Hindawi
BioMed Research International
Volume 2019, Article ID 8420547, 8 pages
https://doi.org/10.1155/2019/8420547

Research Article

Detection of False-Positive Deletions from the Database of Genomic Variants

Junbo Duan,[1] Han Liu,[1] Lanling Zhao,[1] Xiguo Yuan,[2] Yu-Ping Wang,[3] and Mingxi Wan[1]

[1] *Department of Biomedical Engineering, Xi'an Jiaotong University, Xi'an, China*
[2] *School of Computer Science and Technology, Xidian University, Xi'an, China*
[3] *Department of Biomedical Engineering, Tulane University, New Orleans, USA*

Correspondence should be addressed to Junbo Duan; junbo.duan@mail.xjtu.edu.cn

Received 15 August 2018; Revised 24 February 2019; Accepted 4 March 2019; Published 4 April 2019

Academic Editor: Marco Fichera

Next generation sequencing is an emerging technology that has been widely used in the detection of genomic variants. However, since its depth of coverage, a main signature used for variant calling, is affected greatly by biases such as GC content and mappability, some callings are false positives. In this study, we utilized paired-end read mapping, another signature that is not affected by the aforementioned biases, to detect false-positive deletions in the database of genomic variants. We first identified 1923 suspicious variants that may be false positives and then conducted validation studies on each suspicious variant, which detected 583 false-positive deletions. Finally we analysed the distribution of these false positives by chromosome, sample, and size. Hopefully, incorrect documentation and annotations in downstream studies can be avoided by correcting these false positives in public repositories.

1. Introduction

A genomic variant is an alteration of the DNA sequence of an organism. Since an organism's DNA sequence encodes the genetic instructions used in its development, any alteration of this sequence may cause genetic abnormalities or even fatality. According to their sizes, genomic variants are classified into small-scale variants, such as single nucleotide polymorphisms (SNP) and indels, and large-scale variants, namely, structural variations (SV), including copy number variations (CNV), insertions, deletions, inversions, segmental duplications, and translocations [1]. Various complex diseases have been reported to be associated with genomic variants in human genomes [2].

Prior to next generation sequencing (NGS), cytogenetic techniques, such as fluorescence *in situ* hybridization (FISH) and array comparative genomic hybridization (aCGH), were employed to detect SV. However, due to their relatively low genomic resolution (e.g., microscopic scale (Mbp) for FISH, and submicroscopic scale (kbp) for aCGH) [3], most medical and biological research teams have migrated their platforms to NGS, which can provide base-pair level resolution.

Several approaches have been proposed to detect SV from NGS data. Generally, these approaches can be classified into two categories: paired-end read mapping (PEM)-based or depth of coverage (DOC)-based approaches [4]. For PEM-based approaches, if the span of a pair of mapped reads is longer/shorter than a specified cutoff related to the insert size of the sequencing library, a deletion/insertion can be identified [5], whereas for DOC-based approaches, if the local depth of reads is significantly larger/smaller than the global DOC, a duplication/deletion can be identified [3, 6, 7]. PEM-based approaches have advantage of detecting balanced SVs (inversion) and unbalanced SVs (deletion and insertion) of relatively small sizes, whereas DOC-based approaches are good at detecting unbalanced SVs (CNV) of relatively large sizes. Besides PEM and DOC, several other supplementary signatures, such as split read mapping, have been combined into PEM and DOC to improve detection performance, leading to integrative models [8–10].

Despite PEM's advantage in balanced SV detection, the majority of detection approaches use DOC as the primary signature to identify CNVs [11]. However, the DOC signature is biased due to two main factors: GC content [12] and mappability [13].

(i) Since G and C form a triple hydrogen bond (whereas A and T form a double bound), theoretically the melting temperature of GC-rich segments is approximately 2°C higher than that of AT-rich segments [14]. As a result, when the sequencing protocol involves polymerase chain reaction (PCR), GC-rich segments and AT-rich ones are unevenly amplified [15], yielding the correlation between DOC and GC content [16, 17].

(ii) Due to the complexity of the human genome, there are regions in which sequenced short reads cannot be uniquely mapped, e.g., repeated regions, such as retrotransposons (LINE and SINE) [13]. Mappability was introduced to measure the uniqueness of such regions using a score that ranges from 0 to 1, corresponding to highly repeated and unique regions, respectively. From the definition, it is clear that the DOC is correlated with mappability.

In human genomes, both GC content and mappability are distributed unevenly along chromosomes, and, therefore, they introduce biases into DOC. Several methods have been developed to correct these two biases [13, 18, 19].

