Bioinformatics, 33(2), 2017, 184–191
doi: 10.1093/bioinformatics/btw591
Advance Access Publication Date: 14 September 2016
Original Paper

Sequence analysis

Seeksv: an accurate tool for somatic structural variation and virus integration detection

Ying Liang[1,2], Kunlong Qiu[2], Bo Liao[1,*], Wen Zhu[1], Xuanlin Huang[2], Lin Li[2], Xiangtao Chen[1] and Keqin Li[1,3]

[1]College of Information Science and Engineering, Hunan University, Changsha, Hunan 410082, China, [2]BGI, Shenzhen, Guangdong 518083, China and [3]Department of Computer Science State, University of New York, New Paltz, NY 12561, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on January 21, 2016; revised on July 29, 2016; accepted on September 6, 2016

Abstract

Motivation: Many forms of variations exist in the human genome including single nucleotide polymorphism, small insert/deletion (DEL) (indel) and structural variation (SV). Somatically acquired SV may regulate the expression of tumor-related genes and result in cell proliferation and uncontrolled growth, eventually inducing tumor formation. Virus integration with host genome sequence is a type of SV that causes the related gene instability and normal cells to transform into tumor cells. Cancer SVs and viral integration sites must be discovered in a genome-wide scale for clarifying the mechanism of tumor occurrence and development.

Results: In this paper, we propose a new tool called seeksv to detect somatic SVs and viral integration events. Seeksv simultaneously uses split read signal, discordant paired-end read signal, read depth signal and the fragment with two ends unmapped. Seeksv can detect DEL, insertion, inversion and inter-chromosome transfer at single-nucleotide resolution. Different types of sequencing data, such as single-end sequencing data or paired-end sequencing data can accommodate to detect SV. Seeksv develops a rescue model for SV with breakpoints located in sequence homology regions. Results on simulated and real data from the 1000 Genomes Project and esophageal squamous cell carcinoma samples show that seeksv has higher efficiency and precision compared with other similar software in detecting SVs. For the discovery of hepatitis B virus integration sites from probe capture data, the verified experiments show that more than 90% viral integration sequences detected by seeksv are true.

Availability and Implementation: seeksv is implemented in C++ and can be downloaded from https://github.com/qkl871118/seeksv.

Contact: dragonbw@163.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Structural variations (SVs) cover more than 50 bp segments involving DELs, insertions (INSs), duplications, inversions (INVs) and other complex rearrangements compared with the reference sequence. Compared with single nucleotide polymorphism (SNP), SV accounts for more differences between human genomes because of the number of covered nucleotides (Baker, 2012). Current methods are mostly based on high-throughput sequencing (HTS) reads. These methods can be divided into four categories: depth of coverage (DOC), paired-end mapping (PEM), split-read (SR) and assembly-based (AS) method. All these methods have limitations and are unsuitable for comprehensive SV types. DOC-based methods (Abyzov

et al., 2011; Szatkiewicz *et al.*, 2013; Xie and Tammi, 2009) assume that reads are uniform along chromosomes and the number of reads falling into a region follows the Poisson distribution. cn.MOPS (Klambauer *et al.*, 2012) models read count through multiple samples at each genomic position, which is robust to read count variations caused by technical and biological factors. Read count is an important signature for variation detection. DOC methods are best suited for copy number variations, such as duplications and DELs. PEM-based methods (Abyzov and Gerstein, 2011; Chen *et al.*, 2009; Hormozdiari *et al.*, 2010; Korbel *et al.*, 2009; Sindi *et al.*, 2009; Qi and Zhao, 2011) detect SVs according to two signatures: distance and direction between the mapped paired-end reads. InGAP-sv(Qi and Zhao, 2011) detects and visualizes SVs according to discordant paired-end read signals and several features involving local DOC, mapping quality and associated tandem repeat, which could find larger INSs and complex SVs with lower false discovery rate. The detected copy changed and copy invariant variants are of no exact breakpoint resolution and less power in low complexity region. SR based methods (Li *et al.*, 2013; Wang *et al.*, 2011; Ye *et al.*, 2009; Zhang *et al.*, 2016) utilize paired-end reads in which one end read is uniquely mapped to the reference and the other cannot. The unmapped read is supposed to span over breakpoint so that the part of it can be mapped if split. Sprites (Zhang *et al.*, 2016) uses the whole unmapped read rather than its clipped part to align to the target sequence, which aims to detect DELs with micro homologies or micro INSs. SR based methods can reach a bp resolution at the cost of restricted detection range of SV size. AS based methods (Alkan *et al.*, 2011; Chen *et al.*, 2014; Li *et al.*, 2011; Zhuang and Weng, 2015) should be the best method and robust to detect all types of variation in theory. In view of the complexity of the genome, such methods are generally applicable to the detection of variations in simple organisms and conventional *de novo* assembly methods are not designed to detect the variations (Zhuang and Weng, 2015). None of the above four main approaches is comprehensive (Alkan, *et al.*, 2011), and they are complementary to one another. Current SV detection methods (Bellos *et al.*, 2012; Jiang *et al.*, 2012; Rausch *et al.*, 2012; Sindi *et al.*, 2012) take advantage of two or three signatures, which can avoid weakness of a single signature and achieve a good detection effect. Svclassify (Parikh *et al.*, 2016) combines whole-genome sequencing data sets from different sequencing technologies, and the unsupervised machine learning method is employed to genotyping SVs and One Class Classification is used to classify candidate SVs into likely true or false positives, which forms high-confidence SV and non-SV calls.

