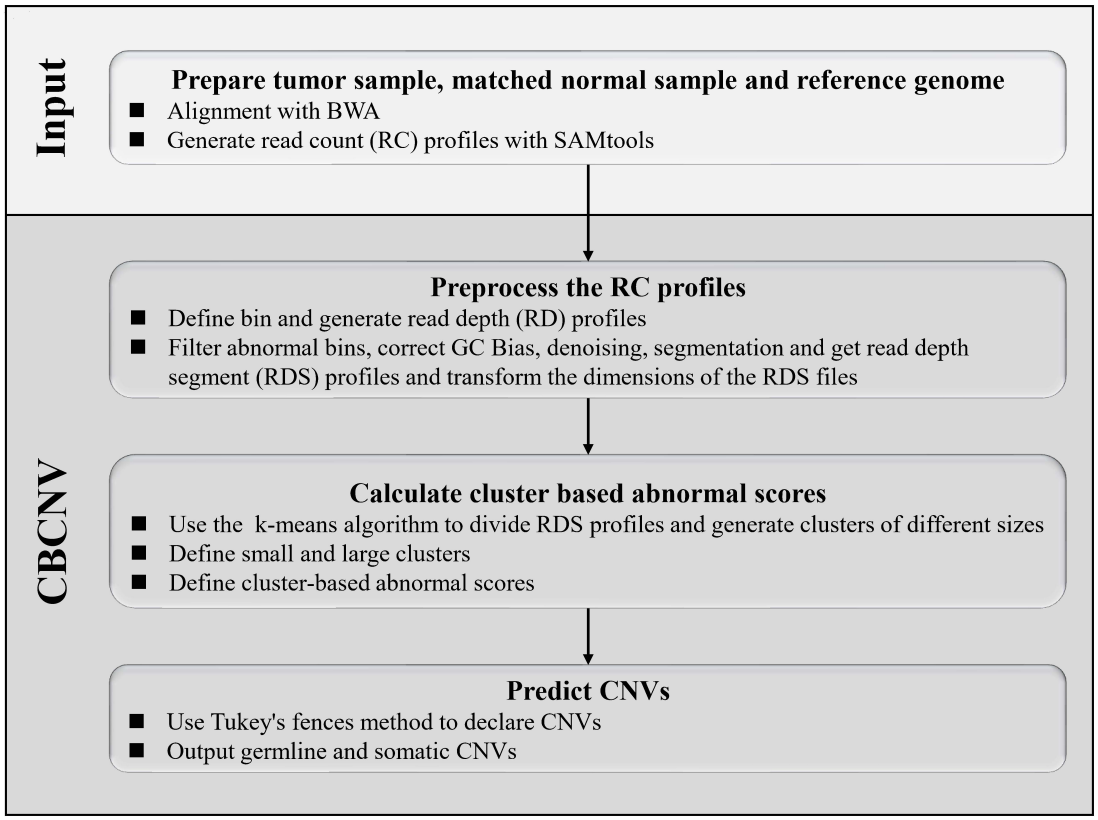

FIGURE 1 | Overview of the workflow of CBCNV. It is mainly composed of four steps, which includes preparing input files, preprocessing read count profiles, calculating cluster-based abnormal scores, and recording CNVs.

Based on the read count profiles, the following four steps are conducted for CBCNV to complete CNV detection. The first step involves defining the bin, dividing the reference genome into continuous and non-overlapping regions according to the bin size, and generating the read depth profiles. In the second step, the abnormal bins are removed, and the GC content bias is corrected. The read depth profiles are denoised and segmented to generate the read depth segment profiles. The dimensions of the read depth segment profiles are converted from one-dimensional to two-dimensional space. In the third step, the preprocessed read depth segment profiles are clustered via the k-means method to form clusters of different sizes. A boundary value is set to divide large and small clusters. The cluster-based abnormal score is defined based on the following two situations (He et al., 2003): (1) if a read depth segment belongs to a small cluster, the cluster-based abnormal score is defined as the distance between the read depth segment and the center of the large cluster that is the closest to it; (2) if a read depth segment belongs to a large cluster, the cluster-based abnormal score is defined as the distance between the read depth segment and the center of the large cluster. Finally, in the fourth step, Tukey's fences method is employed to predict CNVs (Zijlstra et al., 2007). The CBCNV software is developed based on the R and Python languages (Zhao et al., 2019). Its source code is public, and can be downloaded from https://github.com/gj-123/CBCNV/releases, where users can easily install and use the software according to the instructions.

Data Preprocessing

The sequenced reads are aligned to the reference genome with BWA (Li and Durbin, 2010), and the read count profiles are generated by SAMtools (Li et al., 2009). The reference genome is composed of five types of positions ("A", "T", "G", "C", and "N"). Here, "N" indicates the base positions that cannot be determined during the sequencing process. The sequenced reads cannot be matched to the "N" positions, which are often mistaken for CNV deletion regions. To obtain reasonable read count profiles, a binning strategy is adopted to deal with the "N" positions (Yuan et al., 2018). The read count profiles are divided into continuous and non-overlapping areas according to the bin size. The bins that contain the "N" positions are treated as abnormal bins and filtered out. The mean read count value of each bin is calculated to obtain the read depth profiles. Based on the above processing, Eq. (1) is used to deal with GC content bias (Yoon et al., 2009):

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

where RD_i and $RD_i^{'}$ represent the original and revised read depth values of the i-th bin, respectively, RD_m represents the mean

value of the read depth of all bins, and RD_{gc} represents the mean read depth value of the bins, the GC content of which is equal to that of the i-th bin. Sequencing errors and various deviations will lead to a substantial amount of noise in the read depth data, and ultimately false test results. Thus, noise reduction is a necessary step in CNV detection. The fused lasso regression method is adopted to smooth the read depth profile (Tibshirani and Wang, 2008). This method effectively considers the copy number relationship between adjacent read depth signals, which allows a reasonable read depth segment profile to be obtained. Based on the denoised read depth segment profile, Eqs. (2–5) (Li Y. et al., 2019; Liu et al., 2020) are used to transform its dimensions.

