Detecting Pathogenic Structural Variants with Long-Read PacBio SMRT Sequencing

Aaron Wenger, Luke Hickey, Yuan Li, Ben Lerch, Paul Peluso, Jonas Korlach
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

Structural variants (SVs) – genomic differences ≥50 base pairs – contribute to human disease, traits, and evolution.

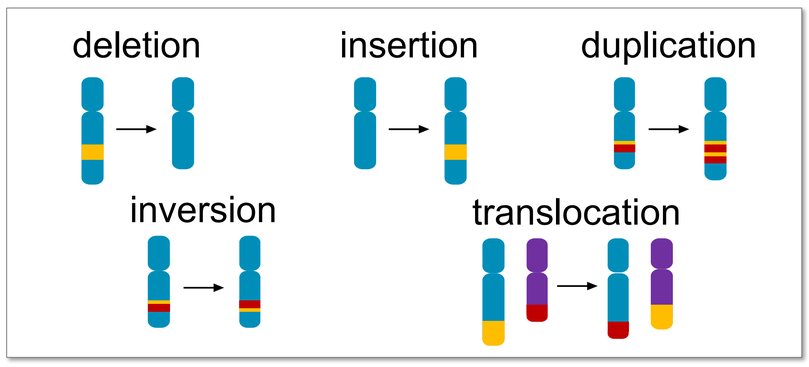

Figure 1. Common types of structural variant.

Compared to single nucleotide variants (SNVs) and indels, structural variants are few by count yet include most of the base pairs that differ between two humans.

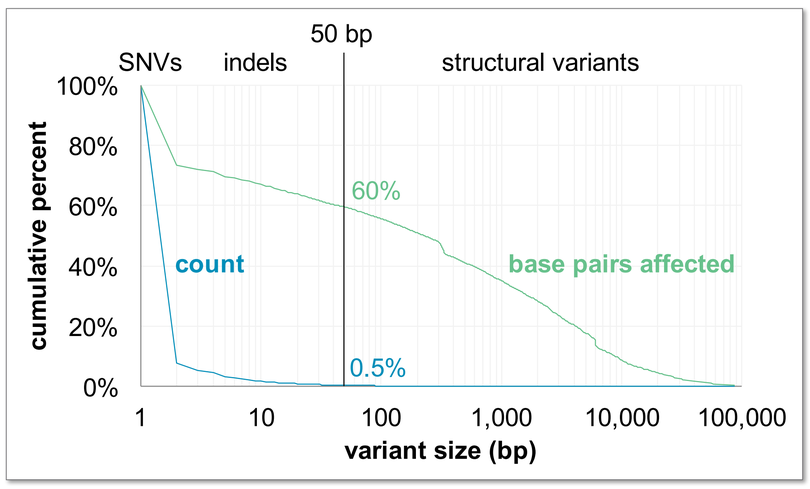

Figure 2. Count of and base pairs affected by variants in a human genome from structural variant, indel, and single nucleotide variant calls in HG00733 against GRCh38 from multiple sequencing technologies.[1] Only 0.5% of variants but 60% of variant base pairs are in structural variants ≥50 bp.

Most human structural variants were detected only by PacBio long reads.

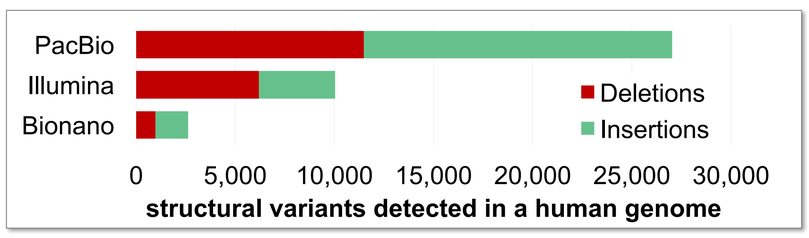

Figure 3. Sensitivity for structural variants in a human genome by technology. Most of the structural variants in HG00733 were detected only by PacBio long reads due to the propensity of structural variants to involve repeats or large insertions that are not spanned by short reads.[1]

The inability to detect structural variants limits the solve rate of exome and short-read whole genome sequencing.[2]

Rationale

We develop and apply a workflow for detecting structural variants in PacBio long reads to improve the solve rate for Mendelian disease cases.

Methods

To detect structural variants, we apply whole genome sequencing on the PacBio Sequel System, align reads with NGMLR[3], and call variants with pbsv.

