BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 21 2010, pages 2684–2688
doi:10.1093/bioinformatics/btq528

Sequence analysis

Advance Access publication September 27, 2010

SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data

Haley J. Abel[1,*], Eric J. Duncavage[2], Nils Becker[3], Jon R. Armstrong[4], Vincent J. Magrini[5] and John D. Pfeifer[3]

[1]Department of Internal Medicine, Division of Genetic Epidemiology, [2]Department of Pathology, University of Utah, Salt Lake City, UT, [3]Department of Anatomic and Molecular Pathology, Washington University, [4]Cofactor Genomics and [5]Department of Genetics, Genome Sequencing Center, Washington University, St. Louis, MO, USA

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Targeted 'deep' sequencing of specific genes or regions is of great interest in clinical cancer diagnostics where some sequence variants, particularly translocations and indels, have known prognostic or diagnostic significance. In this setting, it is unnecessary to sequence an entire genome, and target capture methods can be applied to limit sequencing to important regions, thereby reducing costs and the time required to complete testing. Existing 'next-gen' sequencing analysis packages are optimized for efficiency in whole-genome studies and are unable to benefit from the particular structure of targeted sequence data.

Results: We developed SLOPE to detect structural variants from targeted short-DNA reads. We use both real and simulated data to demonstrate SLOPE's ability to rapidly detect insertion/deletion events of various sizes as well as translocations and viral integration sites with high sensitivity and low false discovery rate.

Availability: Binary code available at http://www-genepi.med.utah.edu/suppl/SLOPE/index.html

Contact: haley@genepi.med.utah.edu

Received on June 11, 2010; revised on August 11, 2010; accepted on September 11, 2010

1 INTRODUCTION

Next-gen sequencing technologies provide an enormous amount of genome data at a cost many orders of magnitude lower than conventional capillary-based sequencing methods. While the use of next-generation sequencing has, to date, been largely confined to complex whole-genome analysis projects (Mardis *et al.*, 2009; Stephens *et al.*, 2009), the technology can similarly be applied to targeted 'deep' sequencing of particular genes or regions of interest. Using this hypothesis-driven approach, selected genes can be analyzed for rare sequence variants, indels and translocations easily and at low cost. This approach is particularly useful in the clinical laboratory where genes/regions of known prognostic significance can be analyzed simultaneously for structural variation. Furthermore, next-generation technologies have the capacity to sequence multiple samples at once through the use of 'barcodes' or indexed labels, further driving down the cost of individual experiments (Mardis, 2008).

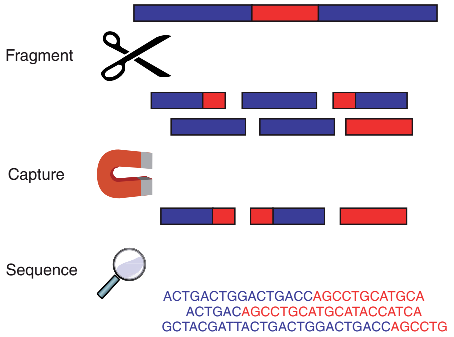
Fragment
Capture
Sequence
ACTGACTGGACTGACCAGCCTGCATGCA
ACTGACAGCCTGCATGCATACCATCA
GCTACGATTACTGACTGGACTGACCAGCCTG

Fig. 1. Hybrid Capture. Human genomic DNA *(blue)* containing integrated viral DNA *(red)* is first fragmented to an average size of 300 bp. Fragments may contain all human genomic DNA *(blue)*, all viral DNA *(red)* or chimeric viral/human DNA *(blue/red)*. The DNA fragments are processed into Illumina paired-end libraries and hybridized with capture probes specific for viral DNA sequences; fragments containing only non-complementary human genomic DNA are washed away. The captured DNA, including fragments with partial viral sequence matches, are then eluted, short cycle PCR-amplified using primers specific for the previously ligated adapters, and sequenced. Using SLOPE, chimeric reads representing the human/virus integration boundaries can be identified.

Next-generation sequence data can be generated from whole genomes, generally with coverage of 20- to 30-fold, or from 'targeted' regions of interest with much higher fold coverage. Target enrichment methods include standard PCR, ligation-mediated PCR or hybrid capture (Mamanova *et al.*, 2010). Unlike PCR-based approaches, hybrid capture enrichment is an amplification-free method that uses labeled, complementary DNA 'probes' to hybridize to regions of interest in genomic DNA. (Fig. 1). The hybridized, captured DNA sequences are then physically immobilized, while non-complementary sequences are washed away, and the captured DNA is eluted and sequenced. A major benefit of hybrid capture over PCR is that non-complementary sequences adjacent to regions of complementarity are also captured, permitting the identification of viral integration sites (as in this case) or translocation breakpoints.

