Structural variant detection in crops using low-fold coverage long-read sequencing

Michelle Vierra and Aaron Wenger
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

Insertions, deletions, duplications, translocations, inversions, and tandem repeat expansions in the structural variant (SV) size range (≥50 bp) contribute to the evolution of traits and often have significant associations with agronomically important phenotypes. However, most SVs are too small to detect with array comparative genomic hybridization and too large to reliably discover with short-read DNA sequencing. While *de novo* assembly is the most comprehensive way to identify variants in a genome, recent studies in human genomes show that PacBio SMRT Sequencing sensitively detects structural variants at low coverage.

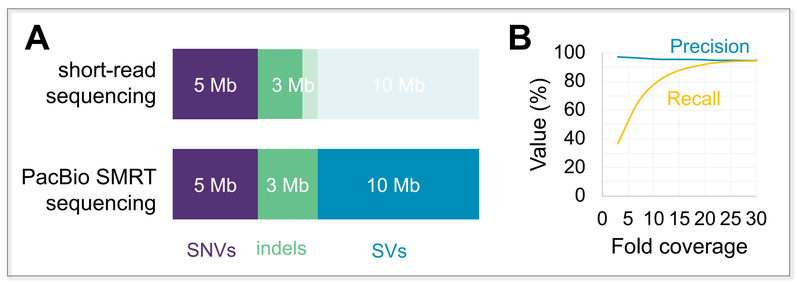


Figure 1. Variation in a typical germline human genome[1]. (A) Most of the base pairs that differ between two human genomes are in indels 1-49 base pairs and in structural variants (SVs), differences ≥50 base pairs. Short-read sequencing has limited sensitivity for indels and SVs, while PacBio long-read sequencing comprehensively detects variants of all sizes. **(B)** Precision (blue) and recall (orange) for SVs in a human genome (HG002) against fold coverage. Recall remains high for ≥10-fold coverage.

Human SV Benchmark

The Genome in a Bottle Consortium[2] has developed a benchmark set of insertion and deletion SVs in a human male, HG002/NA24385. Comparing technologies against this benchmark, PacBio has the highest precision and recall across the structural variant size range, and particularly for insertions.

Technology	Precision	Recall
PacBio	96.13%	95.99%
Oxford Nanopore	83.23%	87.46%
Illumina	85.35%	55.88%
10X Genomics	83.79%	39.83%

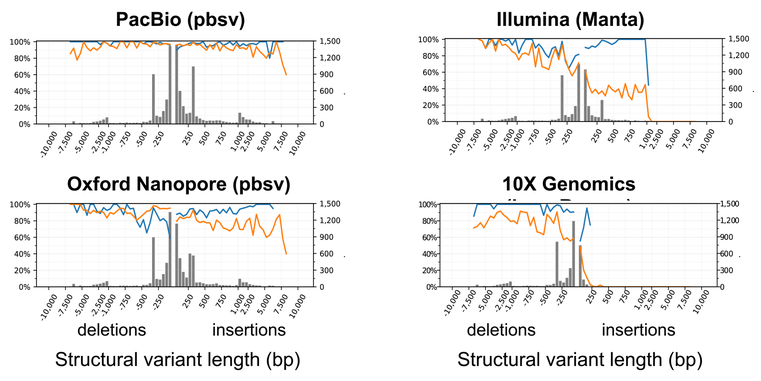


Figure 2. Variant calling performance against the GIAB HG002 v0.6 benchmark. Histograms indicate the number of variants and lines show the precision (blue) and recall (orange) at each variant size for call sets from different technologies.

