Accurate prediction of breakpoints in sequences

Uma Devi Paila, Chun-Song Yang, Bryce Paschal, Aakrosh Ratan
University of Virginia, Charlottesville, VA

Abstract

Structural changes in chromosomes including deletions, insertions, inversions, translocations, copy-number abberations and other rearrangments represent a major source of variation that has been implicated both in phenotypic diversity as well as disease. These variants have to be resolved to the level of precise nucleotide junctions if we want to understand the underlying mutational mechanisms and biases that may play an important role in the etiology of these traits and conditions. Here, we present an inexact algorithm and an accompanying implementation to predict breakpoints in a sample using clipped and unmapped sequences in the dataset. We showcase its utility using simulated sequences replicating the most common use-case of human samples sequenced using short Illumina paired-end sequences. We also use the implementation to identify structural variants in the tumors sequenced as part of the Prostate Cancer Genome Sequencing Project, and report on the putative gene-fusions found in the primary prostate tumors. SRTk (Short-read toolkit) is under active development and the source-code can be downloaded freely from `https://github.com/aakrosh/SRTk`

Overview

	Query	Sequence			GGAGGCAGTAACCAAAGAAATAGCCAGGAAAAAAACAAATTATGAAAAAAAGAGAGAAAACCTATATGCATATTATTAAGAGAAAG
42	85	1	+	69608491	69608534	ATGAAAAAAAGAGAGAAAACCTATATGCATATTATTAAGAGAAAG
50	85	1	+	73063286	73063323	GAGAGAAAACCTATATGCATATTATTAAGAGAAAG
33	63	X	-	151036124	151036161	AAACAAATTATGAAAAAAGAGAGAAAACCTA
⋮						
1	41	1	+	69609965	69610005	GGAGGCAGTAACCAAAGAAATAGCCAGGAAAAAAACAAATT
29	77	19	-	22865979	22866010	AAAAAAACAAATTATGAAAAAAAGAGAGAAAACCTATATGCATATTATTAA

Figure 1: The alignments of the query to the reference genome are shown on the left. The alignments in red on the right show the optimal cover for the query.

We examine m alignments of the query Q of length n against a target genome T using an aligner e.g. LASTZ [RS07] that can find all alignments that score above a pre-determined threshold. These alignments are input to a modified Smith-Waterman algorithm, which then determines the optimal coverage set for the query from among the alignments. A variant with constant jump penalty requires $O(mn)$ time to calculate the optimal alignment score, whereas the variant with variable jump weights requires $O(m^2n)$ time. $O(mn)$ space is required to store the scoring matrix represented by V, as well as a trace-back matrix, which can then be used to recover the alignments for the best coverage set.

Simulations

We simulated 10,000 SVs (equal number of deletions, insertions, inversions and duplications, with size distributions inferred from the results of the 1000 genomes project) with breakpoints that were placed (a) uniformly across the human genome and (b) with a bias towards repeats in the human genome using RSVsim [BD13]. We ran SRTk and the output was supplied to LUMPY [Lay+14] to call the variants. We show the count of true-positive and the false-positive exact breakpoints that were identified as the simulations were repeated 5 times.

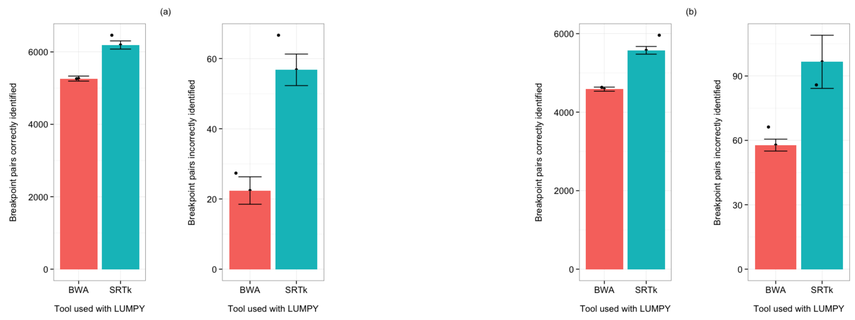

Figure 2: Performance of BWA+LUMPY compared to SRTk+LUMPY when (a) random breakpoints on hg19 are simulated, (b) breakpoints with bias towards repeats on hg19 are simulated.

Prostate Cancer Genome Sequencing Project

37 tumor-normal pairs sequenced as part of the Prostate Cancer Genome Sequencing Project `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000447.v1.p1` were analyzed for breakpoints using SRTk+LUMPY.