$$CN = CN_{norm} \cdot \frac{RDS_i}{RDS_m} \quad 1 \le i \le |RDS| \tag{2}$$

$$RDSR = \frac{RDS_i}{RDS_m} \quad 1 \le i \le |RDS| \tag{3}$$

$$RDSD = \begin{cases} \frac{\sum_{j=i+1}^{i+L} |RDSR_i - RDSR_j|}{L} & i = 1, 5 \le L \le 20 \\ \frac{\sum_{j=1}^{i+L} |RDSR_i - RDSR_j|}{i-1+L} & 1 < i \le L, 5 \le L \le 20 \\ \frac{\sum_{j=i-L}^{i+L} |RDSR_i - RDSR_j|}{2L} & L < i \le |RDS| - L, 5 \le L \le 20 \\ \frac{\sum_{j=i-L}^{|RDS|-1} |RDSR_i - RDSR_j|}{L+|RDS|-i-1} & |RDS| - L < i \le |RDS| - 1, 5 \le L \le 20 \\ \frac{\sum_{j=i-L}^{i-1} |RDSR_i - RDSR_j|}{L} & i = |RDS|, 5 \le L \le 20 \end{cases} \tag{4}$$

$$RDS^{'} = \{CN, RDSD\} \tag{5}$$

In Eq. (2), CN_{norm} represents the normal copy number, and its value is equal to 2. Additionally, RDS_i represents the value of the i-th read depth segment, RDS_m represents the mean across all the read depth segments, and CN represents the set of copy number, which is composed of the copy number of all read depth segments. $|RDS|$ represents the number of elements in the read depth segment set. In Eq. (3), $RDSR$ represents a set that is composed of the ratio between RDS_i and RDS_m. In Eq. (4), L represents the number of left and right neighbors of the i-th element of $RDSR$, and is set to 10 by default. $|RDSR_i - RDSR_j|$ represents the absolute value of the difference between $RDSR_i$ and $RDSR_j$, and $RDSD$ represents the set of differences of each element in $RDSR$. In Eq. (5), $RDS^{'}$ represents a two-dimensional data set, which is composed of CN and $RDSD$. This processing step provides two perspectives to observe read depth segments. The first dimension can approximately reflect the copy number status for each read depth segment, which provides a longitudinal and global perspective to observe the trend of copy number changes. The second dimension indirectly reflects the difference between a read depth segment and its surrounding read depth segments, which provides a horizontal and partial perspective to illustrate the relevance of the copy number status of each read depth segment. Moreover, this processing step provides a valid data set for the calculation of cluster-based abnormal scores, which is elaborated in the next subsection.

Calculation of Cluster-Based Abnormal Scores

Based on the $RDS^{'}$ profile, a cluster-based abnormal score is calculated for each read depth segment. Here, each element of $RDS^{'}$ is regarded as an object O. The cluster-based abnormal score is designed based on the concept of CBLOF (He et al., 2003), and is different from the traditional tumor-normal matched samples based CNV detection methods, which utilize read depth information to fit a statistical model and set a threshold to predict CNVs. The cluster-based abnormal score reflects the isolation degree of the local small cluster relative to the large cluster around it, as well as the deviation degree of each object in the large cluster relative to its cluster center, which indirectly reflects the abnormal degree of the copy number of each object. If the cluster-based abnormal score of an object is higher than those of most objects, it is likely a CNV. To further calculate the cluster-based abnormal scores, the definition is subsequently introduced in detail. First, the k-means algorithm is executed on the data set $RDS^{'}$, and can divide the data set to form clusters of different sizes. Equation (6) is used to describe the clustering result:

$$RDSC = \{RDSC_1, RDSC_2, \cdots, RDSC_{k-1}, RDSC_k\}, \tag{6}$$
$$RDSC_i \cap RDSC_j = \emptyset, 1 \le i \le k, 1 \le j \le k, i \ne j,$$

where RDSC represents a set of k clusters. Second, based on the first step, RDSC is divided into large and small clusters (He et al., 2003), as given by Eqs. (7–11).

$$RDSC^{'} = \{RDSC_1^{'}, RDSC_2^{'}, \cdots, RDSC_{k-1}^{'}, RDSC_k^{'}\} \tag{7}$$

$$|RDSC_1^{'}| + |RDSC_2^{'}| + \cdots + |RDSC_\theta^{'}| \ \ge \ |RDSC^{'}| \cdot x \tag{8}$$

$$\frac{|RDSC_\theta^{'}|}{|RDSC_{\theta+1}^{'}|} \ \ge \ \mathrm{y} \tag{9}$$

$$LRDSC^{'} = \{RDSC_i^{'} | 1 \le i \le \theta\} \tag{10}$$

$$SRDSC^{'} = \{RDSC_j^{'} | \theta < j \le k\} \tag{11}$$

$$RDSC_1^{'}| \ge \ |RDSC_2^{'}| \ \ge \cdots \ge \ |RDSC_{k-1}^{'}| \ge |RDSC_k^{'}|,$$
$$1 \le i \le k, 1 \le j \le k, i \ne j$$

In Eq. (7), $RDSC^{'}$ represents the sorted cluster set RDSC, which is sorted in descending order. In Eq. (8), $|\ ^*|$ represents the number of elements in a cluster, θ represents the boundary threshold of large and small clusters, and x represents a ratio between the total number of objects in the large cluster and the total number of objects in all clusters. The definition of the Eq. (8) is based on the consideration that most objects in $RDSC^{'}$ are not CNVs. Thus, the clusters that contain most of the objects are considered large clusters. Eq. (9) signifies that the size of a large cluster is at least y times the size of a small cluster, and describes the difference in size between the smallest large cluster and the largest small cluster. In Eq. (10), $LRDSC^{'}$

represents the set of large clusters. In Eq. (11), $SRDSC^{'}$ represents the set of small clusters. Finally, based on the preceding definitions, Eq. (12) is constructed to describe the cluster-based abnormal score.