However, due to the overlook of biases introduced by GC content and mappability, some DOC-based SV studies contain false detection. From the mechanism of PEM and DOC signatures, it is clear that PEM is less affected by GC content or mappability than DOC, and, therefore, PEM can be used to detect false positives. In this paper, we used this idea to verify the entries in the database of genomic variants (DGV). We hope that incorrect documentation and annotations can be avoided in downstream studies by correcting the false positives in this public repository and other related ones such as EMBL-EBI's Genomic Variants archive (DGVa) and NCBI's dbVar.

2. Materials and Methods

2.1. Samples and Data. The database of genomic variants (DGV, http://dgv.tcag.ca/dgv/app/home) [20] provides a comprehensive summary of structural variation (SV) in the human genome. In DGV, SVs are defined as genomic alterations that involve segments of DNA with length larger than 50 bp. In DGV, the 6.4 millions of variants represent collections from 55 thousand healthy control samples in 72 studies. DGV provides a curated catalogue of genomic variations in the human genome, which was integrated into EMBL-EBI's Genomic Variants archive (DGVa) and NCBI's dbVar. Therefore, this database is of tremendous importance to investigators whose study interest is about genomic variance, which is also the focus of the current study.

The 1000 Genomes Project (http://www.1000genomes.org/ [21]) is a well-known international collaborative NGS project, which aims to sequence the genomes of approximately 2500 people from 25 populations around the world. In our study, the BGZF compressed sequence alignment/map (SAM) data files (BAM) of most samples were downloaded from the website of this project as the primary dataset.

Sequence read archive (SRA, https://www.ncbi.nlm.nih.gov/sra) is the NCBI database that stores raw sequencing data obtained from NGS technology. The aim of SRA is to make NGS data available to researchers to both improve reproducibility and enable new discoveries. This database includes data from most common sequencing platforms and most NGS studies. The BAM file of some specific samples in our study was not available, so the SRA files were downloaded as the primary dataset.

2.2. Methods. The steps conducted in the study are shown as follows, and the pseudocode is listed in Pseudocode 1 to illustrate the logical structure of these steps.

(1) The latest spreadsheet of supporting variants was downloaded from the DGV website. The GC content and mappability profile of each chromosome of hg18 were downloaded from the readDepth website (https://github.com/chrisamiller/readdepth) [18], which is an NGS-based CNV detection software package.

(2) GC content profiles were smoothed with LOESS [22], and segments with size larger than 500 bp and an average GC content lower than $th_1 = 0.26$ or greater than $th_2 = 0.59$ were obtained as suspicious regions.

(3) Mappability profiles were smoothed with LOESS, and segments with size larger than 500 bp and an average mappability lower than $th_3 = 0.92$ were added to suspicious regions.

(4) All supporting variants in DGV were resolved one-by-one to collect the fields needed in the current study, including variant accession ID, chromosome, genomic location (starting and ending loci), variant subtype, reference, method, and samples used.

(5) Variants associated with the *sequencing* method and *loss* or *deletion* subtype, size smaller than 10 kbp, and nonempty sample fields were filtered for further analysis.

(6) Duplicated variants were merged.

(7) GDV variants that overlapped (F-score greater than 0.9) with suspicious regions were identified as suspicious variants.

(8) For each suspicious variant, the corresponding BAM file was downloaded. If no BAM file was available, the corresponding SRA file was downloaded, aligned with BWA, compressed, and sorted with SAMtools to obtain a BAM file.

(9) For each suspicious variant, both the PEM and DOC signatures were extracted from the corresponding BAM file.

(10) The PEM signature was used to verify whether this suspicious variant was a false-positive or true variant (see Section 2.3).

