Effect of coverage depth and haplotype phasing on structural variant detection with PacBio long reads

Aaron Wenger, Steve Kujawa, Luke Hickey, Jason Chin, Jonas Korlach

Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Introduction

De novo human genome assemblies using high-fold coverage PacBio long reads have revealed tens of thousands of structural variants per genome, most of which are not discoverable with short reads.

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 human genomes *de novo* assembled from PacBio long reads.

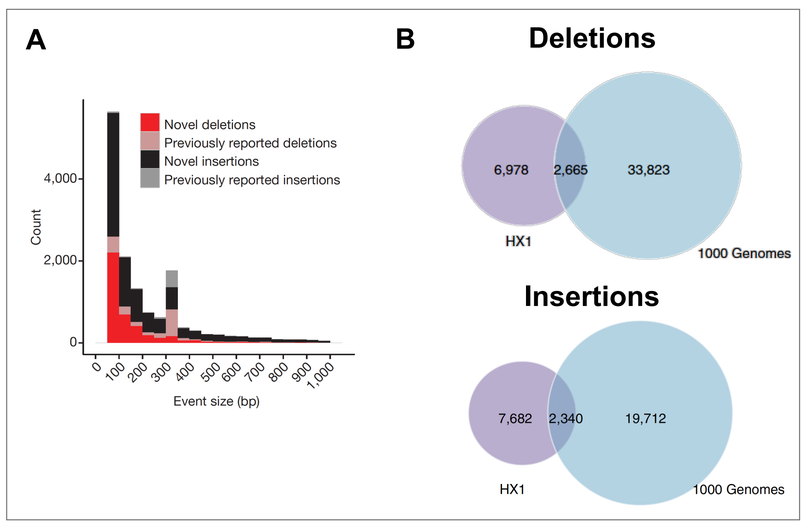

Figure 1. Long reads show increased sensitivity for structural variants. (A) 85% of structural variants identified in CHM1 are novel against prior short read variant databases[1]. (B) A single PacBio genome, HX1, reveals one quarter of the number of variants discovered in 2,504 short read genomes[2].

Higher coverage depth directly translates into both increased sensitivity and increased cost. As efforts turn to analysis of structural variants in large patient and control population cohorts, it is important to strike a balance between cost and sensitivity. Low-fold coverage sequencing on the PacBio Sequel System has the potential to be an effective and affordable solution for structural variant discovery in human genomes.

NA12878 on PacBio Sequel System

To evaluate the power of low-fold coverage PacBio sequencing, we sequenced the human reference sample NA12878 to about 10-fold coverage on the PacBio Sequel System and evaluated the structural variant call set.

Performance at 10-fold Coverage

MAP READS	CALL VARIANTS	EVALUATE
NGM-LR	PBHoney[4]	1000G[5] + GIAB[6]

DNAnexus

Set	Platform	Deletions	Insertions
1000 Genomes	Illumina	1,910	1,090
Genome in a Bottle	Multiple	2,668	n/a
10-fold PacBio	**Sequel**	**7,386**	**7,445**

Table 2. NA12878 structural variant sets.

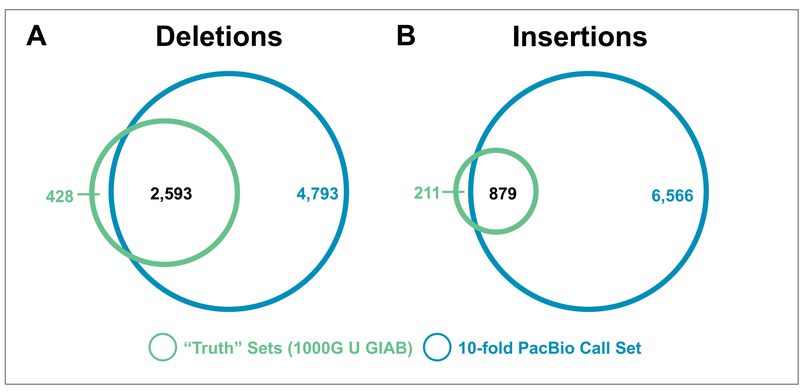

Figure 2. Overlap with truth sets. The 10-fold PacBio call set recovers (A) 86% of true deletions, and (B) 81% of true insertions. The 10-fold PacBio set also includes thousands of novel variants, most of which are directly confirmed by a FALCON-Unzip *de novo* assembly from 60-fold RS II coverage.