$$CBAS(O) = \begin{cases} min(dist(O, RDSC_i^{'}))O \in RDSC_j^{'}, RDSC_i^{'} \in LRDSC^{'}, RDSC_j^{'} \in \\ SRDSC^{'}, 1 \leq i \leq \theta, \theta < j \leq k \\ dist(O, RDSC_i^{'}) \quad O \in RDSC_i^{'}, RDSC_i^{'} \in LRDSC^{'}, 1 \leq i \leq \theta \end{cases} \tag{12}$$

In Eq. (12), CBAS (O) represents the cluster-based abnormal score of object O, which is defined in two cases: (1) if the object O originates from a small cluster, the distance between O and the center of the closest large cluster is considered as the cluster-based abnormal score of O; (2) if the object O originates from a large cluster, the distance between O and the center of the large cluster is considered as the cluster-based abnormal score of O.

Predicting CNVs

Based on the cluster-based abnormal score profiles, the abnormal objects must be identified. For this step, the traditional methods analyze the abnormality of each object, and the users directly select an appropriate threshold to cut off the abnormal objects according to the application scenario. In the proposed method, Tukey's fences method is adopted to determine the abnormal objects. The prediction of abnormal objects consists of the following five steps. (1) the cluster-based abnormal scores of all objects are sorted from smallest to largest. (2) Eq. (13) is defined to evaluate an extreme outer limit:

$$T = CBAS_{Q_3} + w \cdot (CBAS_{Q_3} - CBAS_{Q_1}), \tag{13}$$

where T represents the upper limit of fences, w represents an abnormal weight, $CBAS_{Q_1}$ represents the cluster-based abnormal score of the lower quartile, and $CBAS_{Q_3}$ represents the cluster-based abnormal score of the upper quartile. (3) the basic notion of judging abnormal objects is that the higher the cluster-based abnormal score of an object, the more likely it is to be a CNV. Here, T is used as the baseline to identify abnormal objects. If the cluster-based abnormal score of an object is greater than T, it is considered to be a CNV. If the cluster-based abnormal score of an object is less than or equal to T, it is considered to be a normal area. (4) after the candidate CNVs are determined, their mutation modes (gain or loss) are determined. If the read depth value of a CNV area is greater than or equal to the mean read depth value of all normal areas, it is considered to be a gain area. If the read depth value of a CNV area is less than the mean read depth value of all normal areas, it is considered to be a loss area. (5) finally, somatic CNVs and germline CNVs are further identified. A germline CNV is a genetic variation that may originate from an individual's parents or family. If a CNV exists in both the tumor-normal matched samples, it is regarded as a germline CNV.

Parameter Setting of CBCNV

To effectively use CBCNV, it is necessary to further explain the settings of related parameters, which include the bin size, the number of neighbors (L), the number of clusters (k), and the ratios of large clusters (x), multiples (y), and abnormal weight (w). In this study, the bin size and L are set to 2,000 bp and 10 by default, respectively. Additionally, the values of k, x, and y are set to 5, 0.9, and 5, respectively, which are adopted by referencing published article (He et al., 2003). In Tukey's fences method, w is generally set to 1.5 (Zijlstra et al., 2007). In the proposed method, w is set to 1.5 as the default value. The settings of these default parameter values in the proposed method were determined according to experience and related methods. Users can also adjust these parameters according to their actual needs and application scenarios.

RESULTS

It is necessary to design a reasonable experimental plan to verify the effectiveness and reliability of the proposed method. Aiming at this point, simulation and real experiments were conducted. A simulation experiment is an effective and objective evaluation strategy, which can provide a comparison criterion to quantify the performance of the proposed method. In the simulation experiment, three popular published algorithms (BIC-seq2 (Xi et al., 2016), SeqCNV (Chen et al., 2017), and CNVkit (Talevich et al., 2016)) that can be used to effectively detect tumor-normal matched samples were selected for comparison with CBCNV. The performances of these methods are evaluated from three perspectives. First, the sensitivity and false discovery rate (FDR) of the four methods are evaluated at six CNV length levels. Then, the sensitivity and FDR of each method in the CNV gain and loss regions are analyzed and discussed. Finally, three indicators (recall, precision, and F1-score) are used to comprehensively evaluate the performance of each method. In real data applications, the proposed algorithm was used to detect two pairs of matched breast cancer whole-genome sequencing samples. Because the ground truths of the real data sets are unknown, the number of overlapping events and number of predicted events are adopted to evaluate the performance of each method. To further verify the performance of the proposed method, we use overlapping density score method to quantify performance of each method. The experimental results demonstrate that CBCNV is powerful CNV detection tools.

Application of Simulation Data

Many CNV simulation softwares have been developed and applied to generate next-generation sequencing data. In this study, IntSIM software was selected to generate simulation data sets (Yuan et al., 2017). Before its use, some settings were conducted: (1) the reference genome was prepared; (2) the tumor purity (TP) and sequencing coverage (SC) were set; (3) the number of repetitions of the sample under the configuration of each group was selected. Chromosome 21 of hg19 was entered into the software as a reference genome. The tumor purity was set to 0.2 and 0.3, and sequencing coverage was set to 10× to

generate simulated data sets of different configurations, in which 50 samples were generated. Each sample was embedded with 22 regions of variation, which were composed of 12 gains and 10 losses (four heterogeneous losses and six homogeneous losses). The length of the CNV regions ranged from 2 to 100 kb. To fairly evaluate the performance of each method, the default parameters were used for all methods to detect each set of data.