Several programs, including BreakDancer, Pindel, VariationHunter, MoDIL and PEMer (Chen *et al.*, 2009; Hormozdiari *et al.*, 2010; Korbel *et al.*, 2009; Lee *et al.*, 2009; Ye *et al.*, 2009), to locate structural variants from sequenced whole genomes exist. However, none of these software packages is

[*]To whom correspondence should be addressed.

ideally suited for the analysis of targeted sequence data, especially in the clinical laboratory setting. Pindel, for example, detects certain indels with single-base resolution but is not designed to detect translocations. The remaining programs above rely on discordant mapping of paired-end reads to detect structural variants. Both members of each pair are aligned to the full reference genome; an unexpectedly large or small separation between the two members of a pair suggests an insertion or deletion, respectively, while a pair that gets mapped in the wrong orientation or onto two different chromosomes may indicate an inversion or translocation. These discordant pair methods cannot, in general, achieve single-base resolution, which may be clinically relevant, and may result in relatively high false positive discovery rates (Mardis *et al.*, 2009). Because they do not take advantage of the known target sequence enrichment, they also require a time-consuming preprocessing step, alignment of the sequence data to the entire human genome. We developed SLOPE to overcome many of these problems including speed and false positive discovery rate. By focusing on a small (a few kb to a few Mb) target reference sequence, SLOPE can perform fast and flexible split-read alignments and determine ‘chimeric’ sequences with single-base resolution.

SLOPE aims to detect sequence breakpoints from only one side of a split read, and therefore does not rely on the insert size for detection. In contrast with, for example, Pindel, which permits no mismatches with respect to the reference genome, SLOPE allows for multiple mismatches and gapped alignments between the query and reference genomes. Together, these account for the two major advantages of SLOPE: its speed and its tolerance to structural variation. Both of these are critical to its application in a clinical setting as clinical tests require timely results. Furthermore, DNA translocations, viral insertions, etc., may occur in regions of genomic instability, particularly in cancer genomes (Mitelman *et al.*, 2007). SLOPE attains much of its speed by aligning to only the short target sequence of interest, which requires many orders of magnitude less computation than aligning to the entire genome. While this approach limits SLOPE to the detection of sequence variation in the regions of interest, translocations or viral integration sites that involve these regions can be located in non-targeted areas by alignment of output sequences to the entire genome. SLOPE achieves its flexibility by placing little restriction on the allowable gaps and SNPs. The alignment of short, imperfect partial alignments to the reference genome will necessarily have a low specificity, however. SLOPE minimizes noise and detects the most likely chimeric sequences using a weighted least squares regression approach.

We demonstrate the utility of SLOPE by locating Merkel cell polyomavirus (MCPyV) integration sites in the human genome. Briefly, Merkel cell carcinoma (MCC) is a rare cutaneous tumor that has recently been noted to, in most cases, harbor a novel, clonally integrated polyomavirus, MCPyV (Duncavage *et al.*, 2009; Feng *et al.*, 2008). As MCC is a rare tumor, the characterization of integration sites has been difficult due the paucity of frozen tissue and high-quality DNA needed for traditional methods such as RACE, ligation-mediated PCR or inverse PCR. Using hybrid-selection and next-generation paired-end sequencing, we identify MCPyV viral integration sites from low DNA quality, archival, formalin-fixed, paraffin-embedded (FFPE) tissue specimens using SLOPE. We also test the SLOPE program on simulated data and compare its performance with that of the analysis packages BreakDancer and Pindel (Ye *et al.*, 2009).