However, these methods aim to detect germline genomes based on a single sample and somatic SV, which can alter normal gene function (Yang *et al.*, 2013) and lead to tumor formation (Carter *et al.*, 2012). Somatic SV is implicated in cancer gene overexpression or underexpression, exhibiting a causative role in cancer initiation. CREST (Wang, *et al.*, 2011) utilizes soft-clipped reads to precisely locate breakpoint. It is particularly well suited for detecting somatically acquired SVs but with some flaws:

CREST calls CAP3 (Huang and Madan, 1999) software to assemble soft-clipped reads based on sequence similarity without considering the location of its mapped part in the reference genome. Suppose two soft-clipped reads originate from different sequences, such as read-M and read-N, in which the red is the soft-clipped sequence.

read-M:CCCTAACCCTAACTGGCGTCCCAAATTGCAAAAGCGTATCA
TGCGGCTACGTATCGTTTT

read-N:CCCTAACCCTAACTGGCGTCCCAAATTGGCTACGTATCGTT
TTAAAAAAAAAAAAAAA

CREST can obtain a long sequence GCAAAAGCGTATCATGCG GCTACGTATCGTTTTAAAAAAAAAAAAAAAA as the assembled result, which should not be simply assembled.

CREST assesses variation types according to single breakpoint information; thus two breakpoints of one SV may be mistakenly judged to two different types of variation. For example, a large INS may be inferred as DEL and INS by error.

In this paper, we propose a new SV detection pipeline named seeksv which is developed for somatic case-control SV detection but also can be used for single germline genome analysis. Seeksv does not depend on any assembly software and uses Burrows-Wheeler Aligner (BWA) to align the assembled contig back to reference. CREST uses BLAT for alignment which requires the high configuration of computer. Various detection signals, namely, SR signal, discordant PEM signal, DOC signal and fragment with both ends unmatched are comprehensively used by seeksv. Both unmapped ends may imply breakpoint clues that require attention. Seeksv extracts the fragment with two ends unmapped from original alignment results and invokes COPE (Liu *et al.*, 2012) to obtain a long single-end read and then aligns back to the reference sequence. In general paired-end sequencing reads are exploited to analyze SV, but seeksv accommodates various types of sequencing data, such as paired-end sequencing reads and single-end sequencing reads. To the best of our knowledge, seeksv is the first tool to utilize single-end reads for SV detection. Complex genomes contain many homologous sequences with maximal similarity, and some SVs would be missed if breakpoints are located in the homologous region with multi-alignment. Seeksv develops a rescue model to deal with SVs located in the repetitive region. When the mode is turned on, seeksv preserves the detailed results of the multi-alignment to calculate the position of breakpoints.

The integration of viral sequences and host genomes is a special SV. Tumor viruses invading host cells may result in cell proliferation and uncontrolled growth, and eventually induce cell transformation and tumor formation. The integration relationship between the virus and host should be analyzed for clarifying the mechanism of tumor occurrence and development. The SV tool can detect viral integration sites that are vital for tumor genesis studies. Seeksv possesses excellent ability to identify viral integration sites with whole-genome sequencing data or probe capture data. Compared with traditional research methods, the resolution of seeksv achieves single base level and simultaneously detects all the integration events with higher accuracy and efficiency.

2 Materials and methods

2.1 Term description

The following terms and their descriptions are used by the proposed method.

1. Mate reads: At paired-end sequencing, a fragment will produce two paired-end reads, namely, read1 and read2; read1 is the mate read of read2 and read2 is the mate read of read1.
2. Soft-clipped reads: In BWA alignment results, the read at first alignment fails, but a portion of it is successfully aligned after applying the mate-SW algorithm. According to the relative positions of aligned and clipped reads, the soft-clipped read can be divided into two categories: left soft-clipped read and right soft-clipped read.
3. Left soft-clipped read: The reads with the left part clipped after realigning back to the reference sequence.

4. Right soft-clipped read: The reads with the right part clipped after realigning back to the reference sequence.
5. Left soft-clipped sequence: Short for left soft-clipped consensus sequence, merged by left soft-clipped reads that are clipped by the same junction.
6. Right soft-clipped sequence: Short for right soft-clipped consensus sequence, merged by right soft-clipped reads that are clipped by the same junction.

Seeksv depends on the input BAM file (Li *et al.*, 2009) to extract discordant mate reads and soft-clipped reads. The mapping situation can be divided into four categories, namely (i) concordant mate reads, in which mate reads r_1 and r_2 are both mapped and the distance, orientation and order of mapped are concordant with the expectation; (ii) discordant mate reads, in which r_1 and r_2 are both mapped but at least one of the situations (distance, orientation and order) is discordant with the expectation; (iii) split mate reads, in which one of the mate reads is uniquely mapped but the other one is a soft-clipped read and (iv) unmapped mate reads, in which both of the mate reads fail to align. Seeksv detects four types of variants: INS, DEL, INV and inter-chromosome transfer (CTX). The four categories contain many other types of variants. For example, tandem duplication is regarded as INS. The PEM and SR signatures from the second and third categories of the above mentioned SVs are illustrated (Supplementary Fig. S1).

