Structural Variation Detection from Pacific Biosciences Data Using NextGENe®LR Software

November 2020

Megan McCluskey, Jacie Wu, Lidong Luo, Jonathan Liu

Introduction

Structural variants, variants greater than 50bp in size, contribute significantly to the understanding of a variety of disease types (1). Gene fusion events, for example, play a critical role in many cancers (2). Gene fusions can occur due to translocation events, as well as from deletion or inversion events within a chromosome. Detection of structural variants presents greater challenge than detection of SNVs and small indels. The Pacific Biosciences RS, RSII, Sequel, and Sequel II Systems utilize SMRT (single molecule, real-time) sequencing, producing reads of several kilobases in length (2). These long read lengths are well suited for the detection of structural variants (3).

NextGENeLR is for the analysis of long read sequencing data, such as from the Pacific Biosciences systems. NextGENeLR provides fast and accurate alignments of long sequencing reads and enables accurate detection of structural variants, including gene fusions.

Methods

NextGENeLR uses a specialized long read alignment algorithm. This alignment algorithm works by first indexing the reference sequence, compressing homopolymer regions, which allows homopolymer length differences to be tolerated during alignment. A separate index table is saved for each chromosome. Reads are also similarly indexed. Each chromosome table is then searched for matches. The region with the most and longest matches is selected and a local alignment is performed. The long reads can be split to properly align reads containing structural variants. For each read the alignment positions for each section are compared to the reference. Any differences between the reads and reference are evaluated to determine the structural variant type, as shown in Figure 1 below. If sections of a read align to different chromosomes than a translocation is reported. If sections of a read align in different directions, then an inversion is reported. Similar breakpoints, within 5bp, are merged.

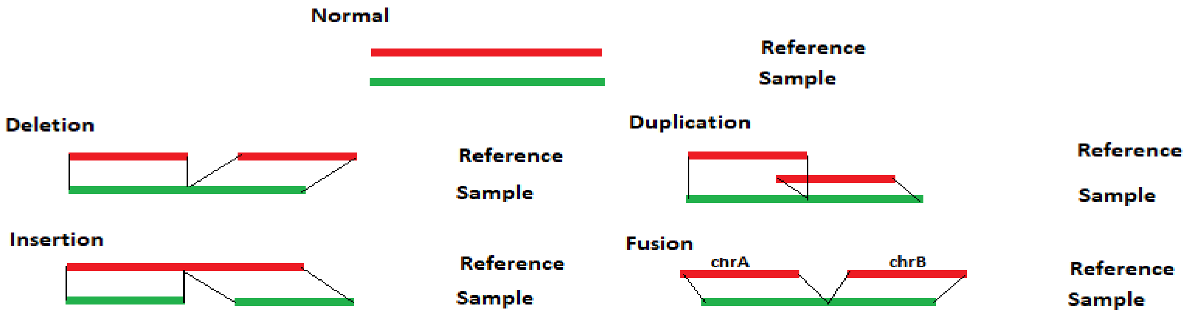
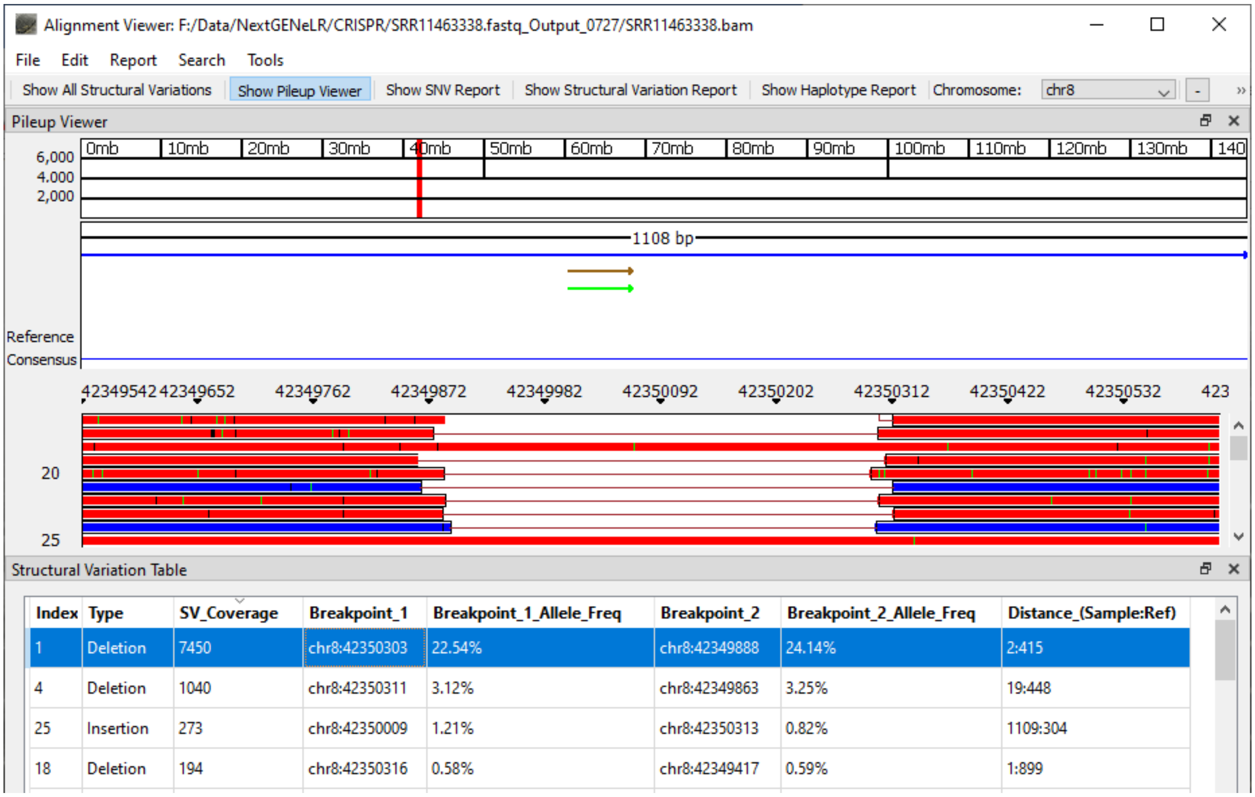