2 METHODS

SLOPE is implemented in C^{++} and accepts as input either preprocessed paired-end data or unprocessed FASTQ single-end reads. In the case of paired-end data, improved computational efficiency is attained by preprocessing: the FASTQ sequence files should be first aligned to the short targeted reference genome using alignment software, such as MAQ (Li *et al.*, 2008), and converted to either the Maq mapview (Li *et al.*, 2008) or standard SAM format (Li *et al.*, 2009). This step reduces the computational burden, as SLOPE then processes only ‘orphaned sequences’, unmapped sequences whose mates did map to the target reference genome; however, SLOPE remains agnostic as to the orientation and position of the mapped partner in order to allow for unknown structural variation. In the case of single-end reads, no preprocessing is necessary, and SLOPE processes all sequences. SLOPE attempts partial alignments between either the 5′ or 3′ end of each unmapped read and the target reference genome: it determines the highest scoring quality-weighted ungapped alignments between either end of the sequence and the target genome, allowing up to five mismatches. Ties in high-scoring alignments are broken at random. In order to reduce noise due to spurious short partial alignments, alignments comprising less than 25–30%, depending on read length, of the total read length are disregarded. To improve computational efficiency, partial alignments representing more than 80–90% of the total read length, which are relatively uninformative about the putative variant, are also discarded. These cutoffs serve only to improve algorithm performance, and do not qualitatively affect the results. The remaining partial alignments are then clustered according to orientation and position on the reference genome to determine the set of potential breakpoints. For each unique read, the initial partial ungapped alignment is refined by the Smith–Waterman algorithm (Smith and Waterman, 1981) to allow for gaps. The Smith–Waterman refinements are restricted to a neighborhood (within 10 bp to either side of the initial, crude alignment) in order to reduce the computational burden. The unique reads in each cluster are then sorted in order of mapped position.

Within each sequence cluster, let x_i be the 5′ position of the i-th unique sequence. Let S_i be set of putative breakpoints, consisting of both indels and terminal ends of partial alignments, identified by the Smith–Waterman algorithm; let z_i be the median of S_i. Finally, let $y_i = z_i - x_i$, the ‘length’ of the i-th partial alignment. For a true chimeric junction, in the absence of noise, the 5′ positions and partial alignment lengths would correlate exactly. Furthermore, the line $y = \beta_1 x + \beta_0$ would have slope $\beta_1 = -1$ (Fig. 2A). However, non-specific alignments of short, imperfect partial reads, sequencing error, PCR replication error and variance in the best-fitting ‘breakpoints’ identified by the Smith–Waterman algorithm will affect the estimated slope. A weighted least-squares regression, with weights given by the Smith–Waterman alignment scores, is used to assign greater importance to longer, higher quality partial alignments. For each cluster representing a potential chimeric read, the slope $\hat{\beta}_1$ is determined and 99% bootstrapped confidence interval for β_1 estimated (Efron, 1979). This resampling approach is robust to error introduced in the PCR and sequencing step as well as to spurious short, partial alignments and provides that clusters representing indels or translocations should contain -1. Our results were also robust to the choice of weight function. For example, the weight function $w_i = 1/(1 + \max S_i - \min S_i)$, an estimate of the precision of the putative breakpoint, provided qualitatively similar results. SLOPE outputs all clusters whose confidence intervals contain -1, within a user-specified tolerance. Other clusters likely result from non-specific alignments of short sequence fragments.

3 RESULTS

3.1 MCPyV insertion site data

To determine the MCPyV integration sites from FFPE MCC specimens, 10μg of genomic DNA was first extracted from tissue blocks using standard methods (Duncavage *et al.*, 2009). To enrich

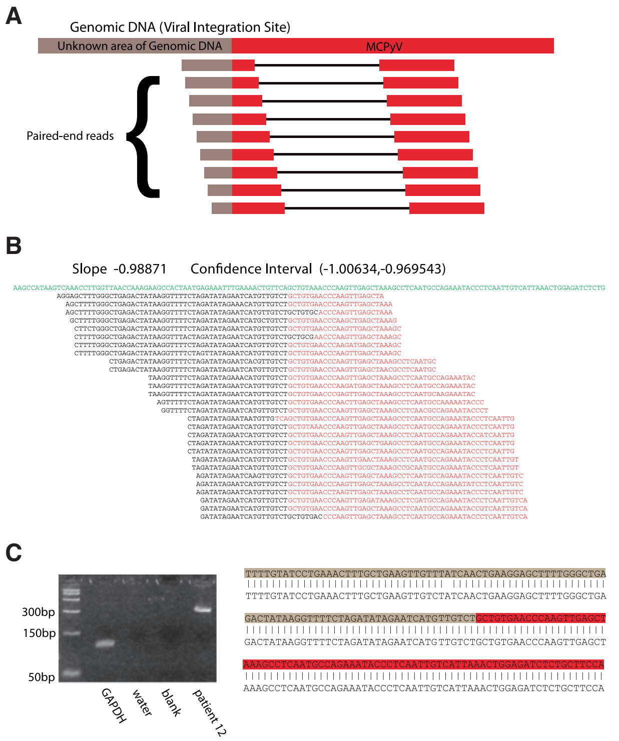
A
Genomic DNA (Viral Integration Site)
Unknown area of Genomic DNA
MCPyV
Paired-end reads
B
Slope -0.98871
Confidence Interval (-1.00634,-0.969543)
C
300bp
150bp
50bp
GAPDH
water
blank
patient 12
TTTTGTATCCTGAAACTTTGCTGAAGTTGTTTATCAACTGAAGGAGCTTTTGGGCTGA
TTTTGTATCCTGAAACTTTGCTGAAGTTGTCTATCAACTGAAGGAGCTTTTGGGCTGA
GACTATAAGGTTTTCTAGATATAGAATCATGTTGTCTGCTGTGAACCCAAGTTGAGCT
GACTATAAGGTTTTCTAGATATAGAATCATGTTGTCTGCTGTGAACCCAAGTTGAGCT
AAAGCCTCAATGCCAGAAATACCCTCAATTGTCATTAAACTGGAGATCTCTGCTTCCA
AAAGCCTCAATGCCAGAAATACCCTCAATTGTCATTAAACTGGAGATCTCTGCTTCCA