Seeksv utilizes soft-clipped reads to locate breakpoint precisely. Breakpoints are surrounded by some discordant mate reads and read counts spanning the breakpoint are significantly different. Extra discordant PEM information and read count data are employed to verify the reliability of the discovered breakpoint.

2.2 Realign

Soft-clipped reads cannot be mapped to the reference as a whole but their mate reads are uniquely mapped. Soft-clipped read is distinguished according to the CIGAR sign in BAM file. BWA automatically splits the soft-clipped read into parts to map one of them back into the reference. The CIGAR field indicates base-level alignment information, which is very crucial for subsequent soft-clipped read analysis. As an example, let us use the reference sequence AGCCTTCAATCCGGTATCAT. If the read is TCATCCAGGCAT, then the alignment situation is as follows:

```
reference: AGCCTTCAATCCGGTATCAT
read:          TCATCCAGG CAT
```

The CIAGR would be 3M1D3M1I2M3D3M, and the soft-clipped read is marked 'S' in CIGAR. Sometime the marked soft-clipped read would be marked 'S' by mistake. If the clipped part of the soft-clipped read is aligned to the reference again, its mapping position is adjacent with the tail end of the mapped part. This result indicates that the read should be mapped to the reference with full-length, but it is marked as a soft-clipped read by mistake. Although the soft-clipped read will be aligned back to reference after merging, exogenic virus sequence INS detection depends on breakpoint information. Thus the soft-clipped information should be accurate and realign back to the reference before merging is necessary. If the clipped part can be successfully mapped to the end of the mapped part successfully, the clipped information should be discarded and no longer used to deduce breakpoint. Clipped parts are also considered as being successfully mapped to the end of its mapped part within a few bp, which is controlled by the threshold value and can be adjusted by the user.

2.3 Soft-clipped read assembly

A junction is defined as two regions that are close to each other in the individual genome but separate in the reference genome (Supplementary Fig. S2). The soft-clipped read can be left clipped or right clipped. In general, the mapped part of the clipped read is closer to its mate unique mapped read (Supplementary Fig. S3.). The soft-clipped read can be grouped according to the (i) clipped direction and (ii) mapped coordinate position of its mapped part. The soft-clipped read with the same clipped direction and close mapped coordinate position is thought as from one junction and will be merged into a long consensus sequence if sequence similarity exceeds a certain threshold. The long consensus sequences from left-clipped reads are called left-clipped sequences, and right-clipped sequences are from right-clipped reads. For the long consensus sequence, seeksv generates k-mer substrings of its mapped part and clipped part as an atomic unit for matching. In general, the mapped part should have high similarity to the clipped part of the matched sequence and vice versa. The long consensus sequences and their matched partner sequences together determine a junction. Most junctions are supported by the two consensus long sequences. Some special junctions only have unilateral clipped reads (left or right), so only one long assembled sequence exists. Seeksv needs to align its clipped part back to the reference, if uniquely mapped; the junction can also be depicted. Seeksv concludes SV type according to the relative position and clipped direction of two matched long consensus sequences. When clipped part of soft-clipped sequence has multiple locations, all of them would be recorded by seeksv with more stringent filter conditions in order to find SV as much as possible without decrease of precision. Seeksv revokes a rescue mode to deal with the following conditions. (i) Clipped part of left soft-clipped sequence or right soft-clipped sequence has multiple locations, the clipped part would be used as supporter of SV detected by its matched sequence if consistency. (ii) Clipped part of left and right soft-clipped sequences have multiple locations, seeksv locates the clipped part with the help of additional information of mapped part of its matched sequence, the detected breakpoint is also reliable if consistency. Details of the algorithm are shown in Algorithm 1. Each soft-clipped read r_i has four attributes *dir, pos, mapped and clipped*. *Dir*(r_i) means the clipped direction of r_i. *Pos* (r_i) means the mapped position of r_i.*Clipped* (r_i) means the clipped sequence of r_i. *Mapped* (r_i) means the mapped sequence of r_i. Function *con()* is used to measure the consistency of the two sequences. *Merge()* is used to merge two sequences, and *chr()* is the chromosome information of the mapped position, and α is sequence similarity threshold, which is tunable for users, default is 85%.The input BAM file should be sorted and virus genome reference sequence is necessary if seeksv is used for virus integration detection. Algorithm 1 takes the junction detection procedure of DEL as an example. Steps 1–6 merge the soft-clipped reads which are considered as from one junction into a long consensus sequence according to their clipped direction and mapped coordinate position of mapped part. Steps 7–19 look for matched soft-clipped sequence and detect junction from the long soft-clipped sequence set **R.** Steps 7–10 detect the position of junction under the condition that clipped part of left soft-clipped sequence and right soft-clipped sequence are both uniquely mapped to the reference genome. Steps 11–19 detect the position of junction under the rescue mode, steps 11–14 detect junction under the condition (i); steps 15–19 detect junction under the condition (ii) and they would be the support information for related junction if consensus threshold is satisfied. Steps 20–23 detect junction which only have unilateral soft-clipped sequence, if the clipped part of unilateral soft-clipped sequence is uniquely mapped, seeksv deduces junction from

