Research on the Detection Method of Structural Variation based on Next-Generation Sequencing Data

Hai Yang

School of Computer Science and Technology, Shandong University, Qingdao, 266237, Shandong, China

Corresponding author e-mail: yh_sdjtu@163.com

Keywords: detection method; structural variation; Next-Generation Sequencing; high-throughput sequencing; evaluating method

Abstract: The detection method of structural variation is one of the most important research field of bioinformatics. In this paper, next-generation sequencing and high-throughput sequencing technoloy are introduced firstly. Next, the types of genome structural variation such as insertion, deletion, duplication, copy-number variant, inversion and translocation are elaborated. Thirdly, four main detection methods of SV are illustrated in detail including paired-end, read-depth, split-read and assembly. Finally, this paper propose several parameters and a evaluating method of structural variation detection.

1. Introduction

High-throughput DNA Sequencing Data is one of the most important characters of Next-Generation Sequencing (NGS). The birth of high-throughput DNA sequencing technology is a technical revolution in molecular biology field. In contrast to the first generation of sequencing, NGS technology is typically characterized by being highly scalable, allowing the entire genome to be sequenced at once [1]. Furthermore, high-throughput sequencing (HTS) technology also has several advantages such as low-cost, fast, and high-throughput. So that HTS has replaced the traditional sequencing technology.

Usually, it is accomplished by fragmenting the genome into small pieces, randomly sampling for a fragment, and sequencing it by using one algorithm or one combinational algorithm. An entire genome is possible because multiple fragments are sequenced at once in an automated process.

There are several following primary methods when detecting structural variation (SV) by using high-throughput sequencing data. Firstly, detecting SV only using the coverage information which is the earliest method for SV but not used exclusively now. Secondly, detecting SV mainly using the discordant pair in the paired end sequencing data to find structural variation [2]. This kind of method also relies on clustering, but can not find the loci information of SV. Thirdly, the split reads are introduced into the detecting structural variation so that the precision of detecting SV is the most highest one in these three methods. In order to improve the performance of the detection method of SV, researchers usually use the integrated method of above ones.

2. High-Throughput Sequencing

The whole human genome sequencing technology is a key precondition for genomics and bioinformatics. The most well-known and widely used sequencing method in the traditional sense is the Sanger sequencing method, which originated in the 1970s and has been gradually improved. Moreover, the first whole human genome sequencing data has been gotten by using Sanger in 2001. However, the advance in human genomics was a joint effort by a number of research institutes all over the world and it took a lot of time and money.

The need for low-cost and faster genome sequencing technologies is growing higher and higher, the high-throughput sequencing technology is proposed at the right moment. The advent of high-throughput sequencing has greatly reduced the time and cost of whole genome sequencing. At

present most of the original sequencing data were obtained by using the Illumina instrument during detecting structural variation.

There are mainly two kinds of sequencing data generated from Illumina instrument: single end and pair end sequencing data. Especially the pair sequencing data not only includes the information of reads, but also the information of insert size. So that the information of reads and insert size provides new evidence when determining the relative location of two pair-end reads.

Usually, when dealing with the single end reads, the read information and read coverage are mainly used to detecting the structural variations. When dealing with the pair end reads, not only the information same as single end reads, but also the information of insert size can be used. So the efficiency of variation detection is much higher. Nowadays, the pair end reads are used more than the single end reads when detecting structural variation.

3. The Types of Genome Structural Variation

With the progress of human genome sequencing technology, the whole genome data has been increasing in vast scale day by day. Even though two human individuals are the same sex among different races and ethnicities, the difference between their genome is fairly small. But because the number of base pairs in the whole human genome is up to 3 billion, there are many structural variations among those huge scale of base pairs, which lead to the differences between human individuals.

Based on the above background, the study of structural variations is of vital and far-reaching significance in bioinformatics, diseases and medicine fields. The information of differences between reference sequence and testing sequence mainly obtained by genome alignment. The obtained information of differences usually is divided into two kinds: one kind is called single nucleotide polymorphisms (SNP) [3]. A SNP is a variation in just one nucleotide of a genetic sequence; think of it as a spelling change affecting just one letter in an uncommonly long word. The other kind is call genomic structural variation, which is also called structural variation (SV) for short. Structural variation is the variation in structure of an organism's chromosome. It consists of many kinds of variation in the genome of one species, and usually includes microscopic and submicroscopic types, such as insertions, deletions, duplications, copy-number variants, inversions, translocations and so on. For example, the insertion structural variation and deletion structural variation are shown in figure 1 and figure 2:

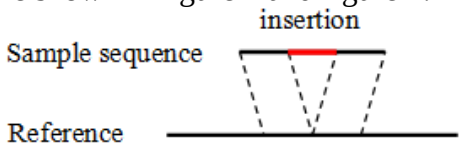

Figure 1. The sketch of insertion.