(11) Finally, the GC content, mappability, DOC, and PEM profiles of each false positive were displayed in an individual figure for visual inspection, and information of all suspicious variants was outputted to a spreadsheet.

```
INPUTS: DGV.xls, chr1~22.gc, chr1~22.map \\step 1
segments_gc = segmenting(smoothing(chr1~22.gc))
FOR i IN 1 TO number_of_items_in_segments_gc
        IF (segments_gc[i].value < th1 OR segments_gc[i].value > th2) AND size(segments_gc[i].loci) > 500
                regions_suspicious=regions_suspicious ∪ segments_gc[i].loci \\step 2
        END
END
segments_map=segmenting(smoothing(chr1~22.map))
FOR i IN 1 TO number_of_items_in_segments_map
        IF segments_map[i].value < th3 AND size(segments_map[i].loci) > 500
                regions_suspicious = regions_suspicious ∪ segments_map[i] \\step 3
        END
END
FOR i IN 1 TO number_of_items_in_DGV.xls
        variant_supporting = {DGV[i].ID, DGV[i].chr, DGV[i].loci, DGV[i].subtype, DGV[i].ref,
                                   DGV[i].method, DGV[i].samples} \\step 4
        IF variant_supporting.method = 'sequencing' AND variant_supporting.sub = ('loss' OR 'deletion')
           AND size(variant_supporting.loci) <10000 AND variant_supporting.sample ≠ empty \\step 5
           AND (∃ j such that F-score(variant_supporting.loci, regions_suspicious[j].loci) > 0.9 \\step 7
                        IF variant_supporting is a duplicated items (chr and loci fields are same)
                                   ID, ref, and sample fields are merged to the existing one \\step 6
                        ELSE variants_suspicious=variants_suspicious ∪ variant_supporting \\step 7
                        END
        END
END
list_samples = ⋃variants_suspicious.samples
\\ download sequencing data of samples listed in list_samples, and preprocess to get BAM files. \\step 8
FOR i IN 1 TO number_of_items_in_variants_suspicious
        calculate PEM[i] and DOC[i] from BAM file(s) of variants_suspicious[i].samples \\step 9
        IF PEM[i] meets false detection criterion \\step 10
                        variant_supporting.false = T; plot figure
        ELSE variant_supporting.false = F
        END
END
OUTPUTS: variants_suspicious, ∗.figure \\step 11
```

Pseudocode 1: Pseudocode of processing pipeline.

Here are some notes that should be addressed:

(i) A *supporting variant* represents a variant called in a single sample/individual, which can also be described as sample level variant [20].

(ii) One sample is the minimal requirement to verify a specific supporting variant, so the sample field should be nonempty.

(iii) The terms *deletion* and *loss* are equivalent in the database [20].

(iv) Duplicated variants are defined in the sense of the same chromosome ID and genomic location.

(v) The *corresponding* BAM or SRA file of a suspicious variant was retrieved according to the *sample* and *reference* field.

2.3. Validation. The validation of variants is based on the PEM signature. First, for each suspicious variant, we extracted all the read pairs in which both ends were mapped within the region of interest (ROI) from the corresponding BAM file. To provide an adaptive zoom, the ROI is defined as the genomic region that extends both upstream and downstream with 1 kbp plus half of the variant length.

Next, the F-score [23] is employed to measure the overlapping quality between the span of a suspicious variant and that of a mapped read pair. The F-scores quantify the overlap quality between two spans, with values ranging from 0 to 1 (see Figure 1, which demonstrates several typical scores). A small value close to 0 means a bad overlap, whereas a high value close to 1 means a good overlap. The F-score is calculated as follows: for a test span, if it has no overlap with the reference span, the F-score is set to 0; otherwise, $F = 2(PR/(P + R))$, where P is the precision (percentage of the test span that overlaps with the reference span) and R is the recall (percentage of the reference span that overlaps with the test span).

In our study, mapped read pairs with F-scores larger than 0.7 were selected, and the average and sum of the mapping quality of all selected pairs were calculated. A suspicious variant was identified as a true positive if the average and sum were above 30 and 90, respectively; otherwise, it was classified

Table 1: The samples and associated studies.

Sample	Study
HuRef	Levy et al. 2007 [24], Pang et al. 2010 [25]
NA10851	Ju et al. 2010 [26]
NA15510, NA18505	Korbel et al. 2007 [5]
NA18507	Bentley et al. 2008 [16], McKernan et al. 2009[27]
YH	Wang et al. 2008 [28]
NA12156	Kidd et al. 2008 [29]

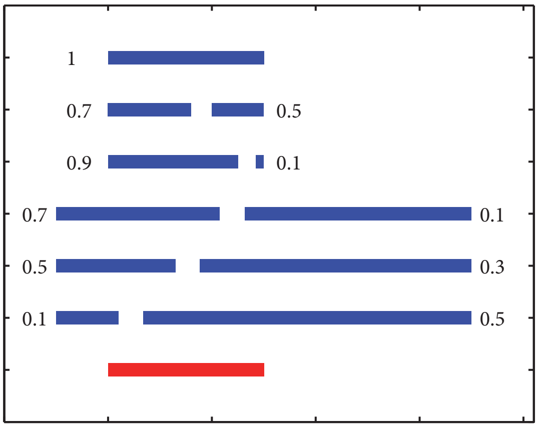


Figure 1: An illustration between F-score and overlapping quality. The bottom red span is the reference, and the 11 blue spans are tests, whose F-scores are shown with respect to the reference, ranging from 0.1 (very bad overlapping) to 1 (perfect one).

as a false positive. Therefore, a pair with a mapping quality of 90, two pairs with mapping quality of 45, or three pairs with a mapping quality of 30 constitute the minimal requirement to confirm a true positive. Figure 2 demonstrates two typical examples. It is shown that both regions have high GC content and low mappability; (a) shows no PEM signature while (b) does. Therefore, (a) is a false positive, and (b) is a true positive.