Algorithm 1 Algorithm to Identify Junction Location

Input: BAM file, Genome reference sequence.
Output: Two position of junction.
1: Extract soft-clipped read set $R(r_1, r_2, \cdots, r_n)$.
2: **if** $dir(r_i) = dir(r_{i+1})$ and $max(length(clipped\ (r_i)), length\ (clipped(r_{i+1}))) - |pos(mapped(r_i)) - pos(mapped(r_{i+1}))| \leq 5$ **then**
3: **if** $con(r_i, r_{i+}\ 1) > \alpha$ **then**
4: Update:
$mapped(r_{i+1}) \leftarrow merge(mapped\ (r_i), mapped(r_{i+1}))$
$clipped(r_{i+1}) \leftarrow merge(clipped\ (r_i), clipped(r_{i+1}\ ))$
$pos(r_{i+1}) \leftarrow max(pos(mapped\ (r_i)), pos(mapped\ (r_{i+1})))$
Delete r_i.
5: **end if**
6: **end if**
7: **if** $clipped\ (r_i)$ and $clipped(r_j)$ are uniquely mapped **then**
8: **if** $con(mapped(r_i), clipped\ (r_j)) > \alpha$ and $con(clipped(r_i), mapped(r_j)) > \alpha$ **then**
9: $match(r_i, r_j)$
$pos_1(junction) = min(pos(mapped\ (r_i)), pos(mapped(r_i)))$
$pos_2(junction) = max(pos(mapped(r_i)), pos(mapped\ (r_j)))$
10: **end if**
11: **else if** $clipped(r_i\ )$ is uniquely mapped **then**
12: **if** $con(clipped\ (r_i), mapped\ (r_j)) > \alpha$ **then**
13: Deduce $pos(mapped\ (r_i))$ according to r_j and go to step 8∼9.
14: **end if**
15: **else if** $clipped(r_i)$ and $clipped(r_j)$ have multiple locations **then**
16: **if** $con(mapped(r_i), clipped\ (r_j)) > \alpha$ and $con(clipped(r_i), mapped(r_j)) > \alpha$ **then**
17: As support information for related junction.
18: **end if**
19: **end if**
20: Realign unilateral $clipped(r_i)$ to reference genome
21: **if** $clipped(r_i)$ is uniquely mapped **then**
22: $pos_1(junction_i) = min(pos(clipped(r_i)), pos(mapped(r_i)))$
$pos_2(junction_i) = max(pos(clipped(r_i)), pos(mapped(r_i)))$
23: **end if**

the position of its clipped part and mapped part or the unilateral soft-clipped sequence would be discarded.

Seeksv also corrects inaccurate mapping of soft-clipped reads automatically. The soft-clipped read appears to be clipped prematurely, which should be clipped at position X instead of $X-n$ or $X+m$. The premature clipped read may contain important clues for junction detection. To solve this problem, seeksv appends some bases from the clipped part to the mapped part and aligns to the merged long consensus sequence (Supplementary Fig. S4). If the concordance rate is greater than threshold T, the long sequence is also supported by the read.

2.4 Paired-end read validation

For candidate SV detected by previous clipped reads, many discordant read pair signatures must be present. To define these discordant read pairs, traversing all aligned reads to calculate the mean insert size (L) and SD is necessary. Seeksv searches for discordant read pairs to support candidate SV with two conditions (Fig. 1):

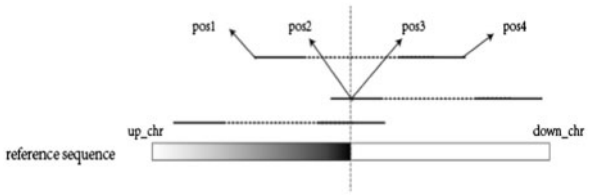

Fig. 1. Illustration of discordant read pairs around candidate SVs. Discordant read pairs can be divided into two conditions according to up_chr and down_chr are same or not

1. if up_chr == down_chr: $mapped\ length(ML) < L\text{-}4\times SD$ or $ML > L+4\times SD$,
2. else: $L - 4\times SD \leq |pos2\text{-}pos1+1| + |pos4\text{-}pos3+1| \leq L+4\times SD$.

The eligible read pair is collected as evidence to validate the previous junction and at least three discordant read pairs are needed (can be adjusted by user). Seeksv also records the average read depth of the whole reference and the region near the junction for further verification.