Methods for SV Detection

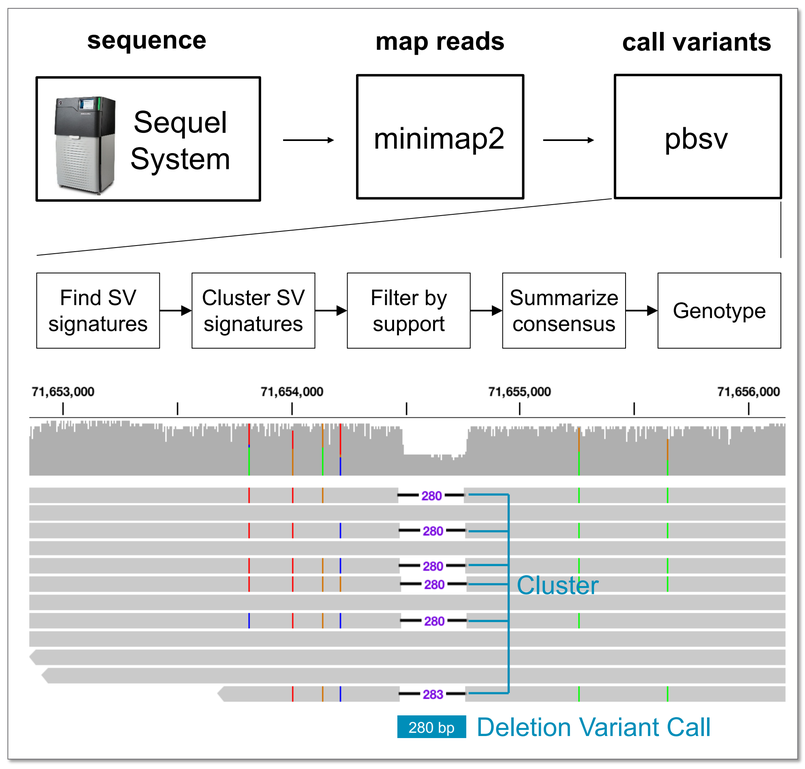
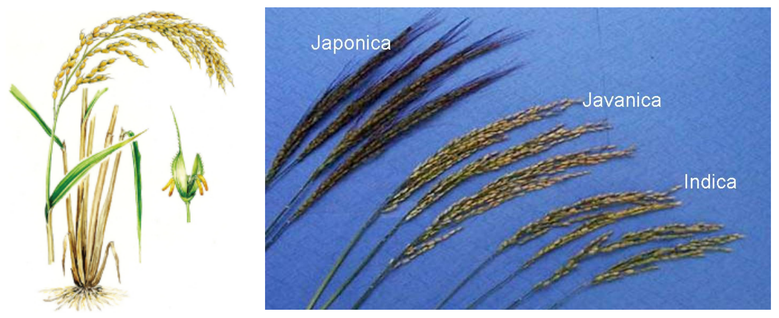


Figure 3. Workflow to detect structural variants from PacBio long reads. To call structural variants, pbsv identifies signatures of structural variation in alignments, clusters nearby signatures with similar length and sequence, summarizes into a consensus call, and assigns a genotype based on read support.

SV Calling in MH63 with Assembly



Cultivar	Accession	Length	N50
Oryza sativa japonica Nipponbare	GCA_001433935.1 (release 7)	374 Mb	7.7 Mb
Oryza sativa indica MH63	GCA_001623345.2 (MH63)	387 Mb	25.6 Mb

Comparison of a *de novo* assembly of the *Oryza sativa indica* cultivar MH63 (ref. 3) to the International Rice Genome Sequencing Project reference assembly of the *Oryza sativa japonica* cultivar Nipponbare[4] provides a baseline of structural variants against which to evaluate variant calling with pbsv. The MH63 *de novo* assembly used 110-fold coverage of PacBio reads.

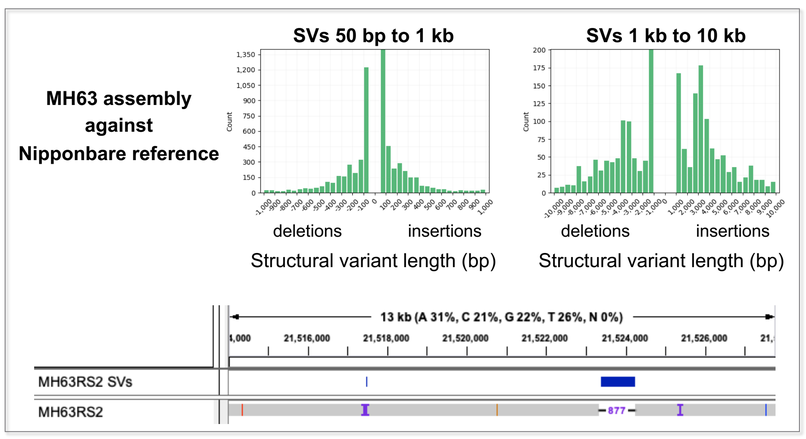


Figure 4. Structural variants in MH63 from assembly-to-assembly comparison. Structural variants were detected by aligning the MH63 assembly to the Nipponbare reference with minimap2 and calling variants with paftools. The assembly-based callset includes 3,739 deletions and 4,328 insertions 50 bp to 10 kb.