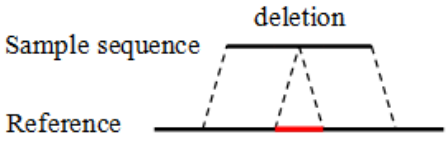

Figure 2. The sketch of deletion.

4. The Main Detection Methods of Structural Variation

Compared to large numbers of DNA sequence alignment methods, the detection methods of Structural Variation are relatively few. Two major reasons for this status quo of SV detection methods are as follows: the study of SV detection method started late and genome structural variations are much more complex than SNP and DNA alignment.

4.1. Paired-end Mapping

Paired-end Mapping is also called Read-pair. First of all, a large number of paired-end reads of individuals are obtained by using high-throughput sequencing technology. In the meantime, the genome information which includes base compositions, read length, the distance of two reads and so on, is recorded. Then these paired-end reads are aligned onto reference sequence by using mapping tools such as BWT, BWA and MAQ etc [4], to obtain the alignment information includes mapping distance and mapping orientation. Finally the structural variations will be obtained by comparing the mapping information and genome information.

If the distance between two reads from the same pair which are aligned onto the reference sequence equals to the distance in the library plus the length of the missing sequence, there is a deletion variation in the genome segment, shown in figure 3 as follows. If the distance between two reads from the same pair which are aligned onto the reference sequence equals to the distance in the library minus the length of the inserted sequence, there is a insertion variation in the genome segment, shown in figure 4 as follows. After aligning two reads in the same pair onto the reference sequence, if one read is aligned in the same direction with the reference sequence and the other read is aligned in the opposite direction with the reference sequence, there is a inversion variation in the genome segment, shown in figure 5. In principle, the paired-end mapping method can detect most kinds of the genome structural variations.

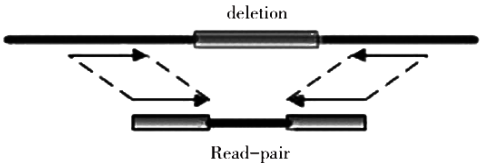

Figure 3. The deletion variation.

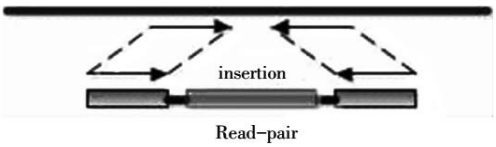

Figure 4. The insertion variation.

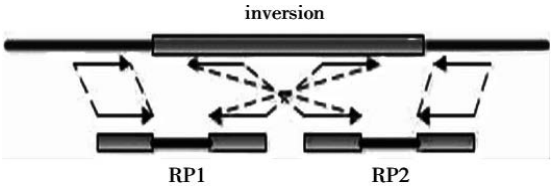

Figure 5. The inversion variation.

4.2. Read-depth Method

Large number of reads can be obtained by using the high-throughput sequencing technology. The length and base composition of these reads are also known. Align these reads onto the reference sequence. If the number of reads in a certain region is much more or less than other regions, there is a duplication or deletion variation with a very high probability, which are shown as figure 6 and figure 7.

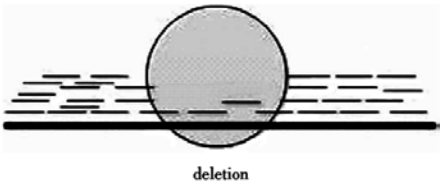

Figure 6. The inversion variation.

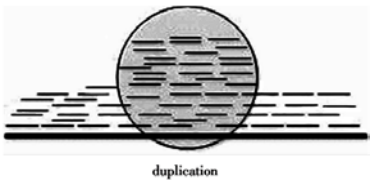

Figure 7. The inversion variation.

4.3. Split-read Method

If one of the two reads in the same pair couldn't be aligned onto the reference sequence, it is called discordant read. In order to mapping onto reference sequence, the discordant read will be split into smaller reads by the breakpoints [5]. Figure 8, figure 9 and figure 10 demonstrate the deletion, insertion and inversion variation.

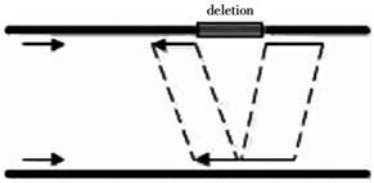

Figure 8. Deletion.

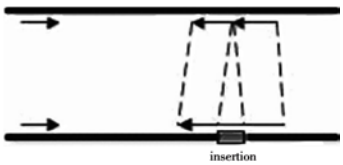

Figure 9. Insertion.