2.5 Viral integration detection

Viral integration detection is similar to SV detection, except for additional adjustments. The analysis pipeline of viral integration involves the following steps: (i) building mixed reference sequence; (ii) scanning the alignment results (BAM file) and merging the reads from the same junction according to the alignment position; (iii) realigning the clipped part of soft-clipped reads back to the mixed reference sequence; (iv) outputting relevant integrated virus sequence and total integration results (see Supplementary Material for details). Seeksv joins the reference sequences of humans and viruses together to build a mixed reference sequence and realigned soft-clipped reads back to the new reference sequence to obtain related virus sequence information. This pipeline cannot only obtain the high accuracy of the viral integration site information, but also has the following functions. (i) The pipeline strictly detects and distinguishes integration direction of the virus sequence. If the mapped and clipped parts of the soft-clipped sequence are both positively aligned back to the reference sequence, the virus integrates into the genome directly forward. If the mapped part is positive and the clipped part is negative alignment. The virus sequence is initially inverted and then integrates into the host genome. (ii) The pipeline identifies the micro homologous sequence and its length if it is within the vicinity of the DNA breakpoint and viral DNA/RNA breakpoint. (iii) The pipeline detects small fragment INSs around the viral integration breakpoint (Supplementary Figs S5–S7).

3 Results

To comprehensively evaluate the performance of seeksv, simulated and real data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2010a) are tested. Single-end reads are also simulated to test the performance of seeksv in and out of rescue mode for comparison. Seeksv, DELLY (Rausch *et al.*, 2012), CREST and PRISM (Jiang *et al.*, 2012) are conducted based on simulated data. DELLY and PRISM mainly focus on germline SV detection. Seeksv and CREST are devoted to somatic SV detection based on pair samples, but they can also be used for single samples.

For single sample analysis, seeksv, DELLY, CREST and PRISM are performed and compared. For pair samples, seeksv and CREST are run for somatic SV analysis, DELLY and PRISM are conducted to detect somatic genome SV, and then make difference set with germline SV to obtain somatic SV. For real data from the 1000 Genomes Project, NA12878 was selected because it has already been analyzed by many other SV tools. The trio samples (NA19238, NA19239 and NA19240) were also comparatively analyzed by the above four detection tools. We also apply seeksv to whole-genome sequencing data from five esophageal squamous cell carcinoma (ESCC) case-control samples which are sequenced by the Beijing Genomics Institute (BGI). For viral integration detection, seeksv is performed based on MyGenostics Company probe and BGI probe sequencing data of five positive samples to detect integration breakpoints; some of the breakpoints are selected to carry out the biological experiments.

3.1 Evaluating seeksv on simulated data

Germline and cancer genomes are simulated based on chr11 of hg19 with three kinds of SVs added, including DEL, INS and INV. Germline genome derives from the reference genome and cancer genome derives from germline genome with a minimal difference in SV, which inherits most of germline SV and derives specific new SV. The variation ratios are 0.001 and 0.00003, respectively, for germline and cancer genomes. For germline SV simulation, a total of 131 large SVs are generated, including 49 large INSs (36 heterozygous and 13 homozygous) ranging from 1 Kb to 4 Mb, 53 large DELs (37 heterozygous and 16 homozygous) ranging from 1 Kb to 4 Mb and 29 large INVs (20 heterozygous and 9 homozygous) ranging from 1 Kb to 2 Mb. For somatic SV simulation, a total of 247 large SVs are generated, including 94 large INSs (62 heterozygous and 32 homozygous) ranging from 1 Kb to 4 Mb, 98 large DELs (64 heterozygous and 34 homozygous) ranging from 1 Kb to 4 Mb and 55 large INVs (35 heterozygous and 20 homozygous) ranging from 1 Kb to 2 Mb. And a total of 116 large SVs only exist in the cancer genome, including 45 large INSs (26 heterozygous and 19 homozygous) ranging from 1 Kb to 4 Mb, 45 large DELs (27 heterozygous and 18 homozygous) ranging from 1 Kb to 4 Mb and 26 large INVs (15 heterozygous and 11 homozygous) ranging from 1 Kb to 2 Mb. Single-end and paired-end sequencing reads are generated from germline and cancer genomes under different read length, coverage and insert size parameter settings with sequence error rate of 0.01.The read is simulated by using the BGI in-house read simulation software called simulate_solexa_reads. The quality profile of simulated read comes from real sequencing read of BGI. The BWA (Li and Durbin, 2009) is used to map all simulated reads to the reference chr11 genome and SAMtools (Li *et al.*, 2009) is used to sort and index artificially synthesized BAM files. In single-end read mode, the read length should be better longer than 50 bp. Seeksv is conducted to detect genome and somatic SVs with general parameters. The detected results are compared with simulated ground-truth SVs and judged to be consistent with simulated events if the distance between predicted breakpoints and simulated events is within 50 bp. The receiver operating characteristic curve of DEL, INV and INS detection under different coverage in single-end mode and paired-end mode under and without rescue mode are illustrated (Supplementary Figs S9 and S10). In paired-end read mode, three another SV detection tools are conducted to compare performance with seeksv, namely delly_v0.6.1 (Rausch *et al.*, 2012), PRISM_v1.1.6 (Jiang *et al.*, 2012) and CREST_v1.0 (Wang *et al.*, 2011) using default parameters. In the detection results of DELLY, lots of predicted SVs are marked as 'LowQual', which would be filtered out for future analysis. The comparison results of germline SV and somatic SV are summarized in Figure 2. The true positive rate (TPR) and precision (P) are calculated to measure the performance of the above four detection tools. Reads are simulated under different settings, which are coverage, insert size and read length. And in each experiment, only one simulated parameter is changed. The default parameters are set as coverage is 20×, read length is 100 bp, insert size and SD is N (500, 30).