Fig. 2. Identification of viral integration sites by SLOPE. (**A**) Diagrammatic representation of sequence data. Seventy-five base pair paired end reads were used to sequence MCPyV-enriched human genomic DNA. Chimeric reads in which one end spans the viral integration site were identified by SLOPE. (**B**) Sequence data from single end reads around the viral integration site. Black sequences represent human genomic DNA, red MCPyV DNA and green MCPyV reference sequence (every 30th unique sequence shown). The slope represents a regression line drawn from through overlapping sequences. (**C**) Results were then verified by designing PCR primers to the proposed integration site, followed by human genomic DNA amplification, subcloning and sequencing by gel capillary electrophoresis. PCR (left) and subsequent sequencing results (right) are shown.

for MCPyV-containing sequences, the extracted genomic DNA was end repaired and ligated to standard Illumina sequencing primers. One microgram of material was denatured with 100 ng of biotin-labeled MCPyV capture probes spanning the 5.3 kb viral genome and subjected to hybrid capture. Post-hybridization, sequence-enriched DNA was eluted from the capture probes and PCR amplified before paired-end sequencing was performed on an Illumina GAII using default parameters. PCR products were analyzed by gel electrophoresis to ensure no distinct bands representing biased amplification were present (data not shown). Four samples, three with 75 bp read lengths and one with 50 bp reads, were enriched and sequenced in duplicate by the above methodology.

The paired-end reads were first aligned to the reference MCPyV genome using Maq (version 0.7.1) under the default parameters (Li *et al.*, 2008). Then the SLOPE program was used to identify chimeric reads. Finally, BLAST (Altschul *et al.*, 1990) was used to align the unmapped ends of the consensus chimeric reads to locate the insertion sites in the human genome.

For each of the three samples consisting of 75 bp paired-end reads, SLOPE successfully located both of the viral/human junctions breakpoints with single-base resolution. These breakpoints were also identified using the BreakDancer program (Chen *et al.*, 2009), and confirmed by PCR across the junction, followed by Sanger sequencing (Fig. 2). In the one sample of 50 bp paired-end reads, SLOPE found only one human/virus junction, which was then confirmed by PCR and Sanger sequencing. BreakDancer was unable to locate any chimeric reads from this sample. SLOPE located the chimeric reads from each sample in ∼10 min, as compared with BreakDancer which required several days.

3.2 Simulated data

Data were simulated by first linearizing the circular MCPyV genome (GI 165973999) and inserting it into a contiguous 1 Mb region of the human X chromosome (build 37 ref). Both the viral linearization breakpoint and the insertion site were varied across simulations. Insertions and deletions of lengths 1–1000 bp were randomly inserted. Maq (version 0.7.1) (Li *et al.*, 2008) was used to simulate paired-end 75 bp reads from this aggregate genome with 750-fold coverage (parameters: -N10 000 000 -r0.001 -R0). Quality scores were simulated according to the distribution of qualities from the real MCPyV data above, and a mutation rate of 0.001 was assumed.

The simulated viral insertion sites were located using SLOPE, just as for the real data. The sites were found using BreakDancer by first using Maq to align the short reads to a FASTA reference file consisting of the viral genome and the X chromosome. Then BreakDancerMax (version 0.0.1-r89, using parameter -e) was run to locate the ‘translocation’ between chromosome X and the MCPyV genome.