Figure 1: NextGENeLR makes structural variant calls by comparing the alignment positions for the sample and the reference. The patterns above illustrate how the structural variant type is reported.

Results

Once analysis is completed the generated BAM file can be viewed using NextGENeLR's Alignment Viewer. The Structural Variation Table is shown by default, with the Pileup Viewer above the table. One chromosome can be viewed at a time with the Pileup Viewer displayed. Near the top right is a dropdown menu to select the chromosome. You can also click "Show All Structural Variations" to view a table with structural variants for all chromosomes.

The Structural Variation Table lists the detected structural variants. For each variant, a type is provided as well as the number of reads supporting the structural variant, the breakpoint positions, the frequency of the variant out of the total reads aligned at each breakpoint position, and the distance in bp for the sample and the reference.



Alignment Viewer: F:/Data/NextGENeLR/CRISPR/SRR11463338.fastq_Output_0727/SRR11463338.bam

Index	Type	SV_Coverage	Breakpoint_1	Breakpoint_1_Allele_Freq	Breakpoint_2	Breakpoint_2_Allele_Freq	Distance_(Sample:Ref)
1	Deletion	7450	chr8:42350303	22.54%	chr8:42349888	24.14%	2:415
4	Deletion	1040	chr8:42350311	3.12%	chr8:42349863	3.25%	19:448
25	Insertion	273	chr8:42350009	1.21%	chr8:42350313	0.82%	1109:304
18	Deletion	194	chr8:42350316	0.58%	chr8:42349417	0.59%	1:899

Figure 2: NextGENeLR Alignment Viewer showing the Structural Variation Table. Highlighted is a 413bp deletion.

Discussion

Alignment of long reads from Pacific Biosciences systems with NextGENeLR software provides a powerful tool for structural variant detection. This can be used as a valuable resource for disease study.

NextGENeLR can also be used for STR expansion analyses and whole genome mitochondrial sequencing analysis, including haplotyping and SNV and indel detection using long read data. Alignments can be set up using an intuitive interface, with batch processing available to quickly set up analyses for multiple samples. The built-in Alignment Viewer provides an interactive tool for visualizing results and viewing reports specialized for each application type.

References

1. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14:125–38.

2. Davies KD, and Doebele RC. Molecular Pathways: ROS1 Fusion Proteins in Cancer.
Clin Cancer Res. 2013; (19) (15) 4040-4045.

3. Eid, J., et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; 323(5910), 133–138.

4. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics. 2015; 13(5): 278-289.