Figure 2 describes the sensitivity of the four methods for the respective detection of CNVs with lengths of 2, 6, 10, 30, 50, and 100 kb under two different configurations, respectively. Two performance indicators (sensitivity and FDR) are adopted to evaluate the resolution of each method. Sensitivity is defined as the value of the number of CNVs correctly detected by a tool divided by the total number of CNVs recorded by the ground truth file. FDR is defined as the value of the number of false positives detected by a tool divided by the total number of CNVs detected by the tool. If a detected event overlaps with the ground truth file by more than 50%, it is considered as a candidate CNV (Hormozdiari et al., 2009). From the figure, it is evident that the sensitivity of each method increased with the increase in tumor purity from 0.2 to 0.3. This demonstrates that tumor purity is one of the key factors that affect CNV detection. In contrast, long CNVs were more easily detected by each method than short CNVs. CBCNV achieved the best sensitivity for all CNV length levels, and BIC-seq2 achieved better sensitivity than the other two methods (SeqCNV and CNVkit) at most CNV length levels. SeqCNV achieved the lowest sensitivity in the cases of CNVs with lengths of 50 and 100 kb, which indicates that it is not sensitive enough to detect long CNVs. CNVkit achieved the lowest sensitivity in the cases of CNVs with lengths of 2 and 6 kb, which indicates that it is not sensitive enough to detect short CNVs. **Figure 3** presents the FDR of each method at the six CNV length levels under two different configurations. In the case of tumor purity = 0.2, CNVkit performed the best in terms of FDR, followed by CBCNV, BIC-seq2 and SeqCNV. Although CNVkit achieved the best FDR, it had the lowest sensitivity. In the case of tumor purity = 0.3, CBCNV performed excellently in terms of FDR, followed by BIC-seq2, CNVkit, and SeqCNV. Considering the two indicators together, CBCNV achieved the best tradeoff between sensitivity and FDR, followed by BIC-seq2, SeqCNV, and CNVkit. Via the preceding analysis and discussion, it can be concluded that CBCNV can detect more CNVs with fewer false positives than the other three methods.

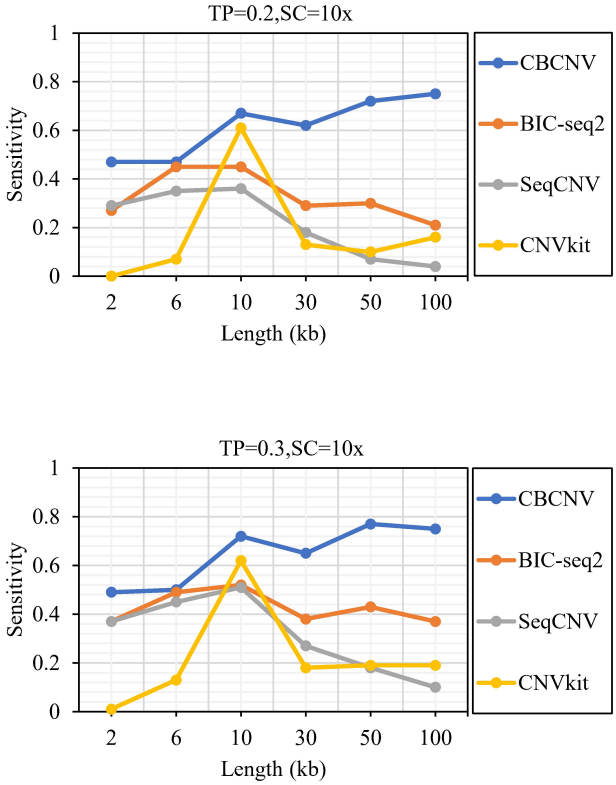

FIGURE 2 | The sensitivity of four methods at the six CNV length levels under two sets of simulation configurations.

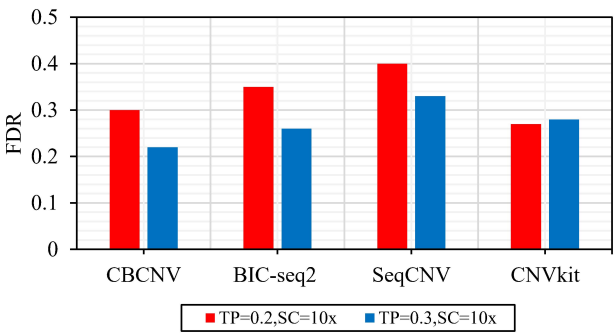

FIGURE 3 | The FDR of each method at the six CNV length levels under two sets of simulation samples.

Based on the simulated data sets, sensitivity and FDR were considered to analyze and evaluate the performances of the compared methods (CBCNV, BIC-seq2, SeqCNV, and CNVkit) in the gain and loss areas, and the averages of the two indicators were calculated across the 50 samples under different setting conditions. In general, the sensitivity of each method was found to increase with the increase in tumor purity, which demonstrates that the performance of each method was very sensitive to tumor purity. Most methods detected the CNV gain areas more sensitively than the CNV loss areas. **Figure 4** describes the sensitivity of each method to the detection of the gain and loss areas under two different sets of conditions. In each set of conditions, CBCNV achieved the highest sensitivity in the gain and loss areas. BIC-seq2 achieved better sensitivity in the gain areas than the other two methods (SeqCNV and CNVkit), and its sensitivity in the loss areas ranked third. The sensitivity of SeqCNV to the detection of the gain areas was between those of BIC-seq2 and CNVkit, and it was insensitive to the detection of the loss areas as compared to the other three methods. CNVkit achieved the lowest sensitivity in the gain areas, but its sensitivity ranked second in the loss areas, which indicates that it is suitable for detecting loss areas. **Figure 5** describes the FDR of each method in the detection of gain and loss areas under two different sets of conditions. When tumor purity was

