BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 1 2015, pages 116–118
doi:10.1093/bioinformatics/btu593

Sequence analysis Advance Access publication September 4, 2014

Genomon ITDetector: a tool for somatic internal tandem duplication detection from cancer genome sequencing data

Kenichi Chiba[1,*,†], Yuichi Shiraishi[1,*,†], Yasunobu Nagata[2], Kenichi Yoshida[2], Seiya Imoto[1], Seishi Ogawa[2] and Satoru Miyano[1]

[1]Laboratory of DNA Information Analysis, Human Genome Center, the Institute of Medical Science, The University of Tokyo, Tokyo 108-8639 and [2]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan

Associate Editor: Michael Brudno

ABSTRACT

Summary: Somatic internal tandem duplications (ITDs) are known to play important roles in cancer pathogenesis. Although recent advances in high-throughput sequencing technologies have enabled genome-wide detection of various types of genomic mutations, including single nucleotide variants, indels and structural variations, only a few studies have focused on ITDs. We have developed an analytical tool called 'Genomon ITDetector' for genome-wide detection of somatic ITDs. After evaluating the sensitivity and precision of the proposed approach using synthetic data, we have demonstrated that it can successfully detect not only common ITDs involving *FLT3,* but also a number of ITDs affecting other putative driver genes in acute myeloid leukemia exome sequencing data.

Availability and implementaion: Genomon ITDetector is freely available at https://github.com/ken0-1n/Genomon-ITDetector

Contact: kchiba@hgc.jp or yshira@hgc.jp

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

Received on November 13, 2013; revised on August 19, 2014; accepted on August 28, 2014

1 INTRODUCTION

Recent advances in high-throughput sequencing technologies have enabled genome-wide detection of various genomic alterations in cancer cells, and a number of analytical tools have been developed for detecting single nucleotide variants, short indels (Shiraishi *et al*., 2013) and structural variations (SVs) from cancer genome sequencing data. On the other hand, it is widely known that an internal tandem duplication (ITD) that involves several tens to several hundreds of nucleotides represents a common type of somatic alteration that plays an important role in cancer pathogenesis through either activating or inactivating affected genes. In particular, ITDs involving FMS-like tyrosine kinase 3 (*FLT3*) are among the most frequent genetic lesions, and are found in ∼30% of patients with acute myeloid leukemia (AML), leading to aberrant activation of the kinase and negatively affecting survival (Kindler *et al*., 2010). However, despite their potential importance in cancer development, somatic ITDs have been poorly focused because most current approaches are not sensitive enough to identify indels of more than several tens of base pairs or SVs within small-scale regions (<1000 bp), rendering ITDs as 'blind spots' for most existing analytical methods. In this article, we present a novel software application, Genomon ITDetector, which is specifically designed for performing sensitive and genome-wide detection of somatic ITDs spanning several tens to several hundreds of base pairs from cancer genome sequencing data.

2 METHODS

When short reads at the positions of ITDs are aligned using tools such as BWA (Li and Durbin, 2009), which support partial alignment, they show characteristic patterns: soft-clipped sequences (i.e. an unmatched fragment in a partially mapped read, see Fig. 1a) are observed outside the breakpoint pairs, corresponding to the sequences inside the breakpoint pairs (which we call ITD breakpoint pairs, ITD-BPPs, see Fig. 1b). The basic strategy for detecting somatic ITDs using Genomon ITDetector is to initially perform genome-wide identification of ITD-BPPs in the target tumor sample, and then filter out uncertain candidates to remove false positives because of the mis-alignments caused by redundancy of the genome, relying on multiple non-matched control samples (Supplementary Fig. S1). Briefly, the procedure adopted in Genomon ITDetector after sequencing alignment is as follows (see Supplementary Materials for details):

(1) All the soft-clipped reads satisfying the conditions in Supplementary Materials S.1 in the target tumor sample are remapped to the human reference genome (see Fig. 1b), and ITD-BPPs are identified (see also Supplementary Fig. S2).