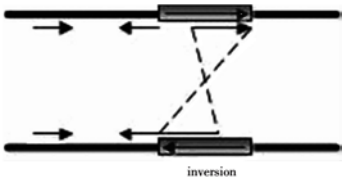

Figure 10. Inversion.

4.4. Assembly Method

By assembling the genome fragments which are obtained by sequencing, the contigs or scaffolds will be formed. Then the contigs or scaffords are aligned onto the reference sequence to find out the

structural variations. In theory, all kinds of structural variations could be detected if the genome fragments obtained by sequencing are accurate enough to make the original assembly algorithm effective. The detection of deletion, insertion and inversion variations by assembly method are shown in figure 11, figure 12 and figure 13.

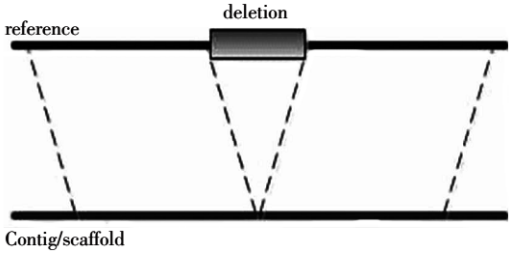

Figure 11. The inversion variation.

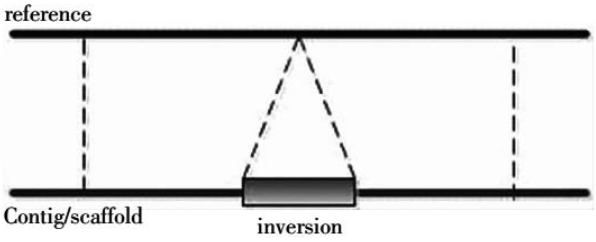

Figure 12. The inversion variation.

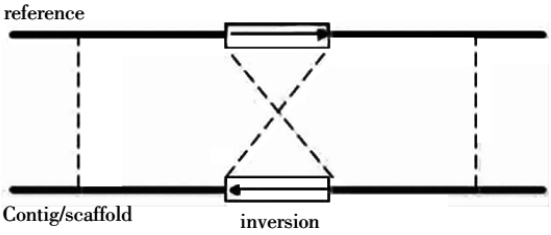

Figure 13. The inversion variation.

5. Evaluating Method of Structural Variation Detection

Denote the set which is composed of the real structural variations as AS, named answer set. Denote the set which is composed of the structural variations obtained by detection methods as RS, named result set. Denote the set which is composed of the structural variations that are both thought to exist and detected as CS, named common set. Then the recall rate and accuracy rate are described as formula 1 and formula 2.

$$recall\_rate = CS / AS \tag{1}$$

$$accuracy\_rate = CS / RS \tag{2}$$

We take 40x coverage on the chr1 chromosome as the simulated data. Then the detection of deletion and insertion variation are carried by using PRISM and Pindel.

Step 1: Load the data of chr1 chromosome from hg18.

Step 2: Inject structural variations of Ventor into the reference sequence and record the information of insertion.

Step 3: Generate the homozygous and heterozygous sequence, then generate the answer information about structural variations.

Step 4: Generate the 40x reads using wgsim or art software.

Step 5: Gain the fasta files of the homozygous and heterozygous reads.

Step 6: Generate the alignment information by using the mapping software.

Step 7: Generate the sam files of the homozygous and heterozygous reads.

Step 8: Generate the related coordinate files of fasta files.

Step 9: Generate the answer set according to the coordinates, reference sequence and so on.

Step 10: Generate the results by Delly, Pindel or PRISM, and then take alignment between result set and answer set.

Step 11: Obtain the final evaluating results.

6. Conclusion

With the rapid development of high-throughput sequencing technology, the research on genome structural variation is entering into the post-genome era. The detection method of SV is becoming

particularly important. In this paper, next-generation sequencing, high-throughput sequencing technoloy and the types of structural variation are introduced. And then, four main detection methods of SV are illustrated in detail including paired-end, read-depth, split-read and assembly. Finally, several parameters and formulas are proposed for a new evaluating method of structural variation detection.

References

[1] SCHUSTER S C. Next-generation sequencing transforms today's biology [J]. Nature Methods, 2008, 5(1): 16-18.

[2] Claudia B C, James R L. Mechanisms underlying structural variation formation in genomic disorders [J]. Nature Review Genetics, 2016, 17(4): 224-238.

[3] Sebat J, Lakshmi B, Malhotra D et al. Strong association of De Novo copy number mutations with autism. Science, 2007, 316(5823): 445-449.

[4] RAUSCH T, ZICHNER T, SCHLATTL A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis [J]. Bioinformatics, 2012, 28(18): 333-339.

[5] ZHANG J. et al. SVseq: an approach for detecting exact breakpoints of deletions with low-coverage sequence data. Bioinformatics, 2011, 27: 3228-3234.