A Method to Detect Large Deletions at Base Pair Level using Next-Generation Sequencing Data

Matthew Hayes[1] and Jeremy S. Pearson[1]

Tennessee State University, Nashville, TN 37209, USA

Abstract. Genomic structural variants (SV) play a significant role in the onset and progression of cancer. Genomic deletions can create oncogenic fusion genes or cause the loss of tumor suppressing gene function which can lead to tumorigenesis by downregulating these genes. Detecting these variants has clinical importance in the treatment of diseases. Furthermore, it is also clinically important to detect their breakpoint boundaries at high resolution. To this end, we have generalized the framework of a previously-published algorithm that located translocations, and we have applied that framework to develop a method to locate deletions at base pair level using next-generation sequencing data. Our method uses abnormally mapped read pairs, and then subsequently maps split reads to identify precise breakpoints. On a primary prostate cancer dataset and a simulated dataset, our method predicted the number, type, and breakpoints of biologically validated SVs at high accuracy, demonstrating its efficacy in variant calling and accurate breakpoint prediction.

1 Introduction

Deletion structural variants (SV) play a role in the onset and progression of cancer. For example, a tumor suppressing gene may be at least partially deleted, or the intergenic region between genes could be deleted, leading to the formation of an oncogenic fusion gene. [1, 2]. The impact of SVs necessitates the development of efficient methods to locate and characterize them. We present a method, termed Pegasus, that finds groups of anomalously-mapped read pairs, and then subsequently aligns the *soft-clipped* portion of local reads to the reference, which could indicate a SV boundary. Pegasus had high sensitivity on a primary prostate cancer dataset and a simulated dataset. It also outperformed another method (Delly [3]) in breakpoint accuracy prediction. We previously presented an algorithm called Bellerophon that applied a similar approach to identify translocations [4].

2 Methods

To find likely variants, Pegasus first finds groups of discordant read pairs that could indicate a deletion. The method defines a discordant pair has having a

mapped distance between read pairs that is greater than $L = mean + k * stdev$, where *mean* is the mean mapped distance between mates, *stdev* is the standard deviation of mapped distance lengths, and k is a user-defined parameter, which for Pegasus is 4 by default.

The program takes read alignment results in SAM format and looks for clusters of overlapping discordant pairs. When a group of overlapping discordant pairs is found, the program then searches for soft-clipped reads that are presumably near the SV breakpoint of the reads on either side of the potential deletion. It then extracts the soft-clipped portion of at least one read and realigns it to the reference genome using BLAT [5][1]. Compared to the size of the reference genome, the size of the cluster region (a few hundred bases) is smaller by several orders of magnitude. Because of this, it is unlikely that even a single clipped subread will realign to the region by chance.

To be predicted as a structural variant, a cluster of overlapping discordant pairs must satisfy two criteria: 1) there must be at least *minD* discordant read pairs in the cluster, which for Pegasus is 3 (by default), and 2) there must be at least *minS* soft-clipped reads from either side of the event that remap within the cluster region, which is the region from the outermost read in the cluster towards the variant breakpoint. For Pegasus, this value is also 3 by default.

2.1 Discordant Read Pair Clustering and SV Prediction Algorithm

Definitions Let $R(p)$ denote the set of reads in a discordant read pair cluster c that are closest to the p-arm telomere. Let $R(q)$ denote the set of those reads in c that are closest to the q-arm telomere. Assume that the reads in $R(p)$ are the mates of the reads in $R(q)$. Thus, the set $S = \{R(p) \cup R(q)\}$ is a discordant read pair cluster that supports a putative deletion, and $|R(p)| = |R(q)|$. Let $S(R(p))$ be a function that returns any soft-clipped reads that map to a coordinate in the range $[min(R(p)), min(R(p)) + k * stdev]$ for $R(p)$, where *min* returns the mapping coordinates with the lowest value among all reads in $R(p)$. Let $S(R(q))$ be a function that returns any soft-clipped reads that map to a coordinate in the range $[max(R(q)) - k * stdev, max(R(q))]$ for $R(q)$, where *max* returns the mapping coordinates with the highest value among all reads in $R(q)$. For all $x \in S(R(p))$ and $y \in S(R(q))$, let $mapped(x)$ and $mapped(y)$ denote the mapping locations of the *aligned* portion of soft-clipped reads x and y. After realigning with BLAT, let $clip(x)$ and $clip(y)$ denote the aligned positions of the *clipped* portion of the soft-clipped reads x and y. The Pegasus algorithm is provided in the Algorithm 1 table.