3. Results

We identified a total of 1923 suspicious variants, which cluster in 7 samples from 8 studies (see Figure 4(b) and Table 1). Among these 7 samples, the BAM files of four samples (NA18507, NA18505, NA12156, and NA10851) were downloaded from the FTP site of the 1000 Genomes Project (ftp://ftp-trace.ncbi.nih.gov/1000genomes/). For other three samples, i.e., YH, HuRef, and NA15510, sequencing data were downloaded: the paired reads of sample YH were downloaded from the FTP site of the YanHuang Project (ftp://public.genomics.org.cn/BGI/yanhuang/), and the SRAs of samples HuRef and NA15510 were downloaded from the NCBI FTP site using the *sra-toolkit* package.

BWA [30] was used to align short sequencing data to the reference genome hg18. Here, human reference genome hg18 was used in order to have a consistent genomic coordinate with the downloaded BAM files from the 1000 Genomes Project. The maximum insert size (-*a* option of BWA) was set to 1e4.

From these 1923 suspicious variants, 583 were detected as false positives. Complete information on suspicious variants and false positives is listed in Supplementary Table S1, and the validation figures of each false positive are shown in the supplementary FIG directory. Two typical examples (a false positive and a true positive) and statistical analysis are shown in the following examples.

Figure 2 shows two typical suspicious variants from NA18507. Figure 2(a) shows the variant with accession essv4528478, whose genomic location is chr1:210537566-210539442 (green bar in all panels). It is shown that both GC content (the magenta curve in the upper left panel) and mappability (the red curve in the middle left panel) profiles at the ROI are abnormal, and hence they yield a valley in the DOC profile (the blue curve in the lower left panel). As a result, a deletion variant was detected in the DOC profile. However, the PEM signature (the right panel) contains no mapped read pairs (the horizontal lines) that overlap with the green bar, suggesting that this suspicious variant is a false positive. In contrast, Figure 2(b) shows the variant with accession essv4968609 (genomic location chr1:154793140-154795767), whose GC content, mappability, and DOC profiles show similar behaviours to those in (a). However, the PEM signature contains well-mapped read pairs (three black lines in the centre) that overlap with the green bar, suggesting a true positive.

Figure 3(a) shows the distribution of the GC content of the human genome (hg18) with a bin size of 100 bp. Based on this distribution, the threshold values $th_1 = 0.26$ and $th_2 = 0.59$ are used to determine the estimated extreme GC content regions, such that both the left and right tail areas cover 5% of the whole distribution.

Figure 3(b) shows the distribution of the mappability of the human genome (hg18) with a bin size of 100 bp. Since exact 1 and 0 mappability values occupy a large portion of the distribution (68% and 7%, respectively), these two values were excluded from the distribution. Based on the remaining values, the threshold value $th_3 = 0.92$ was chosen such that the right area covers 20% of the distribution.

Figure 4(a) shows the distribution of suspicious variants and false positives across chromosomes. It is shown that the number of false positives decreases with respect to the chromosome index number. The correlation analysis between the chromosome lengths and the number of false positives yielded the correlation coefficient $r = 0.83$ and p-value of

