Structural Variant Detection with Low-Coverage PacBio Sequencing

Aaron Wenger, Luke Hickey, Jason Chin, Jonas Korlach
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

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

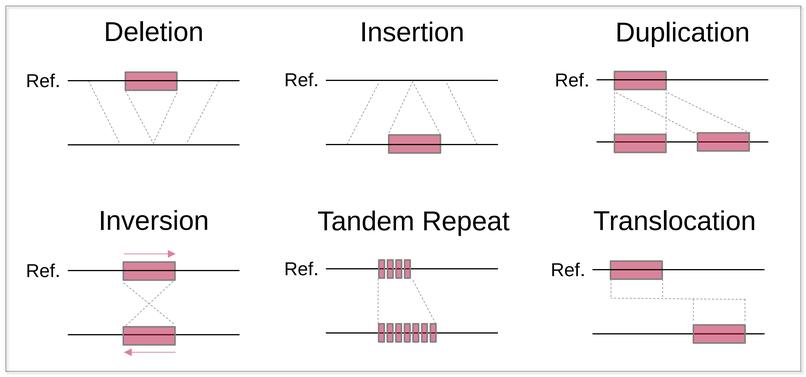
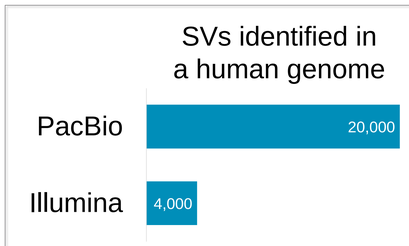


Figure 1. Common types of structural variation.

Most structural variants are too small to detect with array comparative genomic hybridization but too large to reliably discover with short-read DNA sequencing. Recent *de novo* assemblies of human genomes show that PacBio SMRT Sequencing sensitively detects structural variants.

Personal Genome	PacBio Coverage	Deletions ≥50 bp	Insertions ≥50 bp
CHM1[1]	41-fold	6,111	9,638
HX1[2]	103-fold	9,891	10,284
AK1[3]	101-fold	7,358	10,077

Table 1. Structural variants in PacBio *de novo* human genome assemblies.



"Fivefold increase in sensitivity [when compared to short-read sequence data]... from the improved mappability of long-read sequence data to repeat-rich regions (especially STRs and variable number tandem repeats), GC-rich DNA, and low-complexity DNA."[4]

Figure 2. PacBio long reads have 5-fold increased sensitivity for structural variants compared to Illumina short reads.

Rationale

While *de novo* assembly is the ideal method to identify variants in a genome, it requires high depth of coverage. A structural variant discovery approach that utilizes lower coverage would facilitate evaluation of large patient and population cohorts. Here we introduce such an approach and apply it to 10-fold coverage of several human genomes generated on the PacBio Sequel System.

With the Sequel System and the low-coverage analysis workflow, structural variant detection with PacBio reads is now at a price point that supports widespread adoption.

Methods: Map Reads, Chain Alignments, Call Variants

The workflow to identify structural variants from low-coverage PacBio sequencing is: 1) map reads to the reference, 2) chain alignments, and 3) cluster indels to call variants.

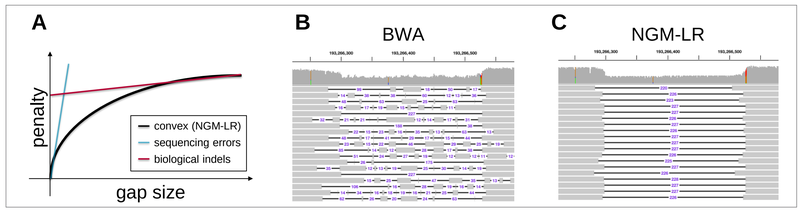


Figure 3. NGM-LR is a read mapper designed for PacBio reads[5]. (A) NGM-LR uses a convex gap penalty to model two sources of alignment gaps: biological indels and sequencing errors. (B) BWA, which uses a standard affine gap penalty, produces fragmented alignments at a deletion variant. (C) NGM-LR aligns the same PacBio reads with sharp boundaries at the deletion.

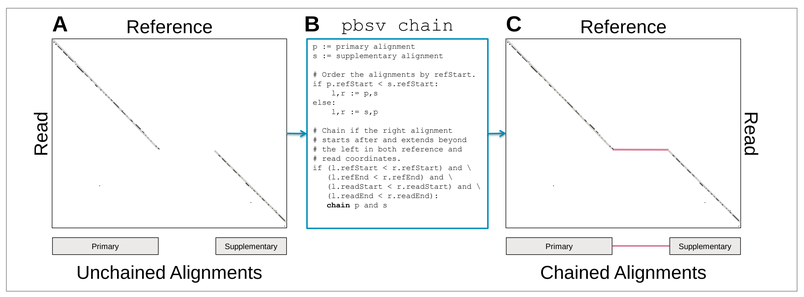


Figure 4. Chaining split alignments. Large gaps split NGM-LR alignments into primary and supplementary segments. Chaining connects colinear segments across large gaps. (A) A large deletion splits alignments of a read into two disjoint segments. (B) Criteria to chain split alignments. (C) Chained alignments directly include a biological deletion, which simplifies visualization and variant calling.