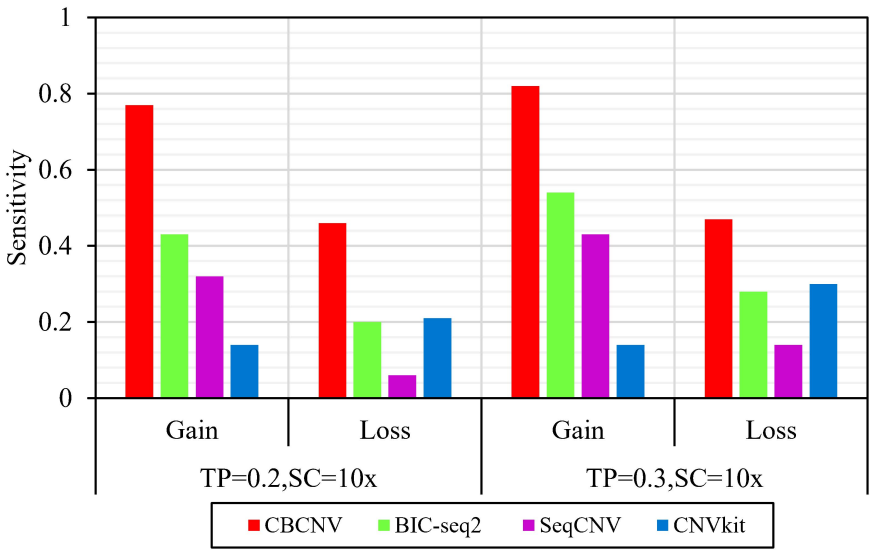

FIGURE 4 | Comparison of the sensitivity of the four methods for detecting gain and loss areas under two sets of simulation settings.

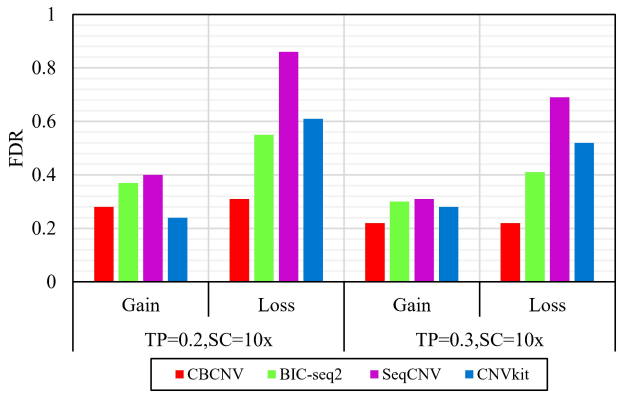

FIGURE 5 | Comparison of the FDR of each method for detecting gain and loss areas under two sets of simulation samples.

equal to 0.2 and 0.3, the FDR of CBCNV ranked second and first in the gain areas, and ranked first in the loss areas. The FDR of BIC-seq2 ranked third and second in the gain and loss areas, respectively. SeqCNV exhibited the largest FDRs in the gain and loss areas, which demonstrates that there were many false positives of the detected CNVs. CNVkit had a medium FDR between those of BIC-seq2 and SeqCNV in the loss areas. In the gain areas, the FDR of CNVkit ranked first when tumor purity = 0.2, and second when tumor purity = 0.3. CNVkit detected the gain areas more accurately than the loss areas. This demonstrates that the DOC-based method detected the gain areas more sensitively than the loss areas. In summary, CBCNV achieved the best tradeoff between sensitivity and FDR in the detection of gain and loss areas.

Three indicators (recall, precision, and F1-score) were adopted to comprehensively evaluate the performance of each method. Recall is defined as the number of correctly detected CNVs divided by the total number of simulated CNVs (Magi et al., 2013). Precision is defined as the number of correctly detected CNVs divided by the total number of detected CNVs (Magi et al., 2013). The F1-score represents the harmonic mean of precision and recall. The three performance indicators are reported as the averages of 50 samples under each set of conditions. **Figure 6** detail the F1-score level of each method, from which it is evident that CBCNV achieved the highest recall, followed by BIC-seq2, SeqCNV, and CNVkit. When tumor purity = 0.3, CBCNV got the best precision rate among all methods. When tumor purity = 0.2, CNVkit performed the best in terms of precision, followed by CBCNV, BIC-seq2, and SeqCNV. Moreover, CBCNV achieved the best tradeoff between precision and recall, followed by BIC-seq2, SeqCNV, and CNVkit, which is consistent with the above experimental results.

Detection of Copy Number Variants From Breast Cancer Samples

To analyze and verify the performance of the proposed method, it was applied to detect four paired whole-genome breast cancer samples (PD4088a, PD4088b, PD4192a, and PD4192b), the details of which were sourced from https://www.ebi.ac.uk/ega/studies under accession EGAS00001000170 (Li Y. Y. et al., 2019). CBCNV was used to detect 22 autosomes in each set of samples, and two well-known methods (BIC-seq2 and CNVkit) were selected for comparison. For a fair comparison, the default parameters were used for these methods to detect the samples. The number of overlapping events and predicted events were used for performance measurement to effectively analyze the advantages and disadvantages of each method. The ground truth file could not be provided in the real data experiment. The number of overlapping events represent the average number of overlapping events for one method and other methods, and