The algorithm works by predicting the precise boundary of the SVs by observing the location of the reference where the clipped subread realigns. There may be several clipped sequences that realign to the region of a structural variant. Due to small sequence polymorphisms, it's possible that all of the sequences may not precisely align to the same location. Thus, the predicted breakpoint within the cluster region is the one to where most of the clipped subreads align.

[1] BLAT was more suitable for realigning the clipped portion of the sub-read due to its speed and ease of use.

Algorithm 1 Pegasus algorithm

procedure PEGASUS($BAMfile$) ▷ bam file containing alignments
  for all $DiscordantPairClusters\ c \in BAMfile$ **do**
    for all $R(p) \in c$ **do**
      Let $A = \{mapped(x) : \forall x \in S(R(p))\} \cup \{clip(x) : \forall x \in S(R(p))\}$
      Let $B = \{mapped(y) : \forall y \in S(R(q))\} \cup \{clip(y) : \forall y \in S(R(q))\}$
      if $S(p) \neq \emptyset$ **and** $S(q) \neq \emptyset$ **then** ▷ At least one side of the boundary contains soft-clipped reads
        if $|S(p)| \geq minD\ and\ clip(x), clip(y) \geq minS$ **then** ▷ minD = minS = 3
          Predict mode(A) *and mode*(B) *as the deletion coordinates*
        end if
      end if
    end for
  end for
end procedure

3 Experimental Design

We conducted two experiments in order to measure the ability of Pegasus to accurately predict deletions and their breakpoints. For both experiments we compared the results of Pegasus against those of the most recent version of Delly, which also detects structural variants through the use of paired reads and local split read alignments. For both experiments Delly's small indel detection was turned off since all deletion variants in both datasets were at least 1000 base pairs. Pegasus parameters were all set to default. BLAT was used for both experiments to realign the clipped portion of soft-clipped reads, for which all parameters were set to their default values.

3.1 Experiment 1: Simulated Data

For the simulated dataset, 2500 synthetic deletion variants were inserted into the human reference genome hg38 using SVSim [6]. The reads were created using Wgsim from the genome containing the synthetic deletions, and the size for the simulated events ranged from 1000 to 100000 base pairs. BWA was used to align the reads to the reference genome hg38. The subsequent SAM file was then analyzed by both Pegasus and Delly, and the results were compared by measuring the sensitivity (SE) and average breakpoint error (ABE) of each. This data had sequence read coverage of 40X and 100 base pair (bp) reads. The average insert size was 400 bp with a standard deviation of 50. The mutation rate was set to 0.001, and of those mutations, approximately 15% were indels. In order to compare the results of each SV method, we measured sensitivity (SE) of deletion predictions, and also average breakpoint error (ABE). For any structural variation (SV) prediction, the breakpoint error is defined as the difference in base pairs between the true variant boundary and the predicted variant boundary.

3.2 Experiment 2: Primary Prostate Cancer Data

The prostate cancer dataset utilized for the second experiment was from a patient (PR-0508) whose genome was analyzed in [7]. The Picard suite was used to deduplicate the alignments, after which it was aligned to the human reference genome hg18 by BWA. In order to preserve consistency hg18 was used, due to the original coordinates having been presented in this older version of the reference genome. Again, the SAM file was analyzed by the algorithms of both Pegasus and Delly, with comparison being made respective to the same categories as those of the first experiment.

3.3 Results on Simulated Data

The results on the simulated dataset are summarized by Figures 1 and 2 below. While it can be seen in Figure 1 that Delly's sensitivity (SE) was higher than that of Pegasus for all deletions, Pegasus also maintained a sensitivity >80% on all deletion sizes. Pegasus did outperform Delly, however, in demonstrating a lower average breakpoint error (ABE) for all deletion sizes, as shown in Figure 2.

