Detecting a Portion of Structural Variation of Hedou12 by Using Some Characteristics of the Next Sequencing Technology

Hui-Qiang Jia¹, Hai-Chao Wei², Da-Ming Zhu¹, Xian-Zhong Feng², Feng-Hua Liu¹, Jing-Jing Ma¹, Rui-Zhi Wang¹, Cheng Chang¹ and Hong-Shuang Li¹

¹Intelligent Algorithm and Software Group, Computer Science and Technology of ShanDong University, 1500 ShunHuaRd, JiNan City, ShanDong Province, China
²Key Laboratory of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology

Keywords: Structural variation, Discordant paired-end read, Split-read.

Abstract. At present, the computable methods which detect structural variations (SVs) of genome have numerous been used in human genome, but are few in plant genome. If we use these methods for plant, many structural variations might not be found. Although we can find some structural variations, the valid SVs in them are very few. To enhance precision of detecting structural variations and reduce the false discovery rate in plant, we design an algorithm based on characteristics of the Hedou12 which is one of the important economic corps. The core of the algorithm composes discordant paired-end read analysis model and split-read analysis model, and set cover based on greedy heuristic method. To increase precision of detecting SVs, we separately apply the probability model of discordant paired-end reads and the probability model of split-reads as optimization selection for SVs. And by validity verification, our algorithm can increase precision of detecting SVs.

Introduction

The SVs of the genome are very important variations in genome and generally include PAV (presence/absence variants) such as large insertions and deletions, and CNV (copy number variance) such as translocations, inversions and tandem duplications [1]. These SVs can change the function or the regulation of genes and also cause various diseases [2], e.g., autism [3], Parkinson’s disease [4], schizophrenia [5] and cancer [6], so it has become increasingly important to apply which method to detect them. Presently, the method of detecting these SVs has two classes. The first class is array comparative genome hybridization (aCGH) [7,8], but it can’t detect reconstruct locus architecture or balanced rearrangements and segmental duplications (SDs) or transposable elements (TEs). The second class is sequencing-based method [9,10,11,12,13], such as paired-end mapping (PEM), which has emerged as a potent alternative to aCGH, but the limit of its practical utility, such as BreakDancer [14], Breakpoints [15], CLEVER [16], GASVPro [17], SVMerge [18], DELLY [19] and LUMPY [20], is the high cost of “long-read” sequencing and the computational difficulties associated with interpreting “short-read” sequence data from complex genomes. This causes we prefer the sequencing-based method. But, presently many sequencing-based methods are used to detect SVs in human genome, for plant genome these methods are few. If we directly apply these methods to check the SVs in plant genome, many SVs may not be found. The main reason maybe the enormously variable regions in plant genome, such as chromosome number, chromosome size [21] and the degree of gene clustering. To reduce the false rate of detecting SVs in the duplication regions of plant genome, we design a new algorithm based on some characteristics of sequencing data of Hedou12. This algorithm comprises these characteristics of paired-end reads and set cover algorithm (PSCA). PSCA mainly applies discordant paired-end reads and split-reads of these paired-end reads to design two classes of cluster algorithms and two probabilities, and also uses these two probabilities as optimization selection for SVs. These two probabilities will help us to improve the precision of detecting the SVs (precision). And to reduce the runtime time, we apply
the set cover algorithm which is designed by Graham Cormode et al. [22] To validate our algorithm, we separately compare DELLY [19] and LUMPY [20], which both incorporate discordant paired-end reads and split-reads to predict SVs of the genome, with PSCA. Through validated verification, PSCA had higher sensitivity and precision.

Methods

PSCA contains three main parts. The first part is that align the reads of the Hedou12 to Williams82 by BWA [23]. After finishing the alignment, PSCA applies mrsFast version3.3.0 short-read alignment tool [24] to produce a file in bam format, which contains all locations and orientations of all discordant paired-end and split-read on Williams82. The second part is that PSCA uses different characteristic of discordant paired-end reads and split-reads on Williams82 to cluster different SVs. And combining some characteristics of paired-end reads which the next sequencing technology generate, PSCA separately designs analysis models for discordant paired-end reads and split-reads. The third part is that we adopt the set cover which is based on greedy heuristic method, to find the most uncovered element sets in initiate set. The purpose of this set cover algorithm is to reduce the runtime and increase efficiency of PSCA. Therefore, the key point of PSCA comprises discordant paired-end read analysis model, split-read analysis model and the set cover algorithm.