In paired-end sequencing mode, all four detection tools achieve more than 85% TPR on almost all parameter settings for germline and somatic SVs detection. And seeksv, CREST and PRISM also have a relatively high precision, but DELLY is about 50% for all parameter settings. Seeksv achieves high precision, which is outperforming its peers with a slight decrease in TPR. For DEL and INV detection of the four detection tools, with the increase of sequencing coverage, the TPR also increases, but the precision has decreases, and INS detection is on the contrary. The change in insert size and SD has little effect on the TPR and precision. For change in read length, the TPR of INS detection is obviously decreases with the increment of read length. The TPR and precision across all SV types demonstrates that seeksv would be a better choice for single-end and pair-end reads.

3.2 Evaluating seeksv on data from the 1000 genomes project

Simulation data are too perfect to show the actual performance of the detection tools. A real data set is needed to evaluate the

	Coverage(X)				Read length(bp)		Insert size	
	15	20	30	40	75	150	N(300,20)	N(600,40)
Germline SV								
Deletions								
seeksv	0.91/1.00	0.91/0.94	0.98/0.98	1.00/0.96	0.75/1.00	0.96/0.93	0.94/0.98	0.92/1.00
PRISM	0.85/0.87	0.92/0.92	0.98/0.84	1.00/0.85	0.85/0.90	0.72/0.93	0.87/0.90	0.92/0.91
DELLY	0.96/0.53	0.96/0.53	0.96/0.52	0.96/0.53	0.96/0.54	0.96/0.52	0.96/0.54	0.96/0.53
CREST	0.91/0.96	0.92/0.91	1.00/0.96	1.00/0.95	0.79/0.89	0.98/0.93	0.96/0.93	0.94/0.94
Inversions								
seeksv	0.97/1.00	0.97/0.97	1.00/1.00	1.00/0.94	0.90/0.90	1.00/0.97	1.00/1.00	1.00/1.00
PRISM	1.00/0.97	1.00/0.91	1.00/0.71	1.00/0.76	1.00/0.91	1.00/0.97	1.00/0.88	1.00/0.94
DELLY	1.00/0.50	1.00/0.50	1.00/0.50	1.00/0.50	1.00/0.51	1.00/0.50	1.00/0.50	1.00/0.50
CREST	1.00/0.97	1.00/0.94	1.00/1.00	1.00/0.97	0.93/0.90	1.00/1.00	1.00/1.00	1.00/1.00
Insertions								
seeksv	0.92/1.00	0.90/1.00	0.86/0.98	0.84/0.98	0.96/0.98	0.88/1.00	0.86/0.93	0.88/1.00
CREST	0.94/0.94	0.88/0.98	0.92/0.75	0.93/0.73	0.92/0.96	0.86/0.93	0.88/0.91	0.84/0.93
Somatic SV								
Deletions								
seeksv	0.93/1.00	0.97/0.98	0.97/0.95	0.98/0.96	0.78/0.97	0.97/0.97	0.95/0.97	0.99/0.96
PRISM	0.78/0.84	0.91/0.86	0.99/0.86	1.00/0.82	0.86/0.88	0.69/0.89	0.86/0.86	0.90/0.85
DELLY	0.92/0.50	0.92/0.50	0.92/0.50	0.92/0.50	0.91/0.50	0.94/0.50	0.91/0.50	0.92/0.50
CREST	0.92/1.00	0.98/0.97	0.98/0.94	0.98/0.92	0.82/0.94	0.98/0.96	0.93/0.96	0.94/0.97
Inversions								
seeksv	0.98/0.96	0.96/1.00	1.00/0.95	1.00/0.96	0.98/0.96	0.98/1.00	0.98/0.98	1.00/0.98
PRISM	1.00/0.96	1.00/0.95	1.00/0.74	1.00/0.70	1.00/0.83	0.96/0.96	1.00/0.95	1.00/0.92
DELLY	1.00/0.50	0.98/0.50	0.98/0.50	0.98/0.50	1.00/0.50	1.00/0.50	1.00/0.50	1.00/0.50
CREST	0.98/0.95	1.00/0.98	1.00/0.96	0.98/0.93	0.98/0.98	0.98/0.98	1.00/0.96	1.00/1.00
Insertions								
seeksv	0.85/0.99	0.84/0.99	0.82/0.97	0.84/1.00	0.91/0.99	0.81/0.99	0.83/0.98	0.82/1.00
CREST	0.89/0.93	0.81/0.96	0.81/0.96	0.79/0.93	0.84/0.93	0.77/0.96	0.81/0.96	0.78/0.97

Fig. 2 TPR and P of the four tools to detect simulated SVs of germline and somatic tumor genomes. The number in the table is represented as TPR/P form

performance of seeksv. Four sample data sets NA19240, NA19238, NA19239 and NA12878 are downloaded from the 1000 Genomes Project. They are all low coverage sequencing data (4×coverage), which were sequenced in Pilot 1. Given the large scale of chromosomal rearrangement of SV and absent gold standard reference set, only DELs greater than 50 bp detection performance of the 1–22 chromosomes are compared. Seeksv and DELLY are conducted, which use both discordant mappings and split reads to detect SVs. Both SVs and 'LowQual' filtered SVs predicted by DELLY are analyzed. The callset released by the 1000 Genomes Project Consortium (2010b) is considered the gold standard of results, which can be downloaded from the Database of Genomic Variants (MacDonald *et al.*, 2014). The predicted results are considered to be consistent if the predicted SV interval overlaps with gold standard, which is less strict than simulation data analysis. Although the 1000 Genomes callset is comprehensive and convincing, it may still miss some actual DELs in an individual's genome. Seeksv and DELLY are run with default parameters, and the results are shown in Figure 3. The tabular form of predicted SVs of four detection tools is also presented (Supplementary Table S1).