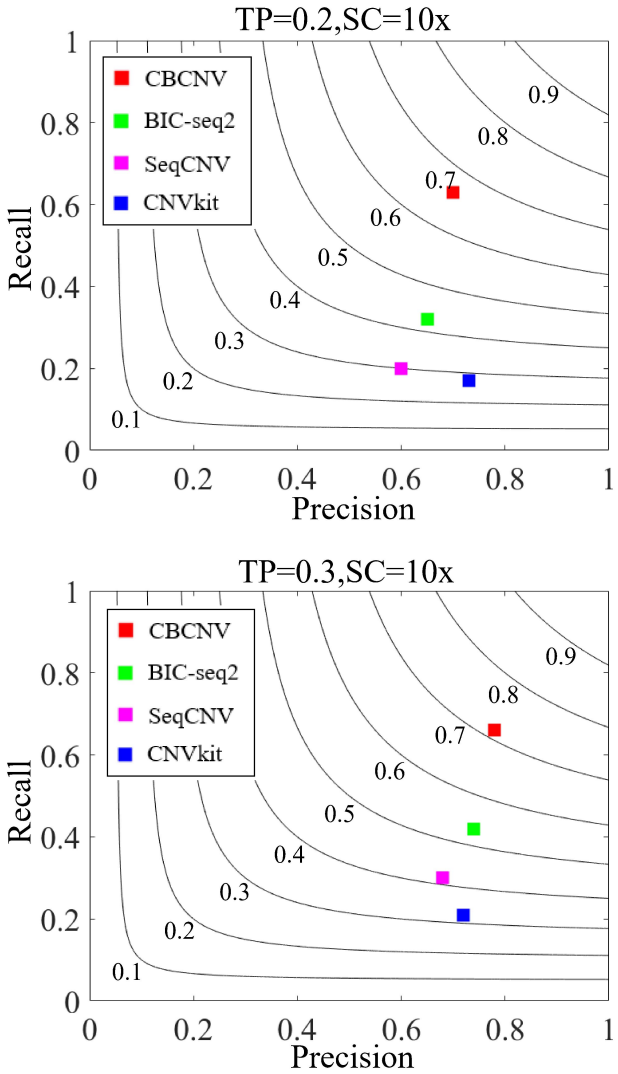

FIGURE 6 | The performance comparison of the four methods in terms of recall, precision and F1-score under two sets of simulation settings. Black curves indicate F1-score levels, which ranges from 0.1 to 0.9. The equations on the left and right sides of the comma represent the tumor purity (TP) and sequencing coverage (SC), respectively.

number of predicted events represents the total number of events predicted by a method. **Table 1** presents the number of overlapping events and predicted events of each method in the 22 autosomes of samples PD4088a and PD4192a, respectively. In the sample PD4088a, it is evident that CBCNV achieved the greatest number of overlapping events and predicted events. BIC-seq2 detects the least number of overlapping events and predicted events, which shows that it is more conservative than the other two methods. CNVkit achieved number of overlapping events and predicted events between CBCNV and BIC-seq2. In the sample PD4192a, CNVkit called a large number of CNV events, but obtained number of overlapping events as many as CBCNV, which means it has detected a large number of non-overlapping events. It indirectly shows that the CNVs detected by CNVkit are likely to contain a large number of false positive events. A small number of overlapping events and predicted events were found by BIC-seq2, performance of which is consistent with the above sample. The number of events detected by CBCNV is between the other two comparison methods, but it obtains the most overlapping events, which fully shows that most of the CNV events detected by CBCNV are true positive events.

In order to further verify the performance of each method, we adopt the evaluation method of overlapping density score (ODS) (Yuan et al., 2018), which is defined by the following equation.

$$ODS = O_m \cdot O_r, \tag{14}$$

Where O_m represents the mean number of overlapping events between one method and other comparison methods, O_r represents O_m divided by the total number of CNV events detected by the method. Here, we use Eq. (14) to calculate ODS for each method, and the comparison results are recorded in **Table 2**. From the experimental results, CBCNV achieve the best ODS in the all samples. ODS of BIC-seq2 are the lowest among all methods. Compared with the above two methods, CNVkit obtain the medium ODS in each group of samples. Overall, CBCNV achieved the best balance between O_m and O_r as compared to the other two methods.

On the basis of the above experiments, we used the catalog of somatic mutations in cancer (COSMIC) database to analyze the biological significance of the detected CNVs. From two pairs of matched breast cancer samples, we found that some of the detected CNVs contained some genes that were related to breast cancer, such as PDZK1 (Kim et al., 2013), XRCC4 (Allen-Brady et al., 2006), Fbxl17 (Mason et al., 2020), ITGBL1 (Li et al., 2015), RORA (Taheri et al., 2017), BAGE (Fujie et al., 1997), AMOTL1 (Couderc et al., 2016), RAP80 (Osorio et al., 2009), PIWIL4 (Wang et al., 2016), CSE1L (Behrens et al., 2001), and USP18 (Tan et al., 2018).

Evaluation of Running Time

Running time is a critical evaluation indicator to evaluate the performance of the methods. For this, CBCNV and the other three methods (BIC-seq2, SeqCNV, and CNVkit) are tested on 50 simulation samples, which are run on a personal computer with

TABLE 1 | Comparison of number of overlapping events (NOE) and predicted events (NPE) for each method on two sets of real samples.

Sample	Indicator	CBCNV	BIC-seq2	CNVkit
PD4088a	NOE	80	19	49
	NPE	510	85	194
PD4192a	NOE	126	20	126
	NPE	482	83	2,156

TABLE 2 | Comparison of ODS for each method on two sets of real samples.

Sample	CBCNV	BIC-seq2	CNVkit
PD4088a	19	6	18
PD4192a	43	7	32

TABLE 3 | Comparison of running time for each method.

Indicator	CBCNV	BIC-seq2	SeqCNV	CNVkit
Running time (s)	39	8	500	182

Intel(R) Core (TM) i7-4710MQ CPU @ 2.50 GHz and 16.0 GB memory. The running time of each method is counted as the averages of 50 simulation samples. As shown in **Table 3**, BIC-seq2 performed the best in terms of running time, followed by CBCNV, CNVkit, and SeqCNV, which shows that CBCNV is a relatively efficient CNV detection tool.