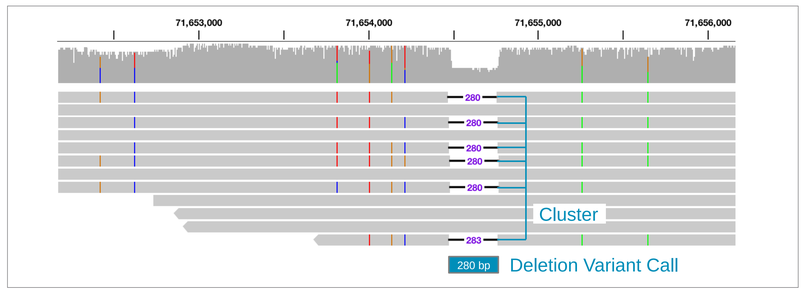


Figure 5. Variant calling. To call structural variants from low-coverage sequencing, identify large deletion or insertion events in chained alignments, cluster nearby events that have similar length and sequence, and summarize into a call.

Benchmarking with NA12878

Set	Platform	Deletions	Insertions
1000 Genomes[6]	Illumina	1,910	1,090
Genome in a Bottle[7]	Multiple	2,668	n/a
10-fold PacBio	**Sequel System**	**8,209**	**11,350**

Table 2. NA12878 structural variant sets.

Figure 6. Overlap with truth sets. The 10-fold PacBio call set recovers (A) 88% of true deletions, and (B) 81% of true insertions. The 10-fold PacBio set also includes thousands of novel variants, most of which are directly confirmed by a FALCON-Unzip *de novo* assembly from 60-fold PacBio RS II coverage.

Mendelian Disease Case Study (J Merker, EA Ashley)[8]

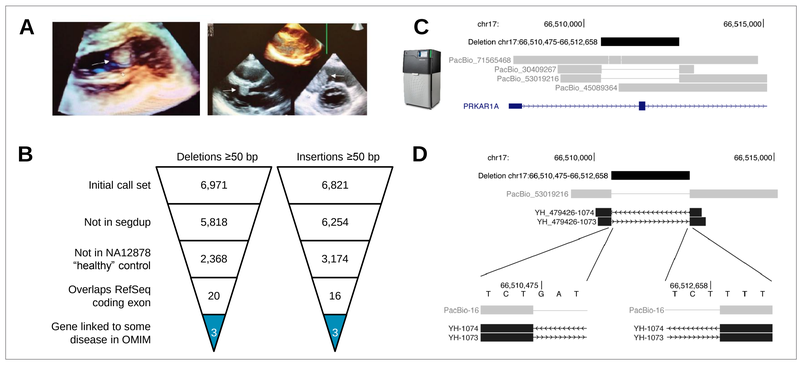


Figure 7. Low-coverage sequencing on the PacBio Sequel System identifies a pathogenic structural variant in a Mendelian disease. Targeted gene testing and short-read whole genome sequencing failed to provide a diagnosis for an individual with (A) cardiac myxomata. (B) Low-coverage PacBio sequencing identified thousands of structural variants in the individual, which were filtered to six variants of interest. (C) One of the six 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.

Visualizing Long Reads in IGV – http://igv.org

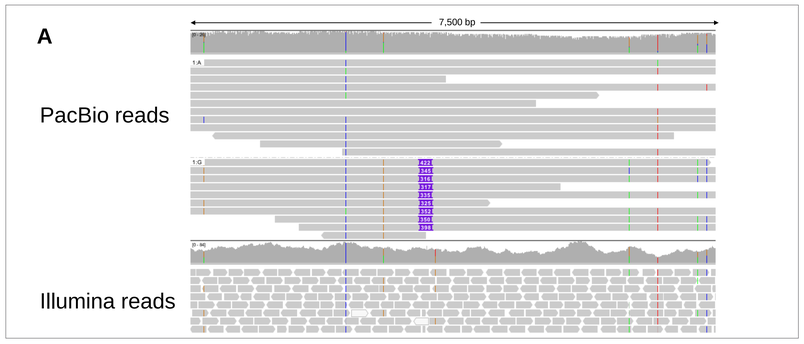


Figure 8. Structural variants in IGV. Improved support for PacBio long reads in IGV v3 makes it easy to see structural variants in phase with single nucleotide variants. PacBio reads agree with Illumina at single nucleotides but also show structural variation. (A) insertion at GRCh37 chr13:78,585,000.

Conclusion

- PacBio SMRT Sequencing has 5-fold increased sensitivity for structural variants compared to short reads.
- Low-coverage (10-fold) PacBio sequencing of NA12878 recalls 86% of known structural variants and identifies thousands more not previously seen in short reads.
- Low-coverage PacBio sequencing discovers a pathogenic variant missed by short-read whole genome sequencing.
- Software tools support read mapping, structural variant calling, and visualization for PacBio long reads.

References

1. Chaisson MJ, et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*. 517(7536), 608-611.
2. Shi L, et al. (2016). Long-read sequencing and *de novo* assembly of a Chinese genome. *Nature Communications.* 7,12065.
3. Seo JS, et al. (2016). *De novo* assembly and phasing of a Korean human genome. *Nature.* 538(7624), 243-247.
4. Huddleston J, et al. (2016). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research.* doi:10.1101/gr.214007.116.
5. Rescheneder P, et al. (2017). https://github.com/philres/ngmlr
6. Sudmant PH, et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature.* 526(7571), 75-81.
7. Parikh H, et al. (2016). svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*, 17, 64.
8. Merker J, et al. (2016). Long-read whole genome sequencing identifies causal structural variation in a Mendelian disease. *bioRxiv*. doi:10.1101/090985.

Thank you to Kevin Eng, Christine Lambert, Matthew Boitano, and Primo Baybayan for data generation; and to David Scherer, Wendy Weise, Kathryn Keho, and Kristin Robertshaw for poster production support.