(2) Support reads and their mate pairs for each ITD-BPP are assembled to generate a contig sequence. Then, the contig sequence is checked to determine whether it truly contains the sequence being inside the corresponding ITD-BPP [which we call presumed duplicated nucleotides (PDN)]. For this purpose, alignment of the contig sequence to the reference genome sequence around the corresponding ITD-BPP is performed to extract the excess unaligned part [which we call observed inserted nucleotides (OIN)], and the OIN is compared with the PDN of the corresponding ITD-BPP by pairwise alignment. Based on these results, we assign a degree of confidence to each ITD as follows (see Supplementary Fig. S3):

- ITD grade A: The contig sequence covers (aligns to) the nucleotides to the left and right of the ITD-BPP (i.e. the start and end alignment positions of the contig sequence are exactly the same as the left and right of the ITD-BPP, respectively). Furthermore,

*To whom correspondence should be addressed.
†The authors wish it be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

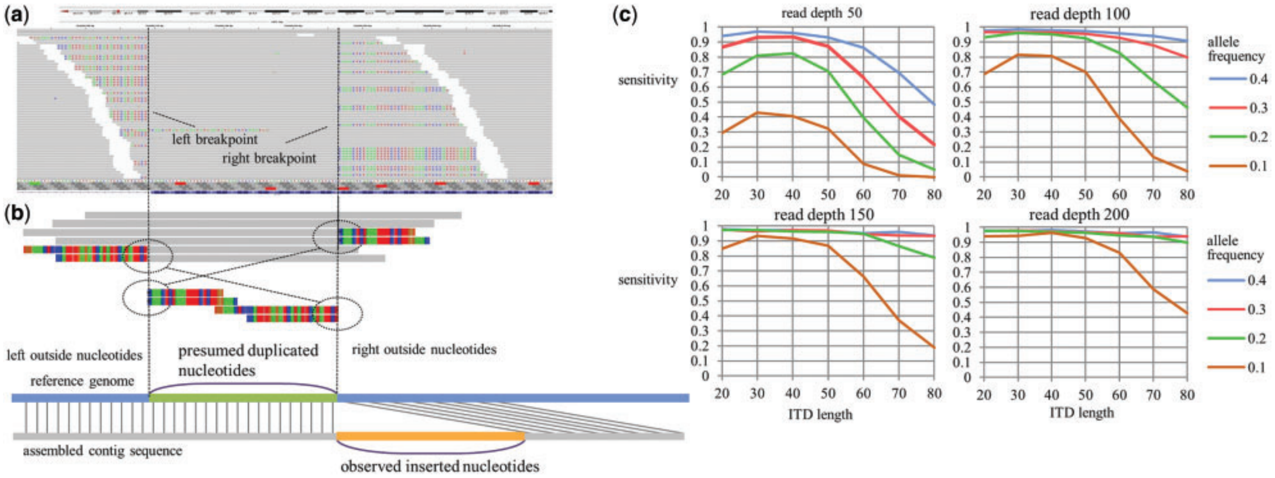

Fig. 1. (**a**) An illustrated view of the alignments around the *FLT3*-ITDs of the TCGA AML sequencing data, using the Integrative Genomics Viewer (IGV; Robinson *et al.*, 2011). The colored part in each read shows a soft-clipped fragment. (**b**) Relationships between soft-clipped sequences around ITDs and the corresponding ITD-BPPs. (**c**) The sensitivity of Genomon ITDetector using synthetic data with varying allele frequencies, read depths and ITD lengths

the number of matched bases between the OIN and the PDN is >90% of the length of the OIN and >80% of the length of the PDN.

- ITD grade B: The contig sequence covers (aligns to) either the nucleotides to the left or right of the ITD-BPP. Furthermore, the number of matched bases between the OIN and the PDN is >90% of the length of the OIN and >80% of the length of the PDN.
- ITD grade C: The contig sequence covers (aligns to) the nucleotides to either the left or right of the ITD-BPP. In addition, the number of matched bases between the OIN and the PDN is >90% of the length of the OIN and >30 bp.
- Not applicable: Candidates that do not fall into either of the above groups are filtered out.

(3) For filtering out polymorphisms and artifacts that commonly occur in multiple control samples, ITDs sharing the same ITD-BPPs as the set of control samples are removed.

3 RESULTS

3.1 Analysis of synthetic data

Using synthetic data, we evaluated the sensitivity and precision of Genomon ITDetector by altering three parameters: allele frequency, read depth and ITD length. See Supplementary Materials for detailed simulation settings. Here, we only counted ITDs identified as grade A.

As shown in Figure 1c, the sensitivity of the method was improved by increasing the read depth and allele frequency. However, even with sufficient read depths and allele frequencies, the sensitivity did not reach 100%, probably because of the difficulty of identifying ITDs located in repeat regions (Supplementary Fig. S4). Despite this minor limitation, Genomon ITDetector achieved sufficiently high sensitivity for practical use. For example, as described in Figure 1c, >90% sensitivity was obtained in the detection of ITDs spanning 20–60 bp at an allele frequency of 30% (which is the minimum condition for acceptable samples in the TCGA project, http://cancergenome.nih.gov/cancersselected/biospeccriteria) and read depths of 100 [close to the mean and median of the exome sequencing data used in the Cancer Genome Atlas (TCGA) AML project (Ley *et al.*, 2013)]. For ITDs spanning >70 bp, sensitivity was significantly increased by adding ITDs identified as grades B and C (Supplementary Fig. S5 and Table S1). Furthermore, the sensitivity decreased as the read length decreased (see Supplementary Materials).