DISCUSSION AND CONCLUSION

In this work, the proposed CBCNV method was developed based on DOC profiles to detect CNVs using next-generation sequencing data, and is suitable for the detection of tumor-normal matched samples. CBCNV uses a local perspective to capture abnormal read depth signals, which are considered to be only a small portion of the overall signals. Its detection concept is different from those of traditional CNV detection methods, which generally construct a statistical model by fitting the read depth signals, then select a reasonable baseline to identify CNVs. Instead, in CBCNV, a clustering algorithm is performed on the read depth segment profile to form clusters of different scales. According to the scales of the clusters, large and small clusters are defined. If a read depth segment originates from a large cluster, its abnormal score is defined as the distance between the read depth segment and the cluster center. If a read depth segment belongs to a small cluster, its abnormal score is defined as the distance between the read depth segment and the center of the closest large cluster. In this way, an abnormal score is assigned to each read depth segment. Based on the abnormal score profile, Tukey's fences method is adopted to predict CNVs (Zijlstra et al., 2007).

Via the analysis of the concepts of the proposed method, the following characteristics are summarized. (1) CBCNV extracts two features of read depth signals, which fully considers the copy number of each read depth segment and the difference in the ratios of adjacent read depth segments. (2) The traditional outlier detection algorithm was effectively converted to detect CNVs. CBCNV uses a local perspective to identify CNVs, which can objectively reflect the actual state of abnormal read depth signals. It does not require the fitting of the distribution of read depth signals, and cluster-based abnormal scores are constructed for each read depth segment signal to effectively identify the copy number status of adjacent read depth signals. (3) Based on the abnormal score of each read depth segment, Tukey's fences method is applied to identify CNVs, which does not require the evaluation of the distribution of abnormal scores.

Simulated data sets were used to evaluate the performance of CBCNV, and three popular algorithms were selected for comparison. First, the sensitivity and FDR of each method for the detection of CNVs of different lengths and in gain and loss regions were analyzed and discussed. Via the analysis of the experimental results, it was found that CBCNV achieved the best tradeoff between sensitivity and FDR. Second, three performance indicators (recall, precision, and F1-score) were adopted to comprehensively evaluate the performance of each method. The experimental results proved that CBCNV achieved the best performance in terms of all three indicators. In real data applications, two sets of whole-genome data were used to evaluate the effectiveness of the proposed method. The experimental results demonstrated that CBCNV achieved the best number of overlapping events and overlapping density scores compared to the other two methods in each group of samples. In summary, CBCNV is an effective and reliable CNV detection tool for using on tumor-normal matched samples.

Some shortcomings of the proposed method were also discovered. For example, the selection of the number of clusters (k) is a very important step that may affect the accuracy of the results. In most application scenarios, the performance of the proposed method was superior under this set of parameter settings, which meets the needs of users in most cases. However, in some unique cases, the performance of this set of parameters may not be suitable. In future research, the data size and characteristics will be fully considered to automatically set the parameter k. In addition, in the present study, only two features of read depth were extracted as the input. In future research, multiple factors of read depth signals will be considered to improve the accuracy of the proposed method. Ultimately, CBCNV will be further expanded (Mao et al., 2021) and proved to effectively detect other types of structural variation in multiple application scenarios.

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 below: https://www.ebi.ac.uk/ega/studies, EGAS00001000170.

AUTHOR CONTRIBUTIONS

GL participated in the design of the algorithms and the experiments. JZ participated in the design of the entire framework of CNV detection and directed the whole work, and helped to revise the manuscript. Both 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 Grants 91853123 and 11674352.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.699510/full#supplementary-material

REFERENCES

Adam, S., and David, M. (2009). Copy number variations and cancer. *Genome Med.* 1, 62. doi: 10.1186/gm62

Allen-Brady, K., Cannon-Albright, L., Neuhausen, S., and Camp, N. (2006). A role for XRCC4 in age at diagnosis and breast cancer risk. *Cancer Epidemiol. Biomarkers Prevent.* 15, 1306–1310. doi: 10.1158/1055-9965.EPI-05-0959

Behrens, P., Brinkmann, U., Fogt, F., Wernert, N., and Wellmann, A. (2001). Implication of the proliferation and apoptosis associated CSE1L/CAS gene for breast cancer development. *Anticancer Res.* 21, 2413–2417.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899–905. doi: 10.1038/nature08822

Buysse, K., Chiaie, B. D., Coster, R. V., Loeys, B., Paepe, A. D., Mortier, G., et al. (2009). Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *Eur. J. Med. Genet.* 52, 398–403. doi: 10.1016/j.ejmg.2009.09.002

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* 39, S16–S21. doi: 10.1038/ng2028

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 Bioinform.* 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

Cook, E., and Scherer, S. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature* 455, 919–923. doi: 10.1038/nature07458

Couderc, C., Boin, A., Fuhrmann, L., Vincent-Salomon, A., Mandati, V., Kieffer, Y., et al. (2016). AMOTL1 promotes breast cancer progression and is antagonized by merlin. *Neoplasia* 18, 10–24. doi: 10.1016/j.neo.2015.11.010

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

Fujie, T., Mori, M., Ueo, H., Sugimachi, K., and Akiyoshi, T. (1997). Expression of MAGE and BAGE genes in Japanese breast cancers. *Ann. Oncol.* 8, 369–372. doi: 10.1023/A:1008255630202

Hartigan, J. A., and Wong, M. A. (1979). A K-Means clustering algorithm. *J. R. Stat. Soc.* 28, 100–108. doi: 10.2307/2346830

He, Z. Y., Xu, X. F., and Deng, S. C. (2003). Discovering cluster-based local outliers. *Pattern Recognition Lett.* 24, 1641–1650. doi: 10.1016/S0167-8655(03)00003-5

Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* 19, 1270–1278. doi: 10.1101/gr.088633.108

Kim, H., Abd Elmageed, Z., Ju, J., Naura, A., Abdel-Mageed, A., Varughese, S., et al. (2013). PDZK1 is a novel factor in breast cancer that is indirectly regulated by Estrogen through IGF-1R and promotes estrogen-mediated growth. *Mol. Med.* 19, 253–262. doi: 10.2119/molmed.2011.00001

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318, 420–426. doi: 10.1126/science.1149504