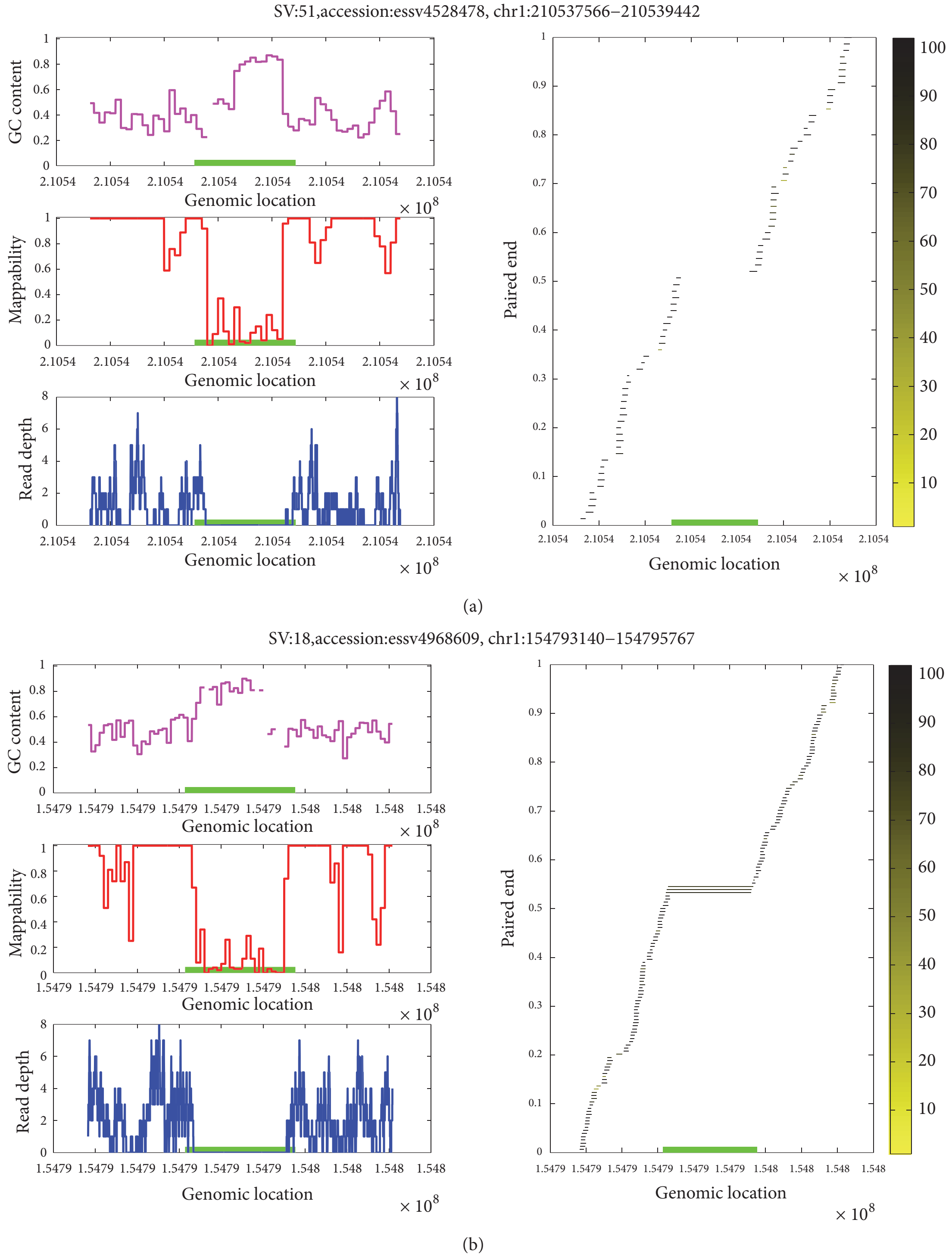
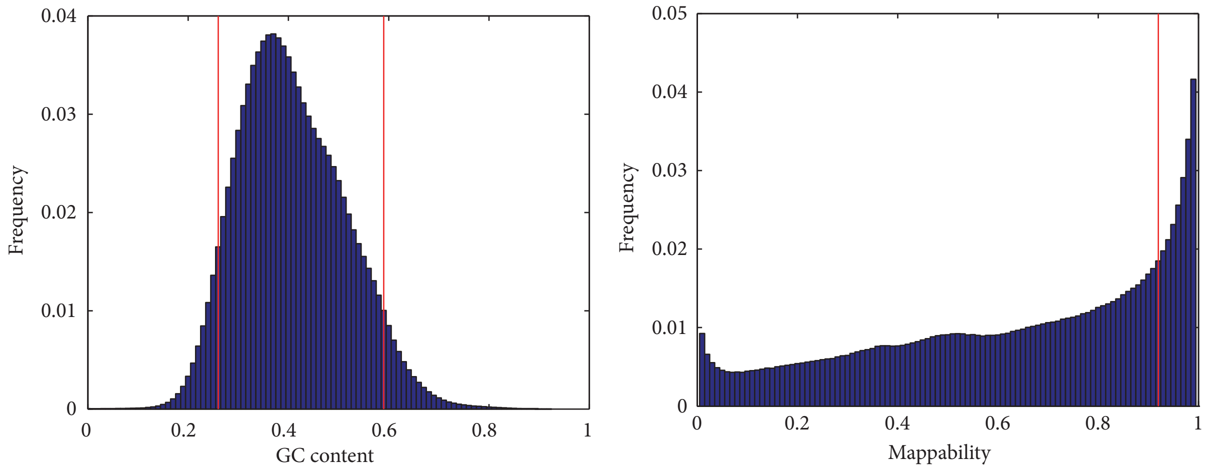


Figure 2: Two examples of suspicious variants. (a) A false positive and (b) a true positive of sample NA18507. The left upper, middle, and lower panels of each subfigure display the profiles of GC content, mappability, and DOC, respectively; the right panel displays the PEM profile, and each horizontal line represents a read pair, where the face colour encodes the mapping quality (yellow and black represent low and high mapping quality, respectively). The green bar in each panel is the studied DGV variant.