3.3 Evaluating seeksv on ESCC samples

To demonstrate the performance of seeksv on real tumor datasets, five whole-genome sequencing ESCC samples are selected; each sample is sequenced to an average of 40× coverage. Seeksv has identified a total of 847 SVs, including 178 DELs, 122 INSs, 114 INVs and 433 CTXs across the five samples (Supplementary Table S1). ESCC have been analyzed by many research teams (Cheng *et al.*, 2016; Gao *et al.*, 2014). 5204 SVs from the 31 ESCC samples SVs have been has identified by Cheng *et al.*, (2016) with an average of 168 SVs per sample. Among SVs detected by seeksv across the five samples, many of which are located in some well-known ESCC-associated genes such as CDKN2A, TP53, RB1 and some newly published ESCC-related genes, including MACROD2, FHIT and PARK2 (Hu *et al.*, 2016). 1275 DELs and 1074 INVs are extracted from Cheng's SV results and compared with the results of seeksv. 140 of 178 DELs from seeksv are overlapping with 1275 DELs from Cheng, 98 of 114 INVs from seeksv are overlapping with 1074 INVs from Cheng. All these overlapped SVs may be the common variations shared by ESCC samples which are expected to be further analyzed.

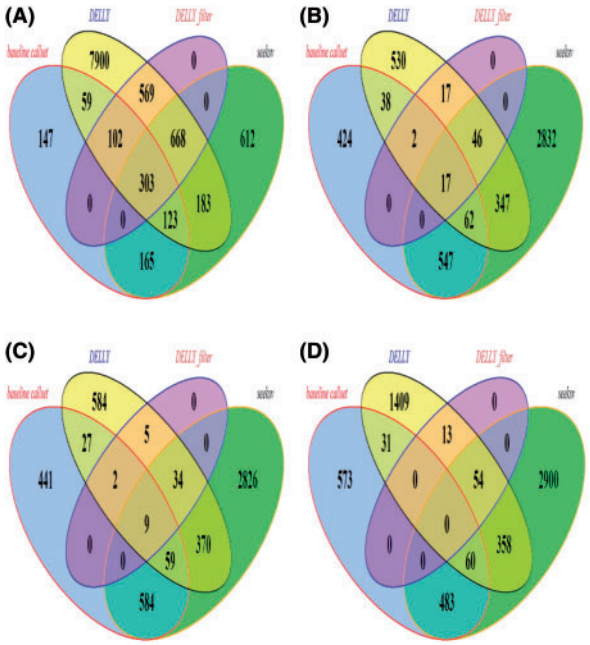

Fig. 3 Private/common calls of four samples. (A) NA12878. (B) NA19238. (C) NA19239. (D) NA19240. The cornflower blue area is SV result of baseline, green area is seeksv, yellow area is DELLY and dark orchid is DELLY after filtered

3.4 Evaluating seeksv on hepatitis B virus integration data

To verify the performance of seeksv in hepatitis B virus integration detection, parallel experiments (case–control) with five samples (695, 728, 807, 815 and 905) infected with HBV are conducted. The sequencing depth of samples is more than 200× using two HBV capture probes from MyGenostics and BGI. Seeksv detects HBV breakpoints based on the two capture platforms. Table 1 lists the detection result of seeksv. The difference in detection results between the two platforms is compared. For sample 695, seeksv detects 17 breakpoints both from MG Company and BGI, and 13 of them are overlapping. For samples 807, 905 and 728, seeksv detects more BGI unique breakpoints, which may be caused by the difference between the capture platforms. For sample 815, the detection situation is similar to sample 695, which has the highest percentage of breakpoints overlapping.

Some of the detected breakpoints are selected for experimental verification. As the BGI probe has more unique breakpoints, 20 detected breakpoints from the BGI platform and 18 breakpoints from MG Company are selected. There are 14 common breakpoints between them. Six and four remaining unique breakpoints were from BGI and MG Company, respectively. The verification results are shown in Table 2.

Among the 14 overlapping breakpoints, 13 of them are experimentally validated and 3 of the remaining 4 MG probe unique breakpoints are validated. For the BGI probe, six unique breakpoints are detected by seeksv, and all of them are validated. Seeksv has more unique breakpoints, but these unique breakpoints are true. This finding shows that seeksv possesses very high accuracy. In terms of the breakpoints that are not detected by seeksv from the MG probe platform, the detection performance of breakpoints may be related to the technique of the probe capture. Seeksv shows good compatibility and stable viral integration detection under different probe capture platforms. Seeksv can also detect viral integration breakpoints in the scale of the whole genome.