SV Calling in MH63 with pbsv

MH63 was re-sequenced to 30-fold PacBio read coverage and downsampled to generate lower coverage. Structural variants were called with pbsv and compared to the calls generated from *de novo* assembly.

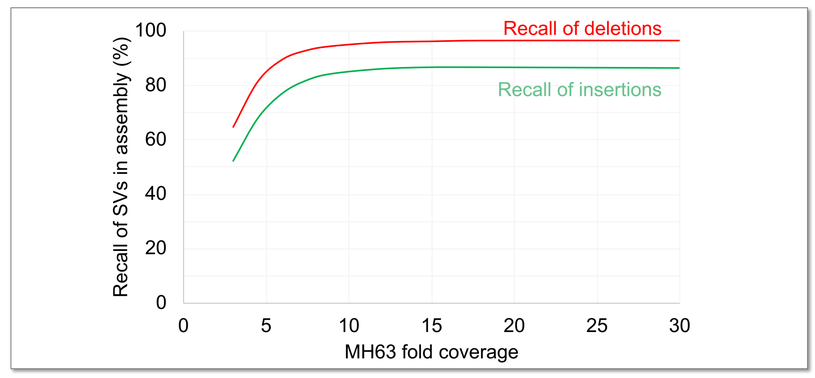


Figure 5. Recall of pbsv at various coverage levels. Recall remains high for coverage ≥10-fold, with the primary limit of sensitivity being large insertions.

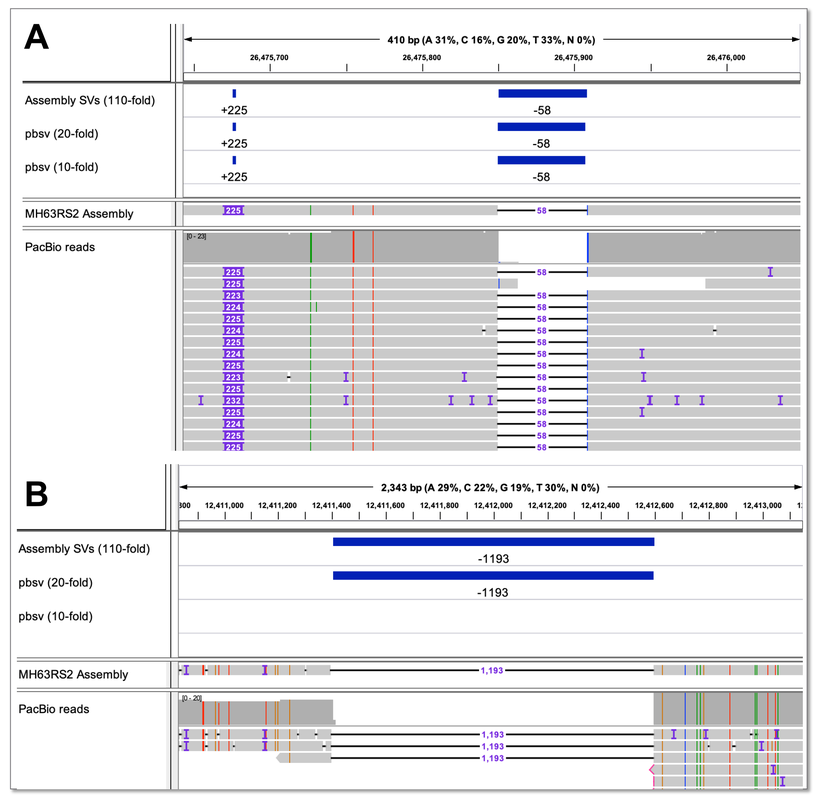


Figure 6. SVs detected by assembly and pbsv. (A) Insertion and deletion detected at 10- and 20-fold coverage. **(B)** Deletion detected at 20- but not 10-fold.

Conclusions

- PacBio sequencing has high precision and recall for SVs in plant genomes.
- SV calling is effective at lower coverage than is *de novo* assembly.
- The workflow to detect SVs is simple and efficient.

References

1. Chaisson MJ, et al. (2017). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. doi:10.1038/s41467-018-08148-z.
2. Zook JM, et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 3:160025.
3. Zhang J, et al. (2016). Extensive sequence divergence between the reference gnomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *PNAS*. 113(35):E5163-71.
4. Kawahara Y, et al. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical mapping. *Rice*. 6(1):4.

Thank you for Dave Scherer, Pamela Bentley Mills, and Kristin Robertshaw for help with poster generation.