1.8e-6, indicating that the false positives are distributed evenly among chromosomes.

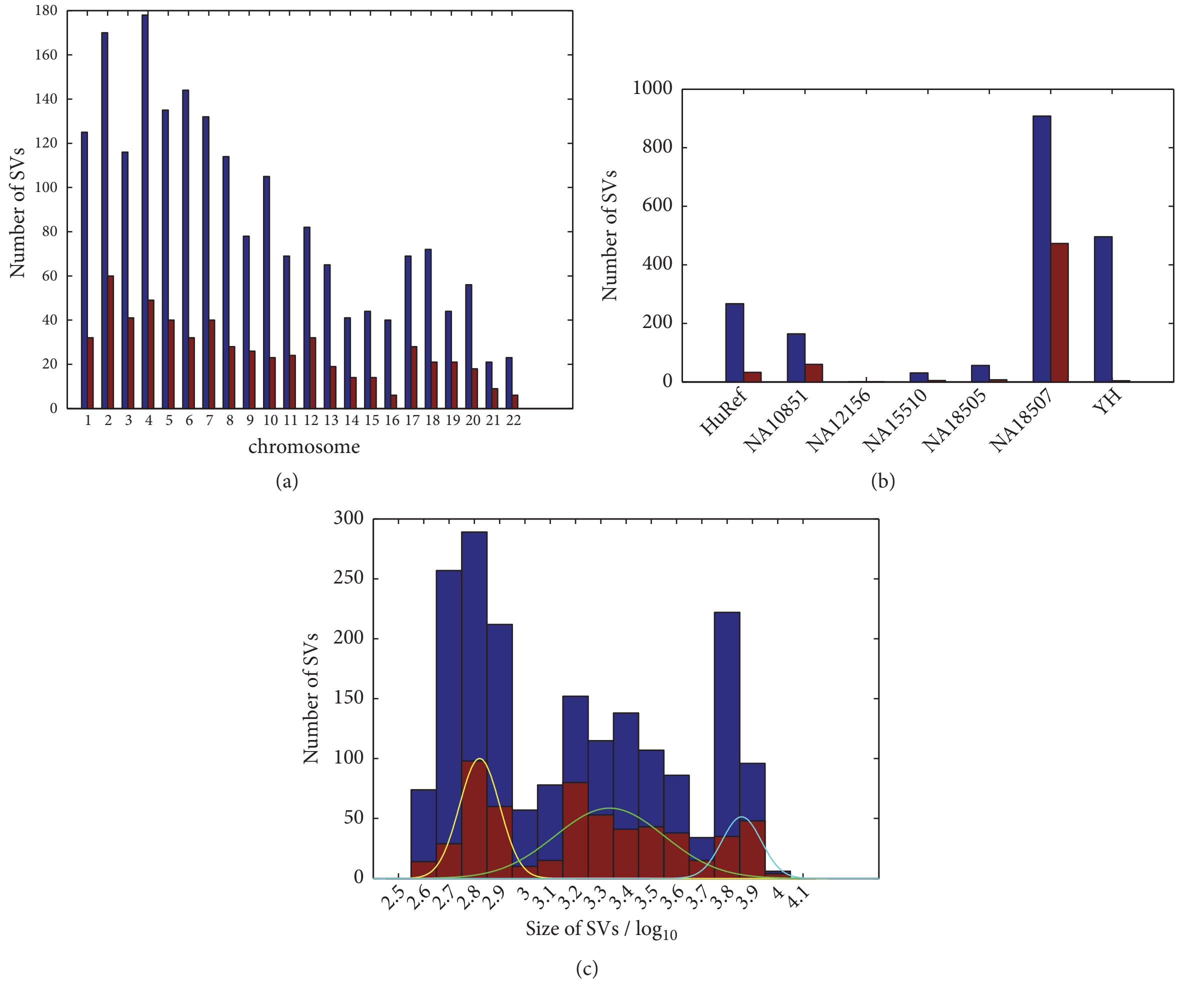
Figure 4(b) shows the distribution of suspicious variants and false positives across samples. It is shown that NA18507 contains more suspicious variants (908) and false positives (473) than the other samples. We further analysed the distribution of suspicious variants and false positives across studies and found that the studies by 'Bentley et al. 2008'



(a)

(b)

Figure 3: The GC content (a) and mappability (b) distribution of human genome (hg18). The three vertical red lines represent thresholds $th_1 = 0.26$, $th_2 = 0.59$, and $th_3 = 0.92$.