Table 1 Detection results of HBV integration

Sample	MG_uniq_number	Overlapping_number	BGI_uniq_number
695	4	13	4
807	2	13	11
815	2	10	2
905	7	6	15
728	0	4	17

All breakpoints have supporting reads ≧3.

Table 2 Verification results of two probes

Capture platform	Total breakpoints	Verified	Verified rate
MG_probe	18	16	89%
BGI_probe	20	19	95%

4 Discussion

In this study, we have developed a novel SV calling methodology that involves DEL, INS, INV and CTX. This methodology can also detect viral integration breakpoints if users offer a virus reference sequence. Seeksv comprehensively uses four different detection signals, namely, SR signal, discordant PEM signal, DOC signal and the fragment with both ends unmatched to avoid the weakness of a single signature. Soft-clipped reads cannot be completely mapped into the reference sequence, but when the read is clipped and partial sequence can be mapped uniquely. Soft-clipped-reads are fully used, which offer base-level breakpoint information. Unlike previous methods, seeksv merges soft-clipped-reads from the same breakpoint into a clipped long sequence individually and does not rely on any of the assembly software. These features reduce configuration requirement of the computer for software. In assessing the type of SV, seeksv is very comprehensive as it combines the information of two junctions instead of only one junction to make judgments, thereby greatly reducing the false positive rate. Even if SV is located in the sequence homology region, which may be missed because of sequence similarity, seeksv has a rescue model to reduce the loss as much as possible. When seeksv is applied to viral integration detection, it can automatically generate a mixed reference sequence of different populations according to the user's input. Tables 1 and 2 show that seeksv possesses very high sensitivity and accuracy. Most of the detected breakpoints are confirmed to be true. Single-end sequencing data, which usually offer DOC signal and are only applied to copy number variation detection, can also be used for SV detection by seeksv. Currently, SV mapping of the human genome remains insufficient and lacks credibility. Only the use of short reads to detect the full range of variations is still challenging, especially in complex areas. The combination of second-generation sequencing data with third-generation sequencing data, which have the advantages of long read length but with high base-call error, would be a better strategy.

Acknowledgements

The authors thank the reviewers for their useful feedback and constructive comments. Many thanks to Zhibo Gao, Qibin Li and Zhao Lin, for they provide and support the idea. Also the authors want to thank Jiayi Du and Cheng Liang for their great help in writing the manuscript.

Funding

This work was supported by the Program for New Century Excellent Talents in University (Grant No. NCET-10-0365 to B.L.), and the National Nature Science Foundation of China (Grant Nos. 11171369, 61272395, 61370171, 61300128, 61472127, 61572178 and 61672214 to B.L.).

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium (2010a) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

1000 Genomes Project Consortium (2010b) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

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.

Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Baker,M. (2012) Structural variation: the genome's hidden architecture. *Nat. Methods*, **9**, 133–137.

Bellos,E. *et al.* (2012) cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biol.*, **13**, R120.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Chen,K. *et al.* (2014) TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res.*, **24**, 310–317.

Cheng,C. *et al.* (2016) Whole-genome sequencing reveals diverse models of structural variations in esophageal squamous cell carcinoma. *Am. J. Hum. Genet.*, **98**, 256–274.

Gao,Y.B. *et al.* (2014) Genetic landscape of esophageal squamous cell carcinoma. *Nat. Genet.*, **46**, 1097–1102.

Hormozdiari,F. *et al.* (2010) Next-generation variationhunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350–i357.

Hu,N. *et al.* (2016) Genomic landscape of somatic alterations in esophageal squamous cell carcinoma and gastric cancer. *Cancer Res.*, **76**, 1714–1723.

Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Jiang,Y. *et al.* (2012) PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, **28**, 2576–2583.

Klambauer,G. *et al.* (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.

Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.*, **10**, R23.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,S. *et al.* (2013) SOAPindel: efficient identification of indels from short paired reads. *Genome Res.*, **23**, 195–200.

Li,Y. *et al.* (2011) Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotechnol.*, **29**, 723–730.

Liu,B. *et al.* (2012) COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*, **28**, 2870–2874.

MacDonald,J.R. *et al.* (2014) The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42(Database issue)**,D986–D992.

Parikh,H. *et al.* (2016) svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*, **17**, 64.

Qi,J. and Zhao,F. (2011) inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res.*, **39(Web Server issue)**,W567–W575.

Rausch,T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.

Sindi,S. *et al.* (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**, i222–i230.

Sindi,S.S. *et al.* (2012) An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.*, **13**, R22.

Szatkiewicz,J.P. *et al.* (2013) Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation. *Nucleic Acids Res.*, **41**, 1519–1532.

Wang,J. *et al.* (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods*, **8**, 652–654.

Xie,C. and Tammi,M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

Yang,L. *et al*. (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, **153**, 919–929.

Ye,K. *et al*. (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.

Zhang,Z. *et al*. (2016) Sprites: detection of deletions from sequencing data by re-aligning split reads. *Bioinformatics*, **32**, 1788–1796.

Zhuang,J. and Weng,Z. (2015) Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes. *Nucleic Acids Res*., **43**, 8146–8156.