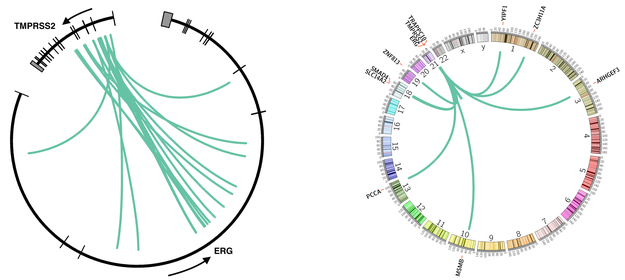

Figure 3: Location of breakpoints in TMPRSS2 and ERG found in 13 of the tumor samples. Also shown are the other genes that share breakpoints with TMPRSS2 in at least one of the samples.

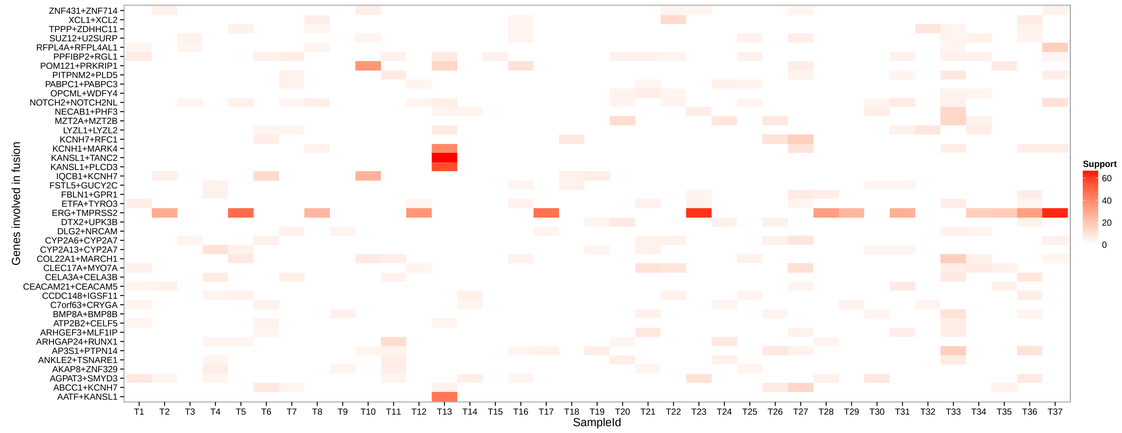

Figure 4: Putative gene fusions in the 37 tumor samples. Only the location of the breakpoint and direction of transcription were considered in making these predictions

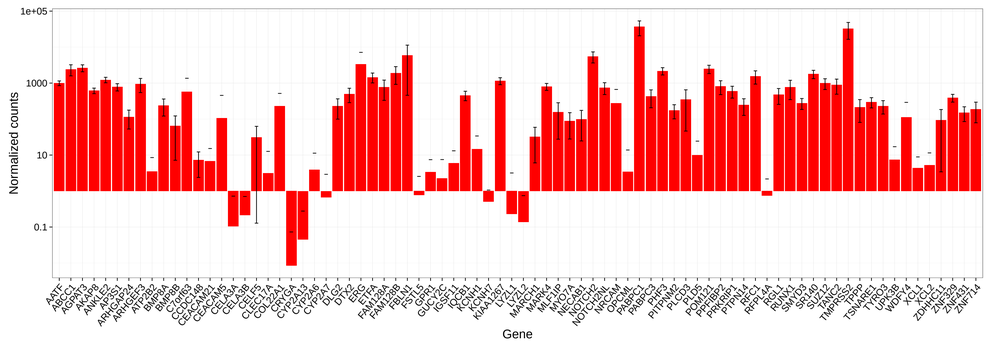

Figure 5: Normalized read counts for genes (80 out of the 81 predicted fusion genes) averaged over 178 tumor samples from a TCGA RNAseq dataset. A large fraction of the candidate genes are expressed in the tumor samples.

References

[RS07] Harris RS. "Improved pairwise alignment of genomic DNA". PhD thesis. Pennsylvania State University, 2007.

[BD13] Christoph Bartenhagen and Martin Dugas. "RSVSim: an R/Bioconductor package for the simulation of structural variations". In: *Bioinformatics* 29.13 (2013), pp. 1679–1681.

[Lay+14] Ryan M Layer et al. "LUMPY: a probabilistic framework for structural variant discovery". In: *Genome Biol* 15.6 (2014), R84. DOI: `10.1186/gb-2014-15-6-r84`. URL: `http://dx.doi.org/10.1186/gb-2014-15-6-r84`.