

Introduction

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

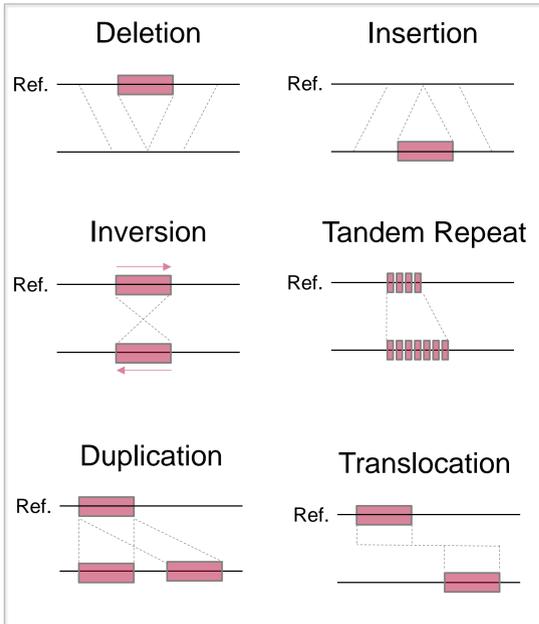


Figure 1. Common types of structural variation.

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	Deletions ≥ 50 bp	Insertions ≥ 50 bp
CHM1 ¹	6,111	9,638
HX1 ²	9,891	10,284
AK1 ³	7,358	10,077

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

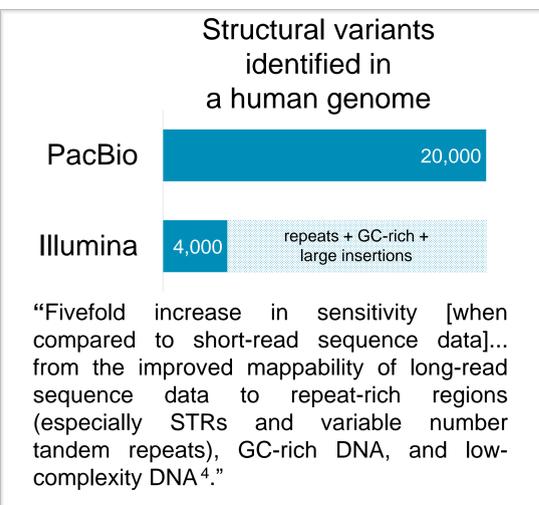


Figure 2. PacBio long reads have 5-fold increased sensitivity for structural variants compared to Illumina short reads.

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 larger 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 a low coverage analysis workflow, structural variant detection with PacBio long-read sequencing is now an affordable and cost effective approach for WGS studies.

Protocol

To identify structural variants from low-coverage PacBio long-read sequencing:

- 1) Prepare SMRTbell library from unamplified gDNA (10 μ g) using 20 kb library preparation protocol with size-selection⁵.
- 2) Sequence to 10-fold depth of coverage on the Sequel System with 10 hr movies with 6 SMRT Cells 1M.
- 3) Call variants with the SMRT Link SV Caller.
- 4) Evaluate structural variant calls and breakpoints by examining read support using IGV^{3,6,7}.

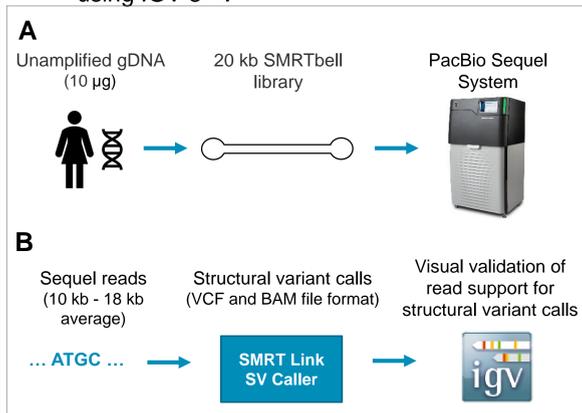


Figure 3. Library preparation, sequencing, and analysis workflow for structural variant discovery and validation.

Visualizing Long Reads in IGV

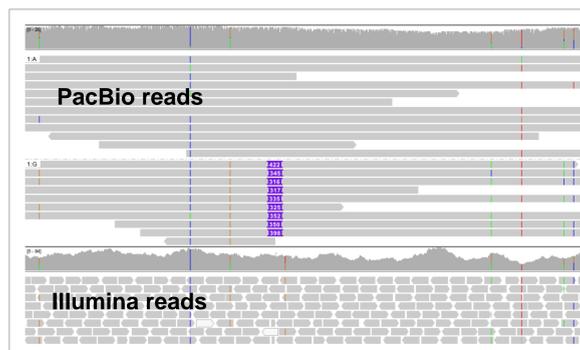


Figure 4. Structural variants in IGV. Improved support for PacBio long reads in IGV 3 makes it easy to see structural variants in phase with single nucleotide variants^{6,7}. PacBio reads agree with Illumina reads at single nucleotides but also show structural variation. (A) insertion at GRCh37 chr13:78,585,000.

Sensitivity vs Coverage

The sensitivity to detect structural variants with PacBio long reads is high even at modest (10-fold) coverage levels.