Krepischi, A., Pearson, P. L., and Rosenberg, C. (2012). Germline copy number variations and cancer predisposition. *Future Oncol.* 8, 441–450. doi: 10.2217/fon.12.34

Kuiper, R. P., Ligtenberg, M. J., Hoogerbrugge, N., and Kessel, A. G. V. (2010). Germline copy number variation and cancer risk. *Curr. Opin. Genet. Dev.* 20, 282–289. doi: 10.1016/j.gde.2010.03.005

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, X.-Q., Du, X., Li, D.-M., Kong, P.-Z., Sun, Y., Liu, P.-F., et al. (2015). ITGBL1 Is a Runx2 transcriptional target and promotes breast cancer bone metastasis by activating the TGFβ signaling pathway. *Cancer Res.* 75, 3302–3313. doi: 10.1158/0008-5472.CAN-15-0240

Li, Y., Yuan, X., Zhang, J., Yang, L., Bai, J., and Jiang, S. (2019). SM-RCNV: a statistical method to detect recurrent copy number variations in sequenced samples. *Genes Genomics* 41, 529–536. doi: 10.1007/s13258-019-00788-9

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

Liu, G. J., Zhang, J. Y., Yuan, X. G., 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

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

Magi, A., Tattini, L., Cifola, I., D'Aurizio, R., Benelli, M., Mangano, E., et al. (2013). EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* 14:R120. doi: 10.1186/gb-2013-14-10-r120

Malek, J. A., Mery, E., Mahmoud, Y. A., Al-Azwani, E. K., Roger, L., Huang, R., et al. (2011). Copy number variation analysis of matched ovarian primary tumors and peritoneal metastasis. *PLoS One* 6:e28561. doi: 10.1371/journal.pone.0028561

Mao, Y. F., Yuan, X. G., and Cun, Y. P. (2021). A novel machine learning approach (svmSomatic) to distinguish somatic and germline mutations using next-generation sequencing data. *Zool. Res.* 42, 246–249. doi: 10.24272/j.issn.2095-8137.2021.014

Mason, B., Flach, S., Teixeira, F., Garcia, R., Rueda, O., Abraham, J., et al. (2020). Fbxl17 is rearranged in breast cancer and loss of its activity leads to increased globalO-GlcNAcylation. *Cell. Mol. Life Sci.* 77, 2605–2620. doi: 10.1007/s00018-019-03306-y

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* 6, S13–S20. doi: 10.1038/nmeth.1374

Osorio, A., Barroso, A., Garcia, M., Martinez-Delgado, B., Urioste, M., and Benitez, J. (2009). Evaluation of the BRCA1 interacting genes RAP80 and CCDC98 in familial breast cancer susceptibility. *Breast Cancer Res. Treatment* 113, 371–376. doi: 10.1007/s10549-008-9933-4

Pei, G., Hu, R., Dai, Y., Manuel, A., Zhao, Z., and Jia, P. (2021a). Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations. *Nucleic Acids Res.* 49, 53–66. doi: 10.1093/nar/gkaa1137

Pei, G., Hu, R., Jia, P., and Zhao, Z. (2021b). DeepFun: a deep learning sequence-based model to decipher non-coding variant effect in a tissue- and cell type-specific manner. *Nucleic Acids Res. [online ahead of print]* gkab429. doi: 10.1093/nar/gkab429

Pei, G., Hu, R., Dai, Y., Zhao, Z., and Jia, P. (2020). Decoding whole-genome mutational signatures in 37 human pan-cancers by denoising sparse autoencoder neural network. *Oncogene* 39, 5031–5041. doi: 10.1038/s41388-020-1343-z

Sebat, J., Lakshmi, B., Malhotra, D., and Troge, J. (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445–449. doi: 10.1126/science.1138659

Sharp, A. J., Locke, D. P., McGrath, S. D., and Cheng, Z. (2005). Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* 77, 78–88. doi: 10.1086/431652

Stone, J. L., O'Donovan, M. C., Gurling, H., and Kirov, G. K. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455, 237–241. doi: 10.1038/nature07239

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature* 458, 719–724. doi: 10.1038/nature07943

Taheri, M., Omrani, M. D., Noroozi, R., Ghafouri-Fard, S., and Sayad, A. (2017). Retinoic acid-related orphan receptor alpha (RORA) variants and risk of breast cancer. *Breast Dis.* 37, 21–25. doi: 10.3233/BD-160248

Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016). CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* 12:e1004873. doi: 10.1371/journal.pcbi.1004873

Tan, Y., Zhou, G., Wang, X., Chen, W., and Gao, H. (2018). USP18 promotes breast cancer growth by upregulating EGFR and activating the AKT/Skp2 pathway. *Int. J. Oncol.* 53, 371–383. doi: 10.3892/ijo.2018.4387

Tchatchou, S., and Burwinkel, B. (2008). Chromosome copy number variation and breast cancer risk. *Cytogenetic Genome Res.* 123, 183–187. doi: 10.1159/000184707

Tibshirani, R., and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* 9, 18–29. doi: 10.1093/biostatistics/kxm013

Wang, Z., Liu, N., Shi, S., Liu, S., and Lin, H. (2016). The role of PIWIL4, an argonaute family protein, in breast cancer. *J. Biol. Chem.* 291, 10646–10658. doi: 10.1074/jbc.M116.723239

Xi, R. B., Lee, S., Xia, Y. C., Kim, T. M., and Park, P. J. (2016). Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.* 44, 6274–6286. doi: 10.1093/nar/gkw491

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. G., Bai, J., Zhang, J. Y., 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, 1141–1153. doi: 10.1109/TCBB.2018.2883333

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

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). PyOD: a Python toolbox for scalable outlier detection. *J. Machine Learn. Res.* 20:96.

Zijlstra, W. P., van der Ark, L. A., and Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behav. Res.* 42, 531–555. doi: 10.1080/00273170701384340

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.