Bioinformatics Advance Access published February 1, 2012

Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads

Ruping Sun [1]*, Michael I. Love [1], Tomasz Zemojtel [1], Anne-Katrin Emde [1], Ho-Ryun Chung [1], Martin Vingron [1] and Stefan A. Haas [1]

[1]Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, 14195 Berlin, Germany.

Associate Editor: Dr. Michael Brudno

ABSTRACT

Summary: We developed Breakpointer, a fast algorithm to locate breakpoints of structural variants (SVs) from single-end reads produced by next-generation sequencing (NGS). By taking advantage of local non-uniform read distribution and misalignments created by SVs, Breakpointer scans the alignment of single-end reads to identify regions containing potential breakpoints. The detection of such breakpoints can indicate insertions longer than the read length and SVs located in repetitve regions which might be missd by other methods. Thus, Breakpointer complements existing methods to locate SVs from single-end reads.

Availability: https://github.com/ruping/Breakpointer

Contact: ruping@molgen.mpg.de

1 INTRODUCTION

Identifing SVs from short sequencing reads remains challenging. Existing NGS-based methods for SV detection are primarily based on the analysis of paired-end reads (PE) assuming that deviations from the expected mapping distance are caused by SVs (Medvedev *et al.*, 2009; Alkan *et al.*, 2011). Alternatively, to characterize SVs from single-end reads (SE), split-read methods can be adopted to generate pseudo PE (Ye *et al.*, 2009; Smith, 2011). However, the short length of the artificial PE limits the mappability and the size of detectable insertions (up to medium size). Alternative SE-based methods are needed to facilitate the discovery of breakpoints of longer insertions and SVs located in repetitive regions.

SVs usually cause specific mapping artifacts in the vicinity of the SV boundaries. Reads slightly crossing the breakpoint of an SV can be mapped only if they contain a few bases of the variant (depending on the allowed edit distance). This case usually leads to consistent misalignments next to the breakpoint. Such misalignments have already been used to clean SNP calls (Li, 2011) or to determine the regions for local realignment (DePristo *et al.*, 2011). By contrast, reads spanning the breakpoint will be unmappable. Consequently, on the left or right side of the SV boundary, fewer read alignments will start or end, respectively (Fig. 1A and Fig. S1). Our tool Breakpointer locates SV breakpoints by analyzing both misaligmnent artifacts and local non-uniform read distribution created by SVs.

*to whom correspondence should be addressed

2 METHODS

Given a small genomic region R of size w, the depth of coverage D is defined as the number of reads (of length l) overlapping R. In addition, we introduce "end-depth" D_e as the number of only those reads starting/ending (summarized as "ends") within R. Assuming uniform coverage, under a given D, D_e will follow a binomial distribution as: $D_e \sim B(n = D, p = \frac{2w}{w+l})$. Regions containing the breakpoint of an SV will have higher D_e than expected, because a lack of mappable breakpoint-spanning reads leads to a depth skew toward ending reads. Some aligned reads slightly overlapping with the SV will generate consistent mismatches around the breakpoint (Fig. 1A). Such mismatches will only occur in the ends of the mappable reads but not in the reads spanning R, e.g. reads from a wildtype allele. We summarize these local mapping features around SV boundaries as "breakpoint signature". To capture this signature, Breakpointer proceeds in three stages (Fig. 1B):

First, using a sliding window of size w ($\ll l$), Breakpointer scans the read alignment along a reference genome, calculates the pileup-corrected sequencing depth D and end-depth D_e for each window. The skewness of depth in each window is represented by the score S_B, which is equal to the negative logarithm (of base 10) of the p value computed from a binomial test. Windows with $S_B > 1$ are selected and then merged into non-overlapping regions. For variable read length, Breakpointer groups reads based on their length and generates S_B based on weighted p values (Supplementary Note).

Second, mismatch screening is performed on each merged region, with the aim to enrich for those regions likely to encompass SV breakpoints. We only consider mismatches located less than 10bp from the read ends (referred as "ME"). A score S_M is assigned to a region with c positions showing MEs: $S_M = \sum_{t=1}^{c}(-\log_{10}(P_t))$ and $P_t = \binom{n}{k}\prod_{i=1}^{k}\epsilon_i \cdot \prod_{j=k+1}^{n}(1-\epsilon_j)$, where P_t is the probability of seeing k MEs out of n reads at a position t in a merged region, assuming that MEs are sequencing errors. ϵ_i is the maximum of the Phred base-error rate at position t in read i or a local error rate computed from the number of MEs in this region. We use local error rate because in real data there are regions containing many mismatches despite high quality scores. Small gaps are treated as mismatches (taking ϵ of surrounding bases for deletions). Regions with no MEs are removed since their depth skewness are likely caused by technological artifacts or mappability.

Given a true SV event, unmappable breakpoint-spanning reads will match the SV boundary including MEs. Thus, in the last step, each candidate region is validated by detecting breakpoint-supporting reads in the unmapped pool. The regions with no supporting unmappable reads are filtered out. Breakpointer sorts the selected regions according to S_B, S_M and generates two rank scores accordingly. A confidence score RC is assigned to each region by combining the two ranks. Breakpointer requires sorted BAM (Li *et al.*, 2009) files as an input and outputs validated regions in GFF format.

