ATLANTIS PRESS
Advances in Engineering Research, volume 118
2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)

Integrated Sequence Assembly-based Approach for Calling Genomic Long Insertion

Lu Ye[a], Jingyang Gao[b*]

College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China

[a]email: changl_ye@163.com, [b*]email:gaojy@mail.buct.edu.cn

Keywords: high-throughput sequencing; long insertions; soft-cilpped reads; sequence assembly

Abstract. With the application and development of high-throughput sequencing technology, the detection methods of structural variants based on sequencing have emerged. However, since the high-throughput sequencing reads is relatively short compared to the previous sequencing reads, it is difficult to detect long insertion. Although assembly-based approach can solve long insertion, the computational resources used for assembly are too complex, resulting in poor results of assembly and final detection. To this end, ISALins was proposed, firstly the initial results of three different detection tools were merged; then high quality soft-cilpped reads and unmapped reads which is the set of most probable reads containing information of insertion were analyzed and extracted around the initial suspect SV breakpoints; finally these reads were assembled using assembly tool based on De Bruijn Graphs. By experimenting on both simulated and real data, we found that the method was superior to the single tool in detecting precision and sensitivity. Compared with the direct combination of call results of multiple tools, in ensuring detection sensitivity of the premise, ISALins significantly improved the detection accuracy.

Introduction

In recent years, high-throughput sequencing (HTS) has become an important technology for the determination of genomic variation. Although nucleotide level variants such as SNPs and indels are numerous, large structural variants, such as deletions, insertion, duplications and inversions, affect more sequence, and as much as 15% of the human genome falls into copy number variable regions [1]. A universal method for detecting variations is resequencing, i.e. to determine the difference of a donor individual with respect to a reference genome sequence. In the past few years, with the continuous development of sequencing technology and the sequencing of lowering cost, SV detection method based on sequencing has rapidly developed, it is mainly divided into the following four categories: Paired-end mapping, Split read, Read depth, Assembly. Insertions, for example, are detected when the distance between mapped paired-end reads is significantly longer than the average size distribution of other mapped read pairs from the same mate-pair sequencing library. Tools that use this method include BreakDancer [2] and VariationHunter [3]. Other tools such as Pindel [4] apply a split-mapping approach where one end of a pair of sequence reads is mapped uniquely to the genome and acts as an anchor, while the other end is mapped so as to detect the insertion SV breakpoint. A major drawback of these methods is the requirement that insertion should be completely contained within a read and correctly identified during the initial read mapping step. This method is effective for small insert detection, but is problematic for detecting longer than the read length. In the long insertion case, reads that support this mutation generally contain too few bases that can match to the reference gene and lead to failure; or the supporting reads may have one end map well to the reference genome but the rest of the bases after the insertion get trimmed or soft-clipped by the NGS aligner [5]. De novo assembly has been used to call insertion larger than the read length. For example, GATK HaplotypeCaller, Platypus [7] and Scalpel [8] employ localized or micro-assembly strategies and FermiKit [8] performs whole genome assembly for variant detection. Even though de novo assembly potentially can identify insertions of any size, the high computational cost and memory requirements have made it difficult

to use in practice. For these reasons we developed ISALins, many studies have shown that genetic structural variants are associated with human performance, including cancer, mental disorders, metabolic disorders, and a variety of incurable diseases, such as the heterozygous variant of gene NRXN1 associated with autism and schizophrenia; Congenital heart defects is closely related to the lack of 22q11.2 area [9]. Second-generation sequencing technology is an important step forward in our understanding of the human genetic structure and explains the relationship between genetic variation and disease. This understanding depends on our ability to accurately detect differences between individual and reference genes. Therefore, the precision and sensitivity of structural variation detection within the genome-wide range is of great significance.

Method

ISALins uses multiple SV-detection methods and tools to find a high-confidence and precise SV breakpoint callset. The novelty of ISALins lies in the combination of the following key ideas: calls reported by multiple methods are generally better quality and the filter of high quality soft-clipped reads and that local assembly can be used to call the long insertion. ISALins proceeds in the following steps (Fig. 1):

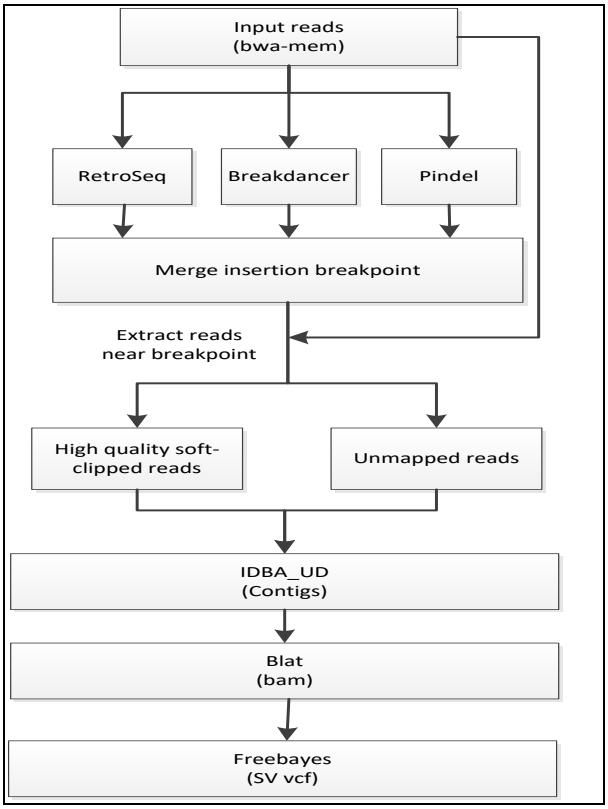

Fig.1. The workflow of ISALins

Insert site detection

We assume that the locations of the insertions are provided to us as input to the assembly method. There are two main methods for determining these locations:

1. Using existing SVs frameworks: Many efficient tools have been developed in past few years to detect the SVs efficiently and accurately [10][11]. We can use the output of their methods as the input to our algorithm.

2. Clustering the OEA Reads: OEA reads are indicator of unique insertion, we will cluster the OEA reads and pick the cluster set which has the most number of OEA reads. Clustering the OEA reads will increase our confidence level if an insertion has occurred in the donor genome. Furthermore, it will reduce the estimated number of unique insertions in the donor genome, which follows the maximum parsimony [12].

Different detection methods of structural variation have been developed, but the precision and sensitivity of each method to different structural variations are different, especially for inserts, and in order to obtain more suspicious breakpoints to achieve ultimate maximization of detection sensitivity and to ensure accuracy, this paper will use a variety of detection methods to get the initial collection of insertion breakpoint. In this paper, we use three efficient tools: BreakDancer, Pindel and RetroSeq. Firstly, we filter the detection results of each tool, that is, we extract the SV type, coordinates and size in the output of each detection tool, and then merge the detection set of the three tools through the SV type and chromosome coordinates to obtain a non-redundant SV set, that is, a collection of suspicious variant breakpoints. For example, all insertion calls from BreakDancer, Pindel, and RetroSeq were compared; if the coordinate spans from a BreakDancer insrtion call and Pindel insertion call overlapped, then the calls were merged. The insertion calls from RetroSeq were then compared to the merged BreakDancer and Pindel set to identify additional insertions not detected by either BreakDancer or Pindel.

Filter of reads for assembly

Because for long insertion, its the information of base can only be distributed in the OEA and Orphan, so we here mainly to screen the two kind of reads, as shown in Fig 2.

By mapping the fastq file of individual gene to the reference gene, we obtain the SAM file which saves the result of mapping. We classify reads into the follow four categories based on the information in SAM file:

• One-end anchored (OEA): Read-pairs in which one mate maps to the reference genome, and one does not. Soft-cilpped reads is a kind of OEA.

• Orphan: Read-pairs in which neither mate maps to the reference.

• Concordant: Read-pairs in which both reads map to the reference, and the distance between their mapped locations is within the range [m-2*v, m+2*v], m is the insert size, v is the standard deviation. Furthermore, one mate should map in the forward direction, and one in the reverse.

• Discordant: Read-pairs in which both reads map to the reference, but are not concordant.

Since sequencing reads is likely to have erroneous base or artificial data, we must filter soft-clipped reads to improve detection accuracy. Soft-clipped reads are coded as ‘S’ in their CIGAR string [13] in the BAM file. Among them, high quality soft-clipped reads are defined with the following criteria: (i) read mapping quality which denoted by MAPQ in BAM greater than a user-specified value (in practice, MAPQ ≥ 1); (ii) fraction of the soft-clipping part (in practice, ≥20 % of read length);(iii) proportion of high sequencing quality (in practice, minimum Q20) of soft-clipped bases (in practice, ≥80 %).With those filters, we try to exclude the reads that are soft-clipped due to bad sequencing quality or ambiguous alignment and only keep the reads with a long soft-clipped part that may suggest the presence of an insertion within it.

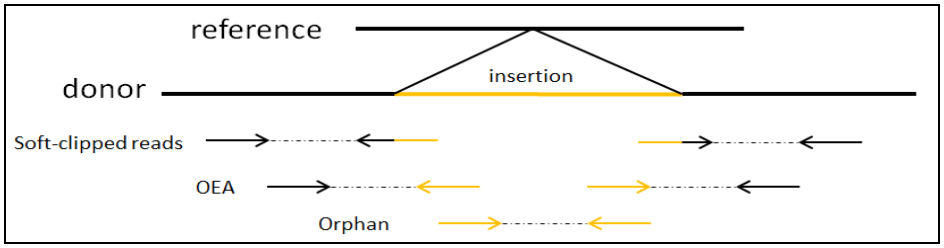

Fig.2. reads contain insertion

Local assembly and contig alignments

In the range of 1000 bp upstream and downstream of each predicted insertion breakpoint, OEA and Orphan were extracted from the BAM files, formatted to FASTA format with interleaved read pairs, and assembled by IDBA-UD [14] and the minimal contig length to be at least one base longer than the read length. Each assembled contig was then aligned against the reference genome by

applying a similar alignment procedure of short reads: BWA-MEM was used to carry out the initial alignment and soft-clipped contigs with breakpoint evidence were were re-aligned with BLAT [15] to refine their CIGAR string.