Both SLOPE and BreakDancer successfully located the viral insertion site in five of five cases. In the case of SLOPE, locating the viral insertion sites on a Macintosh Pro with dual 2.26 GHz Intel Xeon 5500 quad core processors and 16 GB of RAM required 760 s to align to the viral genome using Maq plus 45 s for SLOPE to determine the junctions. Location of the viral insertion sites on the same computer using Breakdancer required 7000 s for alignment to the viral and X chromosome reference sequences by Maq, followed by 3300 s for BreakDancerMax to determine the translocation. We note that in general, without the a priori knowledge that the translocation partner was located on the X chromosome, the computational burden (and runtime) for the BreakDancer method would increase by at least an order of magnitude.

Currently, SLOPE performs optimally given paired-end input. However, SLOPE is also capable of determining breakpoints given a single-end FASTQ file. We used Maq to simulate 75 bp reads as above, now with 50-fold coverage (parameters: -N350 000 -r0.001 -R0) and analyzed only the FASTQ file for a single end. Again, SLOPE successfully located the viral insertion sites in five of five cases, requiring ∼4100 s per case. The output was noisier, however, and the slopes showed larger deviations from −1, with wider confidence intervals.

Finally, since SLOPE (by detecting breakpoints with respect to a reference genome) locates not only translocations but also indels, we compared its ability to detect randomly placed insertions and deletions with that of BreakDancer (combined Max and Mini) and Pindel (Ye *et al.*, 2009). This time, we used Maq to align the reads to the truncated X chromosome and counted the number of insertions and deletions in the ranges small (1–5 bp), medium

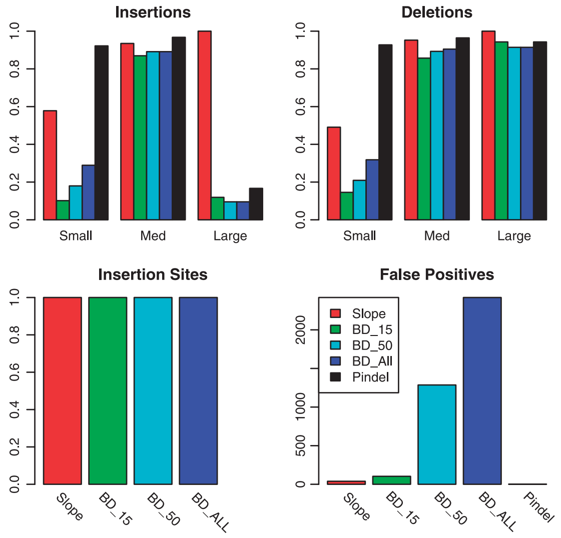

Fig. 3. Comparison with other methods. Slope *(red)* is compared with Pindel *(black)* and all BreakDancer calls *(blue)*, the top 50% of BreakDancer calls *(cyan)*, or the top 15% of BreakDancer calls *(green)*. Success rates for locating small, medium and large insertions *(upper left)* and deletions *(upper right)* from simulated data. *(lower left)* Success rates for detecting viral insertion sites. *(lower right)* Number of false positive discoveries for each program.

(10–50 bp) and large (100–1000 bp) detected by each program. Since SLOPE only detects breakpoints and prints alignments (but does not specify insertion versus deletion), BreakDancer often cannot discriminate between small insertions and small deletions, and all three methods sometimes group neighboring indels together, we defined 'successful location' of an indel as detecting any breakpoint, insertion or deletion covering or within a 10 bp radius of the true position.

Of the three methods, Pindel performed the fastest, requiring 60 s on a Intel Xeon 2.93 GHz powered server with 14 GB of RAM, and successfully identified 92–95% of indels in all categories tested except for the large insertions, of which it detected only 17%, and made only two false positive calls (Fig. 3). SLOPE was intermediate in speed (829 s) and provided highly accurate results (93–100% success rate) for medium and large insertions and deletions. However, SLOPE proved less accurate for small insertions and deletions, attaining success rates of 58 and 49%, respectively, and detected 39 false positive chimeric reads. BreakDancer was the slowest of the three methods, requiring 5100 s to run both BreakDancerMax and BreakDancerMini. It performed well for medium and large deletions and moderately well for medium-sized insertions, attaining accuracies of 72–95%. BreakDancer did not perform as well for small deletions, small insertions or large insertions, attaining accuracies of between 6 and 15%, and produced over 2200 false positives. In order to reduce the large number of false positives called by BreakDancer, we counted only the top 50% or the top 15% (by Breakdancer score) of hits. However, this increase in specificity resulted in a considerable loss of sensitivity (Fig. 3).