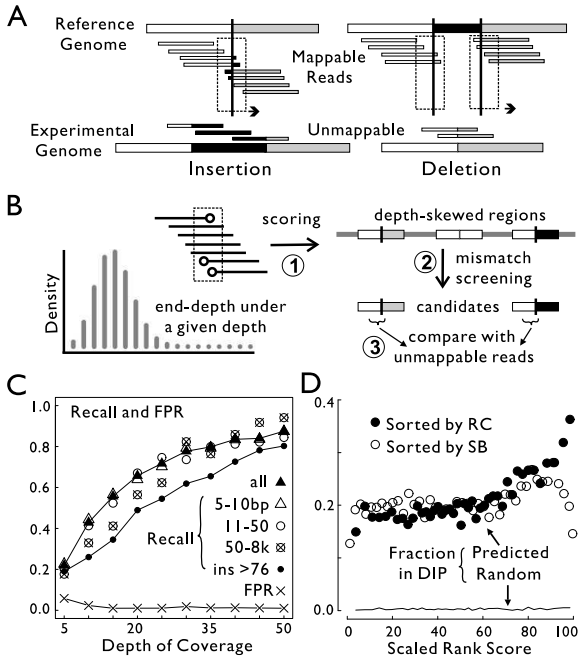

Fig. 1. (A) Depth skewness toward ending reads (small bar) and misalignments close to the breakpoints (black vertical line) of SVs (black bar). (B) Summary of Breakpointer algorithm. (C) Indel recall rate and false positive rate (FPR) at different coverage levels in the simulation. Recall: the breakpoint of an indel is encompassed by a prediction; FP: a prediction is not overlapping with any implanted indels; ins: insertions. (D) The fraction of predicted breakpoints by Breakpointerin the genome of NA18507 overlapping with DIP database. The predictions are grouped according to RC and S_B score, respectively. Also shown is the fraction of random regions overlapping with DIP.

3 RESULTS AND DISCUSSIONS

3.1 Simulation

We characterize the power of Breakpointer to locate the breakpoints of known human indels (Mills *et al.*, 2006) implanted into chromosome X (Supplementary Note). The results, summarized in Fig. 1C and Fig. S2, highlight the ability of Breakpointer to uncover the breakpoints of known indels with various sizes at high sequencing coverages (recall $>$0.8 and FPR: false positive rate $<$0.02 at $>$30x coverage). Breakpointer can discover the breakpoints of insertions longer than the read length, which is beyond the ability of split-read methods (Fig. S2).

3.2 Real data

Breakpointer is also tested on Illumina whole-genome sequencing data from an Yoruban genome (NA18507, Bentley *et al.*, 2008). Predictions by Breakpointer are intersected with external variant sets detected by alternative approaches from the same individual (Kidd *et al.*, 2008) and sets from population studies (Mills *et al.*, 2011a,b). The fraction of Breakpointer predictions overlapping with DIP (deletion/insertion polymorphisms detected by capillary sequencing a part of NA18507 genome) increases with combined rank score RC (Fig. 1D), suggesting that the breakpoint signature represents true SV breakpoints.

Predictions overlapping known SVs were comparable between Breakpointer and other methods using PE (Table S2), indicating that Breakpointer achieves equivalent accuracy to PE-based methods by just using SE. The comparison between the known SVs overlapped by Breakpointer and Pindel (Ye *et al.*, 2009, an anchored split read mapping method) reveals that Breakpointer locates the breakpoints of 1) insertions longer than the read length and 2) many indels in repetitive regions which are missed by Pindel (Table S3, Fig. S3). Breakpointer analyzes the initial small-gapped alignment of the entire read, showing complementarity to Pindel which splits the initial unmapped reads. Besides the breakpoints of indels, Breakpointer also uncovers the breakpoints of some large SVs, e.g. mobile insertions and non-homologous recombinations, by using 36bp SE (Supplementary Note), although the power to detect repeat-mediated SVs is limited due to mapping difficulties in highly repetitive regions and sequence homology around breakpoints.

4 CONCLUSIONS

By evaluating mapping features at the boundaries of SVs, our method locates the breakpoints of a wide range of SVs. Breakpointer does not investigate the SV content; it is designed as a supportive breakpoint discovery tool that ideally should be used in combination with other methods for genotyping SVs. Breakpointer requires a high coverage ($>$20x) to reach an optimal performance. The predictions can be used not only to provide additional support for alternative methods, but also to find breakpoints of SVs that otherwise might be missed by other tools. Thus, by requiring only single-end reads, it complements the current set of SV detection methods.

Funding: Max Planck Society.

REFERENCES

Alkan, C. *et al*. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Bentley, D. R. *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

DePristo, M. A. *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Kidd, J. M. *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

Li, H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, H. (2011) Improving SNP discovery by base alignment quality. *Bioinformatics*, **27**, 1157–1158.

Medvedev, P. *et al*. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.

Mills, R. E. *et al*. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, **16**, 1182–1190.

Mills, R. E. *et al*. (2011a) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Mills, R. E. *et al*. (2011b) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, **21**, 830–839.

Smith, H. E. (2011). Identifying insertion mutations by whole-genome sequencing. *BioTechniques*, **50**, 96–97.

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.