(a)

(b)

(c)

Figure 4: The distributions of variants with respect to the chromosome (a), sample (b), and size (c). Blue and red bars represent the suspicious and false-positive variants, respectively.

[16] and 'McKernan et al. 2009' [27] contained more false positives (289 and 184) than the other studies, and all of these false positives came from the sample NA18507.

Figure 4(c) shows the distribution of suspicious variants and false positives with respect to the size. Three modalities are shown: the left one with size smaller than 1 kbp (3), the middle one with size between 1 kbp and 5 kbp (3.7), and the right one with size larger than 5 kbp, which have proportions of 36%, 46%, and 18%, respectively. These results indicate that most false positives are small- or medium- sized variants. By fitting each modality with a Gaussian curve, the means of the three modalities are 660 bp (2.8), 2.2 kbp (3.3), and 7.2 kbp (3.9) for the left, middle, and right modalities, respectively.

4. Conclusion and Discussion

We proposed an approach to detect GC content and mappability related to false positives from the database of genomic variants. The proposed approach utilized the PEM signature, whose presence is necessary and provides evidence for true positives. 583 false positives were detected by conducting a validation study on the database of genomic variant. The results can avoid incorrect documentation and annotations in downstream studies.

We excluded variants with sizes larger than 10 kbp in this study, and the reasons are as follows: for most NGS alignment/mapping tools, the maximal insert size is limited to thousands of base pairs; e.g., the default values of both the -*a* parameter in BWA and -*X* parameter in Bowtie 2 are 500 bp, and a very large insert size degrades mapping performance. However, large variants (deletions and inversions) do require a large insert size. As a result, there is a conflict between the mapping quality and the maximal size of detectable variants, and a tradeoff has to be taken with caution. Therefore, we confined the maximal variant size to 10 kbp.

In this study, since there are several software packages/algorithms (e.g., smoothing, segmentation, mapping, etc.) and parameters (F-score, thresholds th_1, th_2, th_3, etc.) that are used in the method, the robustness is an important issue. A global optimization of all parameters is not conducted due to the larger number of parameters, but we set each software/algorithm to its recommended setting and tune each parameter separately to a reasonable value (e.g., th1, th2, and th3). We used the F-score to identify whether a segment overlaps with another segment. From Figure 1, we can see that a threshold value of 0.7 is appropriate to determine that two segments are almost overlapping with each other, so we used 0.7 in the validation step to determine an overlap status. When it is increased to 0.75, 644 false positives are detected, and when it is decreased to 0.65, 546 false positives are detected. Therefore, the results are roughly robust with respect to this parameter. In step 7, we used a large threshold value of 0.9 in order to narrow down the total number of suspicious variants to be validated. We also used three threshold values th_i ($i = 1, 2, 3$) to identify suspicious regions. In Figure 3(a), we chose $th_1 = 0.26$ and $th_2 = 0.59$ so that both the left and right tail areas covered 5% (or 10% in total) of the whole distribution, whereas in (b) we chose the threshold value $th_3 = 0.92$ so that the right area covered 20% of the distribution. Since the distribution of GC content is close to a Gaussian distribution, tail areas with 10% are appropriate. However, the distribution of mappability is far from a Gaussian distribution, and values other than 1 are unfavourable. When we adopted the strategy used for GC content to mappability, i.e., the left tail area covering 10%, the resultant threshold value was th_3, which was too tight to identify suspicious variants. Therefore, we chose 20% for the right area, which yielded a much looser threshold.

There are two limitations of the current study. First, since the PEM signature is more straightforward for studying deletion than other types of variants, the current study focuses on only this type of variant. In future studies, we hope to extend the spectrum of variants being studied. Second, this study focused on only the database of genomic variants as a pilot study to validate our method. In future works, we hope to conduct large-scale validations on other well-known public repositories related to structural variants, such as dbVar (https://www.ncbi.nlm.nih.gov/dbvar), which is NCBI's database of human genomic structural variation, and the Database of Genomic Variants archive (DGVa, https://www.ebi.ac.uk/dgva), which is an EMBL-EBI database that archives publicly available genomic structural variants of all species.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was partially supported by National Science Foundation of China (Nos. 61771381, 61401352, and 61571341) and NIH (Nos. R01GM109068, R01MH104680, and R01MH107354).