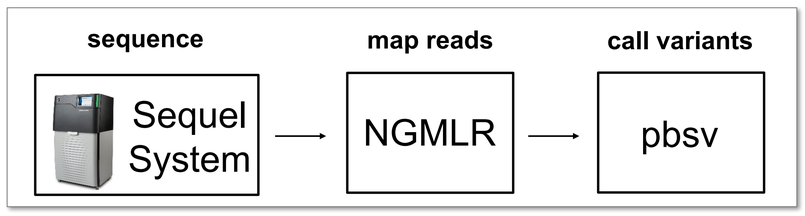

Figure 4. Overall workflow to detect structural variants from PacBio long reads.

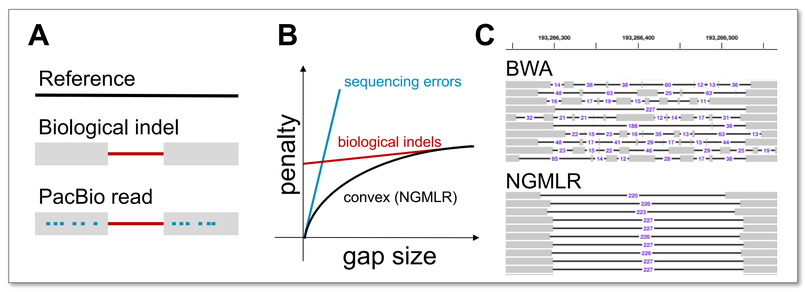

Figure 5. NGMLR correctly aligns PacBio reads around structural variants.[3] (A) PacBio reads have indels both from biological variation and sequencing errors. (B) NGMLR uses a convex gap penalty to effectively model the statistics of both types. (C) The same reads aligned with BWA and NGMLR illustrate how NGMLR produces sharp alignment gaps.

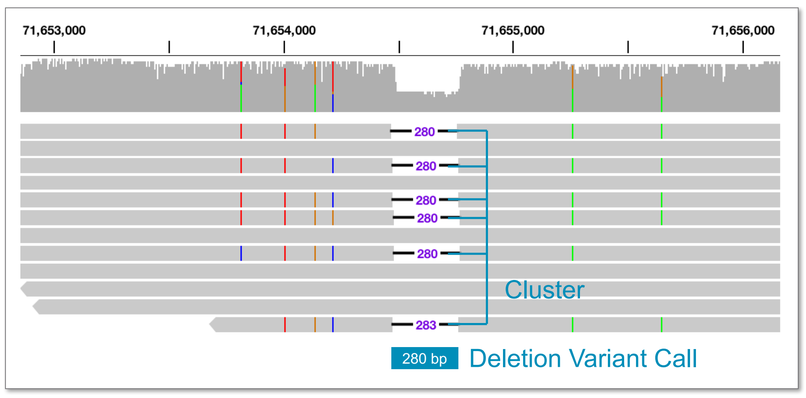

Figure 6. pbsv calls structural variants directly from read alignments. To call structural variants, pbsv identifies large deletion or insertion events in alignments, clusters nearby events that have similar length and sequence, and summarizes into a call.

Sensitivity vs. Coverage Titration

To evaluate the sensitivity of PacBio reads for structural variants vs. coverage, we generated 211 Gb (70-fold) for HG00733, subsampled coverage, and compared the variant calls to the full call set. Sensitivity is high even at modest 5- to 10-fold coverage.

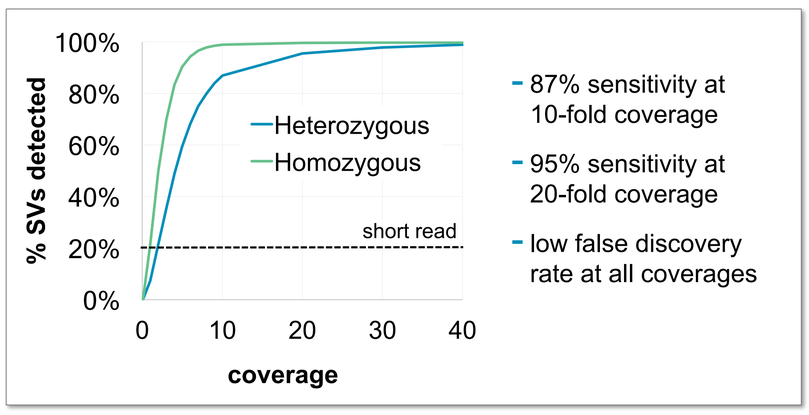

Figure 7. Sensitivity for structural variants at different PacBio coverage levels in HG00733.