Insertion dectect

After soft-clipped read realignment and assembly, ISALins produced a BAM files which is the alignment of assembled contigs after BWA and BLAT tiered mapping. The BAM files were then sorted and indexed and passed as input to the haplotype-based variant caller, FreeBayes [16], for insertion detection.

Test results

Simulate data.

The chromosome 21 of the human reference genome hg19 was used as a reference gene. First we generated a reference file containing the SV information, which contains the SV type, location and length, for the subsequent verification. 1000 insertions were chosen where the inserted positions were not overlapping with each other. The size of these inserts range from 50bp to 1kb, the inserted sequences were randomly generated, and insertions were added to the chromosomes using svsim. We used art_illumina to simulate sequencing reads of fastq file from the generated target genome, by setting the read length to 100bp, insert size to 500, standard deviation to 50,and a coverage of 10X, 30X, and 50X, respectively. The results are shown in Table 1. From Table 1 we can see that ISALins can maintain stable performance regardless of low or high coverage data sets. For 10X, 30X, 50X coverage data, ISALins accuracy can reach almost 100%, while the accuracy is higher than the single tool.

Table 1.insertion accuracy summary for different coverage on simulate data

Coverage	Tool	Call	Benchmark	True Positives	Precision	Sensitivity
10X	Pindel	257	1000	37	0.16	0.037
	Breakdancer	149		61	0.409	0.061
	Restrseq	201		156	0.776	0.156
	ISALins	151		151	1.000	0.151
30X	Pindel	708	1000	49	0.079	0.049
	Breakdancer	217		77	0.355	0.077
	Restrseq	431		399	0.926	0.399
	ISALins	430		430	1.000	0.430
50X	Pindel	1169	1000	68	0.069	0.068
	Breakdancer	243		80	0.329	0.080
	Restrseq	508		469	0.923	0.469
	ISALins	470		470	1.000	0.470

Real data.

Human NITS standard NA12878 was used to validate ISALins on whole genome sequencing (WGS) data. Raw fastq files were obtained from European Nucleotide Archives with the accession number ERA172924.Paired-end reads were aligned to the hg19 human reference using BWA-MEM with default parameters. We split the BAM file containing all reads by chromosome. Each smaller BAM file contains all mapped reads from only one chromosome and all unmapped reads without any mapping information in a large BAM file. All programs were called for each individual BAM file separately and predictions of each chromosome were merged into one final output file. The large novel sequence insertion reference call set was obtained by extracting Cortex identified NA12878 sites from the 1000 Genomes Pilot 1 novel sequences file. The results are shown in Table 2. From Table 1 we can see that ISALins achieved highest sensitivity and precision, which were

92.3% and 45.7% respectively, comparing to all the individual tools analyzed.

Table 2.insertion accuracy summary for 50X coverage on real data

Coverage	Tool	Call	Benchmark	True Positives	Precision	Sensitivity
50X	Pindel	2791	105	28	0.010	0.267
	Breakdancer	0		0	0	0
	RetroSeq	60		43	0.717	0.410
	ISALins	52		48	0.923	0.457

Conclusion

ISALins significantly improves the accuracy of calling long insertion by integrating multiple tools and the analysis of high quality soft-clipped reads compared with the state of the art tools. We consider ISALins as a proof of concept of the effectiveness of using an ensemble approach for calling SVs. The approach is not limited to only using the aforementioned tools, it can be easily adapted to use additional or even a different set of tools.

Acknowledgement

In this paper, the research was sponsored by the National Natural Science Foundation of China (Project No. 61472026).

References

[1] Stankiewicz PC, Lupski JR: Structural variation in the human genome and its role in disease. Annu Rev Med 2010, 61:437-455.

[2] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS,McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ,Wiilson RK, Ding L, Mardis ER: BreakDancer: an algorithm for highresolutionmapping of genomic structural variation. Nat Methods 2009,6:677-681.

[3] Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC: Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 2010, 26:i350-357.

[4] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009, 25:2865-2871.

[5] Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches. Cancer Genet. 2013; 206:432–40.

[6] Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2013; 2014:1–9.

[7] Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee Y, Wang Z, et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nat Methods. 2014;11:1033–36.

[8] Li H. FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics. 2015;31:3694–6.

[9] Lee M Y, Won H S, Baek J W, et al. Variety of prenatally diagnosed congenital heart disease in 22q11. 2 deletion syndrome [J]. Obstetrics & gynecology science, 2014, 57(1): 11-16

[10] Hormozdiari F, et al: Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Research 2009, 19(7):1270-1278.

[11] Medvedev P, et al: Computational methods for discovering structural variation with next-generation sequencing. Nature Methods 2009, 6(11 Suppl):S13-S20.

[12] Hajirasouliha I, et al: Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 2010, 26(10):1277-1283.

[13] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

[14] Peng Y, Leung HC, Yiu S, Chin FY. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):1420–8.

[15] Kent WJ. BLAT - The BLAST-like alignment tool. Genome Res. 2002;12:656–64.

[16] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv Prepr arXiv. 2012;1207:3907.