Supplementary Materials

Supplementary data are available with this article at http://gr.xjtu.edu.cn/c/document_library/get_file?p_l_id=2403541&folderId=2539941&name=DLFE-115097.zip. Table S1 lists the complete information of suspicious variants and false positives, and the FIG directory contains the validation figures of each false positive. *(Supplementary Materials)*

References

[1] J. L. Freeman, G. H. Perry, L. Feuk et al., "Copy number variation: new insights in genome diversity," *Genome Research*, vol. 16, no. 8, pp. 949–961, 2006.

[2] P. Stankiewicz and J. R. Lupski, "Structural variation in the human genome and its role in disease," *Annual Review of Medicine*, vol. 61, pp. 437–455, 2010.

[3] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome Research*, vol. 19, no. 9, pp. 1586–1592, 2009.

[4] A. Magi, M. Benelli, A. Gozzini, F. Girolami, F. Torricelli, and M. L. Brandi, "Bioinformatics for next generation sequencing data," *Gene*, vol. 1, no. 2, pp. 294–307, 2010.

[5] J. O. Korbel, A. E. Urban, J. P. Affourtit et al., "Paired-end mapping reveals extensive structural variation in the human genome," *Science*, vol. 318, no. 5849, pp. 420–426, 2007.

[6] D. Y. Chiang, G. Getz, D. B. Jaffe et al., "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nature Methods*, vol. 6, no. 1, pp. 99–103, 2009.

[7] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinformatics*, vol. 10, no. 1, article 80, 2009.

[8] E. Bellos, M. R. Johnson, and L. J. M. Coin, "cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data," *Genome Biology*, vol. 13, no. 12, pp. 1–11, 2012.

[9] M. Zhao, Q. Wang, P. Jia, and Z. Zhao, "Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives," *BMC Bioinformatics*, vol. 14, Suppl. 11, p. S1, 2013.

[10] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, "DELLY: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics*, vol. 28, no. 18, pp. i333–i339, 2012.

[11] J. Duan, J. Zhang, H. Deng, Y. Wang, and N. Salamin, "Comparative studies of copy number variation detection methods for next-generation sequencing technologies," *PLoS ONE*, vol. 8, no. 3, p. e59128, 2013.

[12] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Research*, vol. 40, no. 10, article e72, 2012.

[13] T. Derrien, J. Estellé, S. M. Sola et al., "Fast computation and applications of genome mappability," *PLoS ONE*, vol. 7, no. 1, Article ID e30377, 2012.

[14] M. F. Polz and C. M. Cavanaugh, "Bias in template-to-product ratios in multitemplate pcr," *Applied and Environmental Microbiology*, vol. 64, no. 10, pp. 3724–3730, 1998.

[15] D. Aird, M. G. Ross, W.-S. Chen et al., "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries," *Genome Biology*, vol. 12, no. 2, p. R18, 2011.

[16] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow et al., "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, pp. 53–59, 2008.

[17] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic Acids Research*, vol. 36, no. 16, p. e105, 2008.

[18] C. A. Miller, O. Hampton, C. Coarfa, and A. Milosavljevic, "ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads," *PLoS ONE*, vol. 6, no. 1, Article ID 16327, 2011.

[19] M.-S. Cheung, T. A. Down, I. Latorre, and J. Ahringer, "Systematic bias in high-throughput sequencing data and its correction by BEADS," *Nucleic Acids Research*, vol. 39, no. 15, p. e103, 2011.

[20] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, "The database of genomic variants: a curated collection of structural variation in the human genome," *Nucleic Acids Research*, vol. 42, no. 1, pp. D986–D992, 2014.

[21] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[22] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[23] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nature Methods*, vol. 6, pp. S13–S20, 2009.

[24] S. Levy, G. Sutton, P. C. Ng et al., "The diploid genome sequence of an individual human," *PLoS Biology*, vol. 5, no. 10, p. e254, 2007.

[25] A. W. Pang, J. R. MacDonald, D. Pinto et al., "Towards a comprehensive structural variation map of an individual human genome," *Genome Biology*, vol. 11, no. 5, 2010.

[26] Y. S. Ju, D. Hong, S. Kim et al., "Reference-unbiased copy number variant analysis using CGH microarrays," *Nucleic Acids Research*, vol. 38, no. 20, article no. e190, 2010.

[27] K. J. McKernan, H. E. Peckham, G. L. Costa et al., "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding," *Genome Research*, vol. 19, no. 9, pp. 1527–1541, 2009.

[28] J. Wang, W. Wang, R. Li et al., "The diploid genome sequence of an Asian individual," *Nature*, vol. 456, no. 7218, pp. 60–65, 2008.

[29] J. M. Kidd, G. M. Cooper, W. F. Donahue et al., "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, no. 7191, pp. 56–64, 2008.

[30] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.