Bioinformatics, 33(17), 2017, 2750–2752
doi: 10.1093/bioinformatics/btx279
Advance Access Publication Date: 8 May 2017
Applications Note

Sequence analysis

rMFilter: acceleration of long read-based structure variation calling by chimeric read filtering

Bo Liu[1,†], Tao Jiang[1,†], S. M. Yiu[2], Junyi Li[1] and Yadong Wang[1,*]

[1]Center for Bioinformatics, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China and [2]Department of Computer Science, The University of Hong Kong, Hong Kong, China

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Inanc Birol

Received on June 28, 2016; revised on March 20, 2017; editorial decision on April 24, 2017; accepted on May 5, 2017

Abstract

Motivation: Long read sequencing technologies provide new opportunities to investigate genome structural variations (SVs) more accurately. However, the state-of-the-art SV calling pipelines are computational intensive and the applications of long reads are restricted.

Results: We propose a local region match-based filter (rMFilter) to efficiently nail down chimeric noisy long reads based on short token matches within local genomic regions. rMFilter is able to substantially accelerate long read-based SV calling pipelines without loss of effectiveness. It can be easily integrated into current long read-based pipelines to facilitate SV studies.

Availability and implementation: The C++ source code of rMFilter is available at https://github.com/hitbc/rMFilter.

Contact: ydwang@hit.edu.cn

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

1 Introduction

Long read sequencing technology such as Single Molecule Real-Time (SMRT) sequencing benefits the studies of genome structure variations (SVs) (Chaisson *et al.*, 2015; Sudmant *et al.*, 2015). However, such studies rely on the alignment of the noisy long reads (Chaisson *et al.*, 2015; English *et al.*, 2014), which is computational intensive (Chaisson and Tesler, 2012). Moreover, it is expensive to handle a large amount of aligned reads in downstream analysis. Herein, we propose a local region match-based filter (rMFilter), to identify chimeric noisy long reads, which likely span the breakpoints of SVs. rMFilter can be used as a pre-processing step to reduce the number of reads to be analyzed. It can help to speed up SV calling pipelines without loss of effectiveness.

2 Materials and methods

Previous study (Chaisson and Tesler, 2012) suggests the number of short token matches between a given read and its true site in reference follows a geometric distribution, given a certain sequencing error rate. We then assume that, when a read is correctly mapped to a local region while the number of local matches is significantly different from those of other reads, this read is partially aligned and tends to span some SV breakpoint(s). Under this assumption, rMFilter processes each read in two steps as followings (refer to Supplementary Notes for more implementation details).

1) For each read, rMFilter employs Regional Hash Table (RHT) index (Liu *et al.*, 2016) to locate a local region which has the highest number of short token (*k*-mer) matches. If the density of the local matches (i.e. the number of the matches divided by the length of the region) is too low, rMFilter will consider the read as a ‘junk read’ with serious sequencing errors and directly discard it. For all the reads not marked as ‘junk’, rMFilter chooses a proportion (5% in default) of them at two-tails as outliers and marks them as ‘chimeric reads’.

2) For each of the remaining reads (i.e. neither ‘junk’ nor ‘chimeric’), rMFilter performs a sparse dynamic programming (SDP) on the maximal exact matches between the read and the local region to investigate whether a skeleton of end-to-end alignment can be built. The reads that rMFilter fails to build the skeleton are also marked as chimeric reads. This step aims to reduce false negatives, since some issues such as small size of the SV event(s) or exceptional sequencing

Table 1. The comparison between the pipelines w/o rMFilter

Benchmarks	Without rMFilter	With rMFilter
Time cost of the various steps (in minutes)		
Read filtering of rMFilter	–	735
Read alignment (step 1)	5408	1110
SV sites inference (step 2)	2729	1338
Local assembly-based SV calling (step 3)	17 473	12 077
Effects on the various steps		
# of reads input to Step 1	24 703 987	11 721 367
# of validated putative SV sites of Step 2	12 731	12 602
# of validated SVs of Step 3	1079	1063

The time of 'Read filtering of rMFilter' includes the time (82 minutes) of transforming the reads from bas.h5 format to fastq format with a Python script (Bash5tools.py) provided by PacBio. The time of 'Local assembly-based SV calling (step 3)' is only for the time of local assembly on Chromosome 1 of human reference genome. '# of input reads to step 1' are the number of reads input into step 1 of the pipelines w/o rMFilter. '# of validated putative SV sites of step 2' and '# of validated SVs of step 3' are respectively the numbers of the putative SV sites and the SV calls output by steps 2 and 3, which can be validated by the callset of (Chaisson et al., 2015) with 1bp overlapping criterion.

quality may also cause a SV-spanning read to have a high number of short token matches.