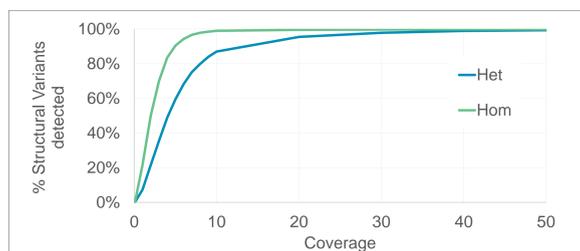


Figure 6. Coverage titration to measure sensitivity of structural variant detection in the diploid human HG00733. Sensitivity to structural variants in a human genome is high even at modest coverage levels. Sequencing was performed on the Sequel System.

Benchmarking with NA12878

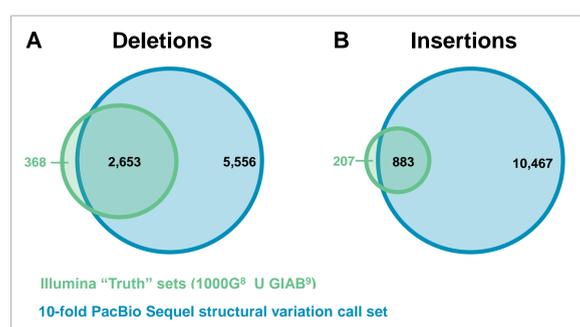


Figure 7. Overlap with truth sets. A 10-fold PacBio call set recovers (A) 88% of true deletions, and (B) 81% of true insertions. The 10-fold PacBio set also includes thousands of novel variants, most of which are directly confirmed by a FALCON-Unzip *de novo* assembly from 60-fold PacBio RS II coverage¹⁰.

Mendelian Disease Case Study (J Merker, EA Ashley)¹¹

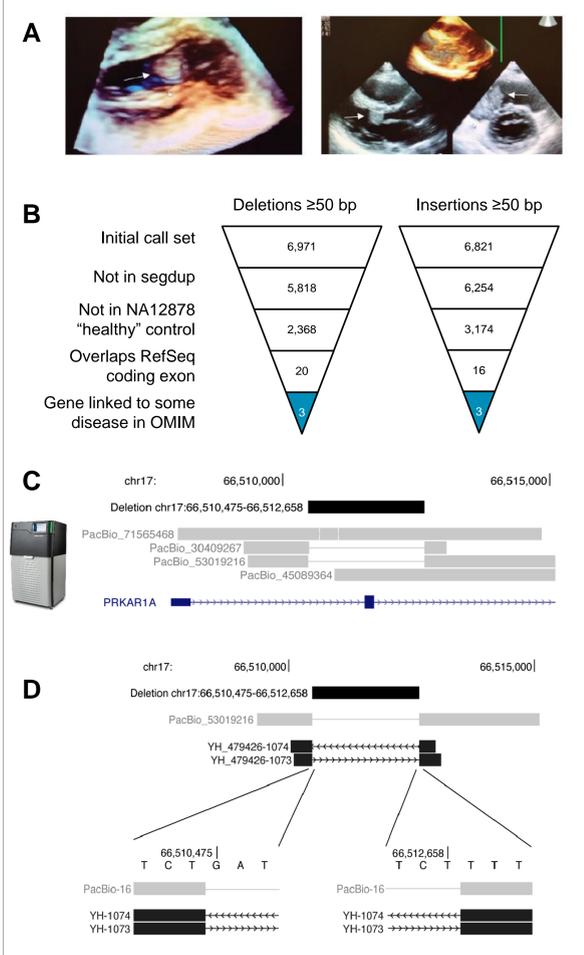


Figure 8. Low-coverage sequencing on the Sequel System identifies a pathogenic structural variant in a Mendelian disease. Targeted gene testing and short-read whole genome sequencing failed to provide a diagnosis for an individual with (A) cardiac myxomata. (B) Low-coverage PacBio sequencing identified 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.

Conclusion

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

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. PacBio Procedure and Checklist – [20 kb Template Preparation Using BluePippin™ Size-Selection System](#)
6. Wenger, A. ["IGV 3 Improves Support for PacBio Long Reads."](#) Web blog post. *PacBio Blog*. PacBio, 29 Mar 2017, Web. 10 May 2017
7. Robinson JT, et al. (2011). [Integrative genomics viewer](#). *Nature Biotechnology*. 29(1), 24-26
8. Sudmant PH, et al. (2015). [An integrated map of structural variation in 2,504 human genomes](#). *Nature*. 526(7571), 75-81.
9. Parikh H, et al. (2016). [svclassify: a method to establish benchmark structural variant calls](#). *BMC Genomics*, 17, 64.
10. Wenger, A. ["Identifying structural variants in NA12878 from low-fold coverage sequencing on the PacBio Sequel System."](#) Web blog post. *PacBio Blog*. PacBio, 19 Oct 2016, Web. 10 May 2017
11. Merker J, et al. (2016). [Long-read whole genome sequencing identifies causal structural variation in a Mendelian disease](#). *bioRxiv*. doi:10.1101/090985.

Acknowledgements

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