An approach for detecting structural variations from NGS paired end reads using Split Reads, Discordant Read Pairs and Local Alignment

M. Ta, P. Cotter, M. Moore: ResearchDx, 5 Mason, Suite 200, Irvine, CA 92618, USA

Introduction

A major challenge in Next Generation Sequencing is the development of efficient algorithms to detect structural variants present in the genome. Several different approaches for the detection of structural variants have been identified. Breakdancer searches for clusters of anomalous read pairs for sites to investigate. Similarly, another analysis tool, SoftSearch, uses the soft clipped read data from the aligner to determine sites of interest and heuristically report potential structural variants around them. Our algorithm, HardSearch, expands on the approach of SoftSearch to further identify the exact break points that support chromosomal structural variations. Paired end reads from DNA-seq with an unmapped mate are collected around each potential fusion site; the unmapped mates are realigned to the reference genome using a local aligner. The segment of each read that aligns with the highest alignment score without gaps is subtracted from the original and the remainder is realigned allowing for the identification of the breakpoint and breakpoint partners.

Methodology

Using the output from the aligner, HardSearch identifies and groups together softclip reads ocurring at the same location. Softclip reads are sequences where a portion of the read does not align properly. Regions containing multiple softclips reads suggests the presence of a structural variant. In each of these regions, we identify all reads that align properly but are paired with an unmapped mate.

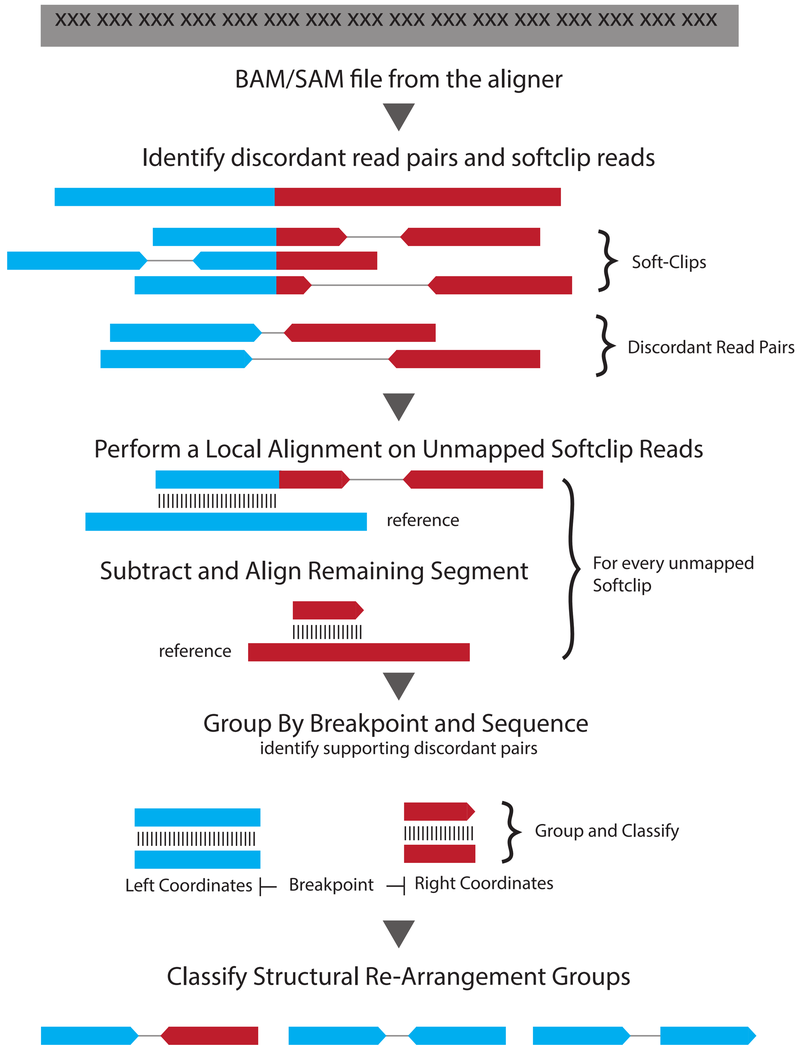
HardSearch realigns the unmapped sequence from the mate using a local alignment tool (BLAST). We run the complete unmapped sequence through the alignment tool and select the best alignment reported without any gaps If a portion of the read aligns, then it is subtracted from the original, unmapped sequence and the remainder is run through the local alignment tool once again.

The breakpoint is determined from the coordinates of the realigned segments separating them. Reads spanning the same breakpoint are grouped together. Chromosomal translocations are identified by realigned reads mapping to separate chromsomes and inversions are identified by realigned reads mapping in opposite orientations within the same chromosome.

Structural Variant Detection Tools

Tool	Method
BreakDancer	discordant read pairs
CNVnator	depth of coverage
Pindel	split read
SoftSearch	softclip reads

Table 1. Structural variation detection tools and their detection methods. HardSearch uses a combination of softclip and split read approach for the detection of translocation and identification of exact breakpoints.



Discussion

We developed HardSearch to look at oncology panels targeting specific genes present in translocations and inversions. By realigning initially unmapped reads, HardSearch can identify both ends of a breakpoint as long as one side is sequenced. Read pairs on either side containing a mapped and an unmapped sequence can be used to identify translocations or inversions present in the genome. In the case of an inversion, both sequences of a paired end aligment will map with the same orientation to the reference genome. Sequences spanning the breakpoint will contain a segment in each direction. Realigning the unmapped reads in the EML4-ALK sample shows that the segment aligning before the breakpoint is in the reverse orientation whereas the portion aligning after is in the forward orientation.

HardSearch offers a good compromise between performance and specificity. The program focuses on regions likely to span a breakpoint by filtering for softclip reads first; then realigning reads with unmapped mates situated near the softclips. Realignment of the unmapped mate allows HardSearch to identify additional reads that support a structural variant overlooked by other detection tools.

Results

EML4-ALK Inversion

Variant reads at each dilution for CRL5935

Sample	SoftSearch	HardSearch
Undiluted	5	132
50%	21	54
25%	1	3
15%	10	48

Table 2. The number of reads reported by each program supporting the EML4-ALK inversion. SoftSearch reported the variant as a novel insertion due to the high number of reads with an unmapped mate in the region. The diluation samples were created from a mixture of tumor cells containing the EML4-ALK inversion and from cells without the variant.

Conclusion

SoftSearch identifies and counts these reads to determine if an inversion ocurred. In the undiluted EML4-ALK sample, SoftSearch found 5 of these reads pairs but reported an additional 307 reads with unmapped mates in the region. Therefore, SoftSearch reported the variant in the region to be a novel insertion. By investigating and realigning the unmapped reads, HardSearch identified 132 supporting read pairs and reported the exact breakpoint as chr2: 42493956 - chr2: 29448092. Realigning the unmapped mate sequences allowed HardSearch to report more reads supporting the EML4-ALK inversion and show higher sensitivity in the diluted samples.