3 Results

One of the state-of-the-art SMRT read-based SV calling pipelines was proposed (Chaisson *et al.*, 2015) to comprehensively analyze the SVs of a *H. sapiens* sample (the CHM1htert cell line) with a 54× coverage SMRT P5/C3 release dataset. This pipeline calls SVs in three steps: (1) aligning the reads to reference genome, by a modified version of BLASR (Chaisson and Tesler, 2012); (2) inferring putative sites of SVs by clustering and analyzing chimerically aligned reads; (3) calling SVs by local assembly around the putative SV sites. We integrated rMFilter into this pipeline by using the result of rMFilter as input of pipeline to assess its read filtering ability. (More detailed information on the implementation of the benchmarking is in Supplementary Notes.)

A major concern of rMFilter is whether it could mistakenly filter SV-spanning reads out and cause the loss of the evidence of SVs. This loss of SVs is most related to the first two steps of the pipeline, because of that the first two steps aim at generating and analyzing the alignments of reads to find the evidence to infer SV sites, and Step 3 emphasizes more on refining the result of Step 2 to produce a higher resolution SV callset. Therefore, we focused more on the comparison between the putative SV sites along the reference genome (the output of Step 2) with and without (w/o) rMFilter. Furthermore, we executed the pipeline to perform local assembly (Step 3) for Chromosome 1 to investigate the overall effect of rMFilter.

Three major results were observed from this benchmark experiment.

First, rMFilter reduces the time of all the steps of the pipeline (Table 1) because of effective elimination of false positive inputs (>50% false positive reads were filtered out). Step 1 (the alignment step) gains the most benefit from this. Considering the slow speed of local assembly, the absolute speed gain of Step 3 is also impressive.

Second, the pipeline maintains the same level of sensitivity when integrated with rMFilter. We investigated the putative SV sites of the whole genome (the output of Step 2) and the SV calls of Chromosome 1 (the output of Step 3). It is observed that, w/o rMFilter, the pipeline produced very similar numbers of putative SV sites and SV calls which can be validated by the SV callset of (Chaisson *et al.*, 2015) (Table 1, Supplementary Table S1). Note that the pipeline requires a certain number (at least 5 in default) of supporting reads to report a putative SV site or a SV call. Having similar numbers of validated SV-sites/SV-calls (w/o rMFilter) suggests that rMFilter does not filter out many useful reads and the remaining reads are able to provide strong SV signals (i.e. enough supporting reads) to keep the sensitivity of the pipeline.

Third, the 'chimeric' reads recognized by rMFilter show strong SV evidence. We investigated the alignment result of BLASR for all the chimeric reads and found that 96% of the reads have the characteristics of SV-spanning read: 34% of them were unaligned, 33% of them have large clippings (>20% of the read length) and 29% of them have at least one large indel (>50 bp) in the alignment. This indicates that the filtering of rMFilter is appropriate.

We also investigated the SVs recorded in the callset of (Chaisson *et al.*, 2015) which are only recovered by the pipeline without rMFilter (Supplementary Table S2). Result shows that rMFilter failed mainly at small SVs (around 50 bp). Due to the serious indel errors of SMRT sequencing, rMFilter mistakenly filtered the SV-spanning reads out with false negative alignment skeletons. Moreover, for a portion of the false negatives, it is observed that the reads around the putative SV sites (the inputs to Step 3) are similar w/o rMFilter while the SVs were discovered only without rMFilter. These failures could also result from the design and implementation of the pipeline.

Three simulated SMRT datasets respectively from *in silico* Chromosome 1, 14 and 22 of human genome (corresponding mean read lengths: 8000, 7500 and 7000 bp) were used to assess the ability of rMFilter on read level. It is observed that rMFilter can recognize proximately 84% of grand truth SV-spanning reads as chimeric reads (Supplementary Table S3). Most of the false negatives are caused by the fact that the reads having very little overlap with the SV events and the reads in the SV regions can be consecutively aligned to the reference. For these reads, the alignment tool (BLASR) also cannot produce alignments with strong SV evidence (Supplementary Notes). Under such circumstance, the pipeline with rMFilter would not be seriously affected. This is also demonstrated by the similar numbers of grand truth SVs being recovered w/o rMFilter (Supplementary Table S4). The only exception is that, rMFilter could still wrongly recognize a few reads spanning some small SVs (round 50 bp) as normal reads, like that of the real dataset, thus a few small SVs were missed (Supplementary Table S5).

In conclusion, rMFilter can effectively filter chimeric reads and speed up the SV calling process without decreasing its sensitivity. It is a useful tool for the analysis of noisy long reads and benefit cutting-edge genomic studies. Please also refer to Supplementary Notes for more detailed results, discussions and examples (especially for the failures) on the real and the simulation studies (the 'Discussion on the ability of rMFilter' part).

Funding

This work has been partially supported by the National Nature Science Foundation of China (Nos: 61301204), the National High-Tech Research and Development Program (863) of China (Nos: 2015AA020101).

Conflict of Interest: none declared.

References

Chaisson,M.J. and Tesler,G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.

Chaisson,M.J. *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608–611.

English,A.C. *et al.* (2014) PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, **15**, 180.

Liu,B. *et al.* (2016) rHAT: fast alignment of noisy long reads with regional hashing. *Bioinformatics*, **32**, 1625–1631.

Sudmant,P.H. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.