3.2 Analysis of TCGA AML data

We used Genomon ITDetector to analyze exome sequencing data from 94 paired AML samples (Ley *et al.*, 2013, see Supplementary Materials for the detailed procedure). Please refer to Supplementary Figure S6 for each of the filtering steps used to reduce putative false positives. The computational resources required were 1 CPU core [Intel Xeon X5675 (3.06 GHz)], 5.0–7.7 GB, for each process. The average calculation time was 4.2 h, with a maximum of 22 h and a minimum of 1.5 h. In total, we identified 719 ITDs in coding regions, of which 45 and 20 were classified as grades A and B, respectively. ITDs involving *FLT3* were identified in 20 samples (grade A: 16, grade B: 1, grade C: 6; see Supplementary Fig. S8). Furthermore, ITDs affecting cancer genes (registered in the Cancer Gene Census, http://cancer.sanger.ac.uk/cancergenome/projects/census/), such as *CEBPA* (Lin *et al.*, 2005), *KIT* (Corbacioglu *et al.*, 2006) and *WT1*, were identified as grade A. These results showed that Genomon ITDetector is able to detect a number of ITDs in putative cancer driver genes with high sensitivity (Supplementary Table S2).

3.3 Comparison with Pindel

Most existing methods focus only on SVs of at least several hundred base pairs in size. Although Pindel (Ye *et al.*, 2009) can detect short (10–200 bp) tandem duplications, it is not specifically designed for detecting 'somatic' changes.

To evaluate the performance of Genomon ITDetector, we devised a simple framework for the detection of somatic ITDs using Pindel (s-Pindel) and compared it with Genomon ITDetector. These results suggested that although s-Pindel detected a similar number of *FLT3*-ITDs to Genomon ITDetector, it produced more putative false positives. Therefore, a simple extension of Pindel is not sufficient for accurate genome-wide screening of somatic ITDs, and Genomon ITDetector is a significant advancement. However, there is room for further improvement in Genomon ITDetector, as several probable ITDs were identified only by s-Pindel. The details of this comparison are shown in Supplementary Material S6. Integration of the merits of Pindel and Genomon ITDetector to produce a more sensitive and accurate software will be investigated in the future.

4 DISCUSSION

The effectiveness of Genomon ITDetector for detecting somatic ITDs was demonstrated using both synthetic and real data. Still, there are several remaining issues. First, the current approach cannot detect ITDs with more than two duplicates. One possible solution would be to check whether the OIN in the assembled contig sequence is double, triple or greater of the PDNs. Second, the proposed approach is for somatic ITD detection only, not germ line detection. For the detection of germ line ITDs, alignment artifacts and germ line ITDs must be effectively distinguished. For this filtering, use of multiple non-matched control samples may provide a solution; however, the data cannot be used directly in our current method, and this problem requires further investigation.

ACKNOWLEDGEMENTS

The supercomputing resources were provided by Human Genome Center, the Institute of Medical Science, The University of Tokyo. The authors would like to thank The Cancer Genome Atlas (TCGA) project.

Funding: Funding for open access charge: Integrative Systems Understanding of Cancer for Advanced Diagnosis, Therapy and Prevention (Grant-in-Aid for Scientific Research on Innovative Areas from the Ministry of Education, Culture, Sports, Science and Technology, Japan (Grant Number: 22134004)).

Conflict of interest: none declared.

REFERENCES

Corbacioglu,S. *et al.* (2006) Newly identified c-KIT receptor tyrosine kinase ITD in childhood AML induces ligand-independent growth and is responsive to a synergistic effect of imatinib and rapamycin. *Blood*, **108**, 3504–3513.

Kindler,T. *et al.* (2010) FLT3 as a therapeutic target in AML: still challenging after all these years. *Blood*, **116**, 5089–5102.

Ley,T.J. *et al.* (2013) Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.*, **368**, 2059–2074.

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

Lin,L.I. *et al.* (2005) Characterization of CEBPA mutations in acute myeloid leukemia: most patients with CEBPA mutations have biallelic mutations and show a distinct immunophenotype of the leukemic cells. *Clin. Cancer. Res.*, **11**, 1372–1379.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Shiraishi,Y. *et al.* (2013) An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.*, **41**, e89.

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.