4 DISCUSSION

SLOPE is a fast and accurate tool for locating translocations and other structural variants from targeted, resequenced DNA. SLOPE was faster and more accurate than BreakDancer in detecting real and simulated insertion sites for MCPyV. SLOPE also required a much faster preprocessing step, alignment of the paired-end data to only the short viral genome, as compared with BreakDancer which required pre-alignment to a much larger genome (in general, the entire human genome). SLOPE also had higher sensitivity and specificity than BreakDancer for detecting indels from simulated data. SLOPE was not quite as fast as or as specific as Pindel for indel detection. However, except in the case of the smallest indels, SLOPE's sensitivity compared favorably with that of Pindel, which also does not detect translocations. In addition to the efficiency attained by aligning only to the targeted reference genome, SLOPE has the advantage of SNP and gap-tolerant alignments. This tolerance may be an important feature for a program used to interrogate genomes of cancerous cells. We do emphasize, however, that SLOPE is suitable only for targeted, resequenced DNA. For whole-genome searches, a program such as BreakDancer, which searches for discordant paired reads to detect structural variants in general, or Pindel, which uses a fast, but SNP-intolerant, pattern-matching algorithm to detect indels, would be preferable (Chen *et al.*, 2009; Ye *et al.*, 2009).

Another useful feature of SLOPE is its ability to provide single-base resolution of chimeric reads. Translocations are thought to underlie a wide variety of tumor types, and specific chromosomal rearrangements have prognostic value (Mitelman *et al.*, 2007). Current methods to detect translocations, including cytogenetics and fluorescence *in situ* hybridization (FISH), provide only a very low resolution indicator of the chromosomal aberration. Furthermore, cytogenetic studies require actively dividing cultured cells, which are not always available, especially in the case of solid tumors. The ability to accurately locate and characterize translocations from a variety of specimen substrates including the ubiquitous FFPE tissue may provide further diagnostic information, and may help to elucidate the mechanisms by which these translocations arise.

Currently, SLOPE performs optimally for paired-end reads, since this allows the sample to be narrowed to only those reads for which one end maps (in its entirety) to the reference genome. The algorithm works in just the same way for single-end reads, however. This is an important advantage in terms of analyzing data for clinical use, as paired-end sequencing takes up to 5 days as compared 3 days for single-end sequencing. Use of single-end reads results in a widened sample space and somewhat noisier output. It also results in a much slower search, although the algorithm is trivial to parallelize. Future work will speed this search using a more sophisticated string matching algorithm.

Given its ability to quickly and accurately define translocation breakpoints from target-enriched sequence data, SLOPE is well-suited to both research and clinical applications. For example, the *Myeloid Leukemia Lymphoma (MLL)* gene on 11*q*23 is commonly translocated in both acute myeloid leukemia and acute lymphoblastic lymphoma. However, as more than 80 MLL translocation partners are known, identification of partner genes by FISH is problematic. Use of a targeted sequence-enrichment approach (in this case for the *MLL* gene), followed by next-generation sequencing, and application of SLOPE could theoretically identify all such translocations regardless of breakpoint location or partner gene. Additionally, as next-generation sequencing methods are readily amenable to FFPE tissue blocks, identification of recurring translocations could easily be accomplished from difficult-to-culture solid tumors or archived specimens, thereby opening up additional diagnostic and research possibilities.

ACKNOWLEDGEMENTS

We thank Professor Elaine Mardis for her advice and generous sequencing support, Professor Alun Thomas for his comments on the manuscript and Xiaopei Zhu for her assistance with cloning and sequencing.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.*et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Duncavage,E.J. *et al.* (2009) Prevalence of Merkel cell polyomavirus in Merkel cell carcinoma. *Mod. Pathol.*, **22**, 516–521.

Efron,B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

Feng,H. *et al.* (2008) Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science*, **22**, 1096–1100.

Hormozdiari,F. *et al.* (2010) Next-generation VariationHunter: Combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350–i357.

Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing. *Genome Biol.*, **10**, R23.

Lee,S. *et al.* (2009) MoDIL: Detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, **6**, 473–474.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mappings quality scores. *Genome Res.*, **18**, 1851–1858.

Li,H. *et al.;* the 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Mamanova,L. *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, **7**, 111–118.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Mardis,E. *et al.* (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.*, **361**, 1058–1066.

Mitelman,F. *et al.* (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Stephens,P. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.

Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired short reads. *Bioinformatics*, **25**, 2865–2871.