For discordant paired-end read analysis model and split-read model, we separately design two cluster algorithms and two probabilities to reduce the false positive of the found SVs of repeat regions on Williams82. These two probabilities are as follows separately:

\[
P(\text{PE}_{\text{Clu}_i}) = \left(\prod_{i=1}^{\text{m}} P(PR_i \in \text{PE}_{\text{Clu}_i})\right) \left(\sum_{i=1}^{\text{m}} \frac{1}{\text{Seq}(PR_i)}\right)
\]

\[
= \left(\prod_{i=1}^{\text{m}} P(PR_i \in \text{PE}_{\text{Clu}_i})P(\text{PE}_{\text{Clu}_i})\right) \left(\sum_{i=1}^{\text{m}} \frac{1}{\text{Seq}(PR_i)}\right)
\]

\[
P(\text{SR}_{\text{Clu}_i}) = \sum_{i=1}^{\text{m}} \frac{1}{\text{Seq}(PR_i)} \sum_{X \subseteq \text{SR}, |X| = m} \left(1 - \frac{1}{10^\sum_{i=1}^{\text{m}} q(i)}\right)
\]

\[
= \sum_{i=1}^{\text{m}} \frac{1}{\text{Seq}(PR_i)} \sum_{X \subseteq \text{SR}, |X| = m} \left(1 - \frac{1}{10^\sum_{i=1}^{\text{m}} q(i)}\right)
\]

\[
\text{Seq}(PR_i) \text{ is the sequence similarity score of } PR_i \text{ that is aligned to the reference. The N is the number of the discordant paired-end read contained in cluster } Clu_i.
\]

The set cover algorithm separately adopts these two probabilities as an optimization selection for the SVs. In here, we require that each probability is more than or equal 0.3. The PSCA framework is as follows.

1 BWA aligner aligns the Hedou12 to Williams82 and produce paired-end reads file in sam format, mrsFast generate split-reads file in sam format.

2 extract all these locations on the Williams82 and generate U={Pe1, Pe2, Pe3,...., Pen,Sp1,Sp2,..., Spn}

3 separately use two different probability models to give each set of the discordant paired-end reads and each set of the split-reads to assign value. And select these sets that the probability is more than or equal 0.3, finally generate two initial set DisS and Sp.

4 merge Sp and DisS, and produce S=Sp∪DisS

5 implement the greedy heuristic set weight cover algorithm

5.1 C ← ∅

5.2 while C ≠ U do

5.3 if P^k ≤ |S_i| < P^{k+1}, assign S_i to subcollection S^{(k)}

5.4 for k → K down to 1
For each set $S_i$ in $S^K$

If $|S_i - C| \geq P^K$ and Wmin: add i to ID and update C

ELSE: let set $S_i \leftarrow S_i - C$ and add the updated set to subcollection $S^{(k')}$, where the new set sizes satisfies $P^{k'} \leq |S_i| < P^{k'+1}$ and $k' < k$

For each set $S_i$ in $S^{(0)}$

If $|S_i - C| = 1$ and Wmin: add i to ID and update C

6 output: Set cover C of detecting the best and minimal set of read alignment from S

P is the number of elements in one set, the value of p is 2, which is that we require that at least two read pairs support a SV. K is the largest k with non-empty $S^{(k)}$.

Results and Discussion

We tested PSCA using Hedou12 and Williams82. And then we separately compared PSCA with LUMPY and DELLY. Through these comparisons (Table 1: compare PSCA with LUMPY and DELLY separately), PSCA was feasible.

Table 1. compare PSCA with LUMPY and DELLY separately.

<table>
<thead>
<tr>
<th></th>
<th>PSCA total</th>
<th>PSCA valid SV</th>
<th>LUMPY total</th>
<th>LUMPY valid SV</th>
<th>DELLY total</th>
<th>DELLY valid SV</th>
</tr>
</thead>
<tbody>
<tr>
<td>deletion</td>
<td>9781</td>
<td>6328</td>
<td>9775</td>
<td>6182</td>
<td>9776</td>
<td>6305</td>
</tr>
<tr>
<td>tandem duplication</td>
<td>4179</td>
<td>1863</td>
<td>4175</td>
<td>1803</td>
<td>4176</td>
<td>1853</td>
</tr>
<tr>
<td>insertion</td>
<td>1329</td>
<td>854</td>
<td>1324</td>
<td>837</td>
<td>1324</td>
<td>848</td>
</tr>
<tr>
<td>inversion</td>
<td>1142</td>
<td>533</td>
<td>1139</td>
<td>513</td>
<td>1140</td>
<td>530</td>
</tr>
</tbody>
</table>

Conclusion

By testing from sensitivity and precision, PSCA was efficient and reliable. However, we also must point out that due to the own complexity of Hedou12 genome [28] and the sequencing error, PSCA still can’t obtain all validated SVs. This may also be due to aligning artifacts in repeat regions, and sequencing error related to Illumina technology such as GC-rich regions. Except these aligning issues, PSCA also has:

- two probabilities only consider the relation between the aligned distance and the insert size and between sequencing base and sequencing similarity score and between both segments of the split-reads, and PSCA only solve some false positives of repeat regions due to existing vast inversion or everted regions in Hedou12 genome.

In the future, we will further improve PSCA to make it detect more SVs and more precision SV breakpoints. And we will also find more precise probability model from the analysis of the Hedou12 genome so that we can increase the sensitivity and precision of PSCA.

Acknowledgments

We thank all authors in this paper. All contribution is the same.

References