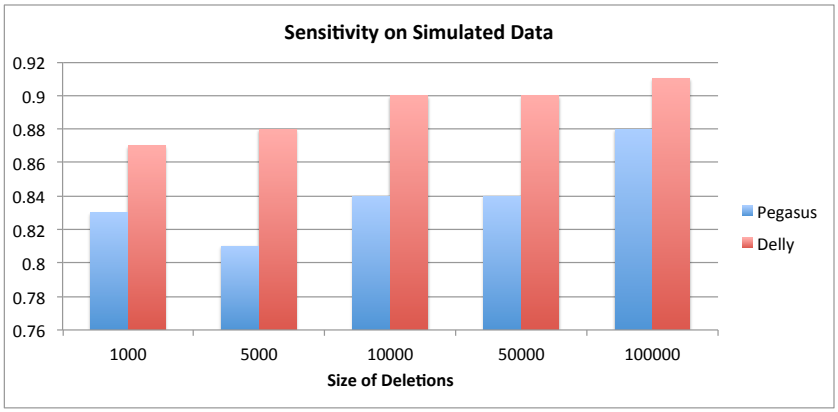

Fig. 1. Sensitivity of predictions on 2500 simulated deletions. The x-axis gives the size of the deletions in base pairs (bp). There were 500 deletions per size category.

3.4 Results on Prostate Cancer Data

There were 22 somatic deletions reported in this sample. Delly again had better sensitivity (SE=1) than Pegasus (SE=.95), though both had greater overall sensitivity for the real dataset when compared to the simulated data results. Pegasus once again displayed a lower average breakpoint error (ABE=.95) than that of Delly (ABE=.98). Though both had low ABE values, Pegasus demonstrated a

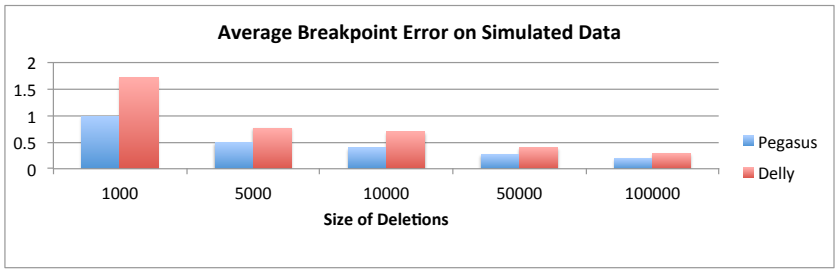

Fig. 2. Average breakpoint error on 2500 simulated deletions. The x-axis gives the size of the deletions in base pairs (bp). There were 500 deletions per size category.

superior ability to call precise breakpoints than Delly, which is significant in targeting specific genomic regions for cancer therapy.

4 Conclusion and Discussion

For both experiments, the sensitivity of predictions was higher for Delly, though Pegasus excelled at predicting precise boundaries. Regarding limitations, Pegasus is not suited for discovering small structural variants or indel polymorphisms, which can also be important markers for cancer diagnostics and therapy. Delly is superior in that regard. Also, Pegasus is currently only suited to find large deletions, while many methods for SV detection can identify several kinds of variants. Lastly, Pegasus is currently not capable of filtering germline variants from SV predictions. Future work will address all of the aforementioned issues.

References

[1] Tomlins S, Bjartell A, Chinnaiyan A, Jenster G, Nam R, Rubin M, Schalken J: ETS gene fusions in prostate cancer: from discovery to daily clinical practice. Eur Urol. 56: 275-286. 10.1016/j.eururo.2009.04.036 (2009).

[2] Mitelman F, Johansson B, Mertens F.: The impact of translocations and gene fusions in cancer causation. Nature Reviews Cancer, 7:233?245 (2007).

[3] Rausch T, Zichner T, Schalttl A, Stutz A, Benes Vladimir, Korbel J: DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28 (2012).

[4] Hayes M, Li, J: Bellerophon: a hybrid method for detecting interchromosomal rearrangements at base pair resolution using next-generation sequencing data. BMC Bioinformatics 14:Suppl 5 (2013).

[5] Kent WJ: BLAT - The BLAST-Like Alignment Tool. Genome Res. 12: 656-664 (2002).

[6] Faust G, SVsim: a tool that generates synthetic Structural Variant calls as benchmarks to test/evaluate SV calling pipelines. https://github.com/GregoryFaust/SVsim. Last accessed 4/25/16.

[7] Berger M M, Lawrence M S, Demichelis F, Drier Y, Cibulskis K, Sivachenko A Y, Sboner A, Esqueva R, Pflueger D, Sougnez C, Onofrio R, Carter S L, Park K, Habegger L A, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos A H, Pugh T J, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardie K, Baldwin J, Simons J W, Kitabayashi N, MacDonald T Y, Kantoff P W, Chin L, Gabriel S B, Gerstein M B, Golub T R, Meyerson M, Tewari A, Lander E S, Getz G, Rubin M A, Garraway L A : The genomic complexity of primary human prostate cancer. Nature 470:214-220 (2011).