Properties of 10-fold PacBio Call Set

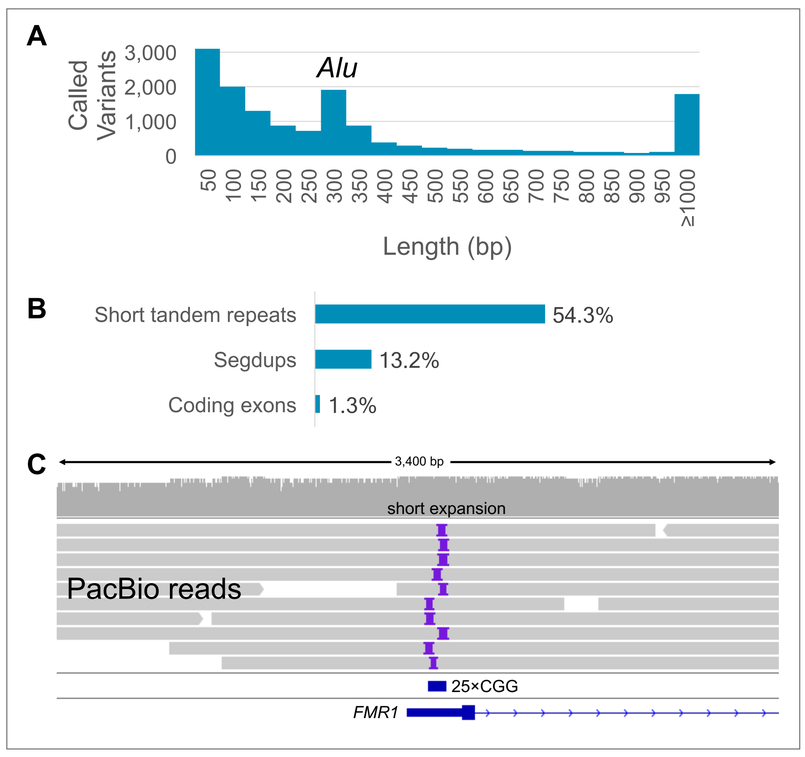

Figure 3. Properties of variants in the 10-fold PacBio call set. (A) The distribution of variant lengths shows an expected peak at 300 bp for *Alu* repeats. (B) Fraction of variant calls that overlap genome annotations. (C) Sensitivity to tandem repeat expansions and contractions is due to the ability of long reads to fully span the repeats.

Visualize and Phase Variants in IGV

Manual exploration is useful to confirm variant calls and evaluate loci of interest. The development version of IGV includes new features to support long PacBio reads.

http://tinyurl.com/pacbioigv

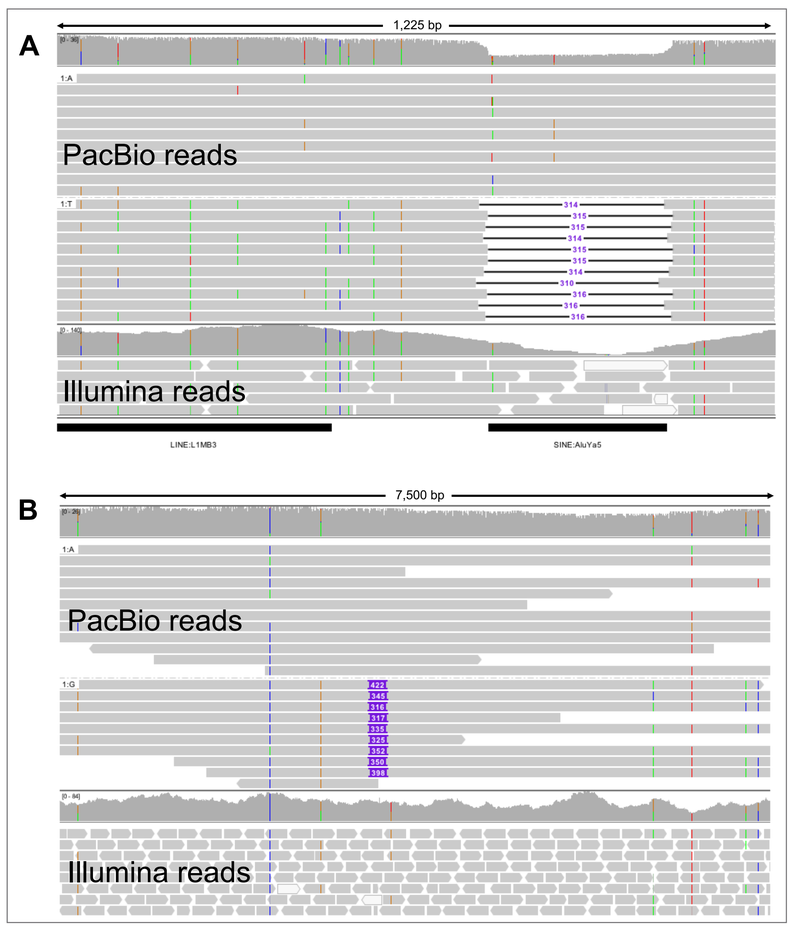

Figure 4. Structural variants in IGV. Improved support for PacBio long reads in IGV makes it easy to see structural variants in phase with single nucleotide variants. PacBio reads agree with Illumina at single nucleotide positions but also show structural variation. (A) deletion at GRCh37 chr16:33,485,217-33,486,446, (B) insertion at GRCh37 chr13:78,581,290-78,588,800.

Conclusion

- Low-fold (10-fold) PacBio sequencing of NA12878 recalls 84% of known structural variants and identifies thousands more not previously seen in short reads.
- Comparison to a *de novo* assembly shows that precision of the 10-fold call set is high, both for known and novel variants.
- Support for PacBio reads in IGV enables exploration and haplotype phasing of short and long variants.

References and Acknowledgments

1. Chaisson MJ, et al. (2015). *Nature*, 517(7536):608-11.
2. Shi L, et al. (2016). *Nat Commun*, 7:12065.
3. Seo JS, et al. (2016). Nature, 538(7624):243-7.
4. English AC, et al. (2014). *BMC Bioinformatics*, 15:180.
5. Sudmant PH, et al. (2015). *Nature*, 526(7571):75-81.
6. Parikh H, et al. (2016). BMC *Genomics*, 17:64.

Thank you to Kevin Eng, Christine Lambert, and Primo Baybayan for data generation; Fritz Sedlazeck and Philipp Rescheneder for access to NGM-LR; and Andrew Carroll for providing PBHoney on DNAnexus.