Mendelian Disease Case Study[4]

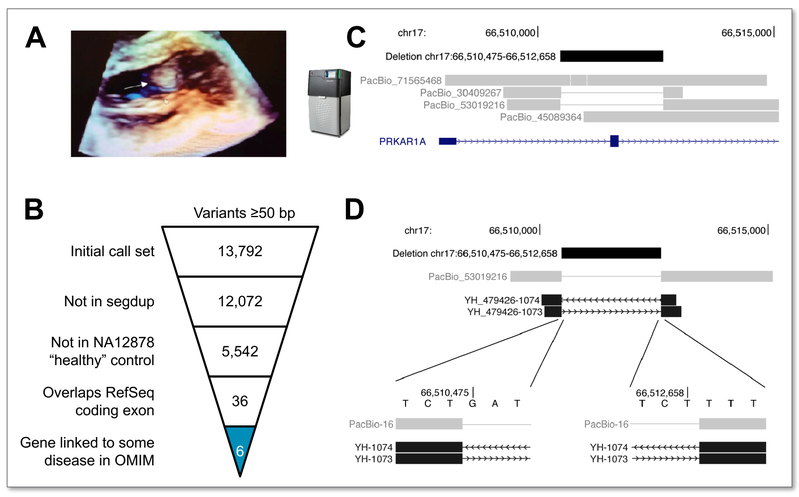

Figure 8. Low-coverage sequencing on the PacBio Sequel System identifies a pathogenic structural variant in a Mendelian disease. Short-read whole genome sequencing failed to provide a diagnosis for an individual with (A) cardiac myxomata. (B) Low-coverage PacBio sequencing identified thousands of structural variants, which were filtered to six variants of interest. (C) One is a heterozygous deletion of the first coding exon of *PRKAR1A*, null mutations in which cause autosomal dominant Carney complex. (D) The deletion breakpoints were confirmed by Sanger sequencing.

PacBio Reads in IGV – http://igv.org

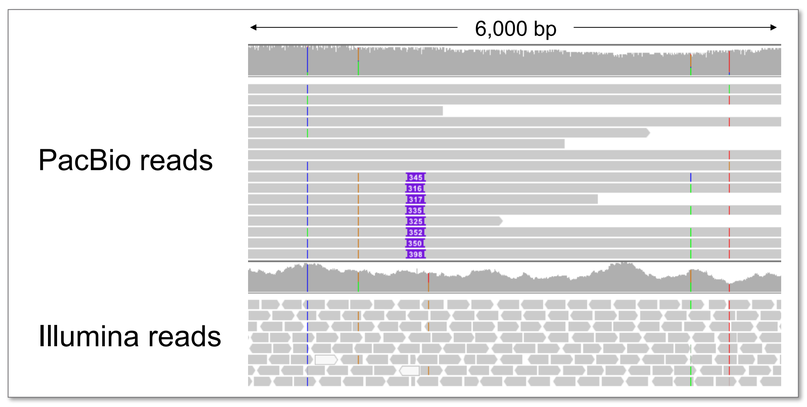

Figure 9. Structural variants in IGV. Improved support for PacBio long reads in IGV 2.4 makes it easy to see structural variants in haplotypes. PacBio reads agree with Illumina at single nucleotides but also show structural variation, such an insertion at GRCh37 chr13:78,585,000.

Conclusion

- Over 60% of human genetic variation is in structural variants ≥50 bp, most of which were detected only by PacBio sequencing.
- The sensitivity for structural variants is high even at low-coverage (5- to 10-fold) PacBio sequencing.
- Structural variant discovery with PacBio sequencing holds promise to increase the solve rate for Mendelian disease cases.

References

1. Chaisson MJ, et al. (2017). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*. doi:10.1101/193144.
2. Biesecker LG, et al. (2011). Exome sequencing: the expert view. *Genome Biology*. 12(9),128.
3. Sedlazeck FJ, et al. (2017). Accurate detection of complex structural variations using single molecule sequencing. *bioRxiv*. doi:10.1101/169557.
4. Merker JD, et al. (2017). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med.* doi:10.1038/gim.2017.86.

Thank you to David Scherer, Jenny Ekholm, Wendy Weise, Kathryn Keho, and Kristin Robertshaw for poster production support.