Structural Variant Detection in Crops Using PacBio SMRT Sequencing

Greg Concepcion[1], Shreyasee Chakraborty[1], Michelle Vierra[1], Emily Hatas[1], Aaron Wenger[1]
1. PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

Structural variants (genomic differences ≥50 base pairs) contribute to the evolution of traits and disease. Most structural variants (SVs) are too small to detect with array comparative genomic hybridization and too large to reliably discover with short-read DNA sequencing.

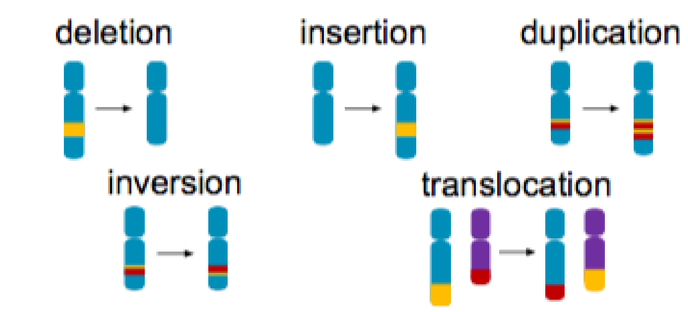

Figure 1. Common types of structural variation

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[1]. Here we present SV characterization in two major crop species grown worldwide, *Zea mays* (Maize) and *Glycine max* (Soy).

Datasets

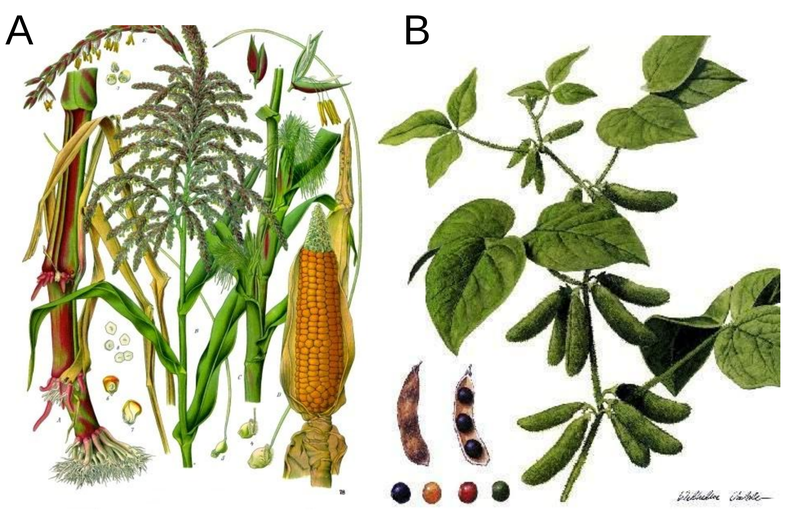

Figure 2. Illustrations; A) *Zea mays* B) *Glycine max*

Species	Cultivar	SRA Accession	Seq Platform	Gbp	Coverage
Zea mays	Mo17	SRR58643 [24:32]	PacBio RSII	48.0	22.7
Zea mays	Mo17	SRR5826129	Illumina	89.9	42.7
Glycine max	Williams	unreleased	PacBio Sequel	24.3	24.9
Glycine max	Wm82	SRR425302 [8:9]	Illumina	34.8	35.6

Table 1. For both species, *Glycine max* and *Zea mays*, two parallel datasets, consisting of both long and short reads, were acquired for comparison.

Methods

Structural variation detection was performed using two parallel pipelines appropriate for the respective technologies. For PacBio long reads, NGMLR[2] was used for mapping and SV detection performed with pbsv, while the standard BWA[3] mapping tool was used followed by SV detection using manta[4] for short reads. Additionally, a subset of the long read data for each dataset was used to investigate the sensitivity of SV detection with low-fold coverage.

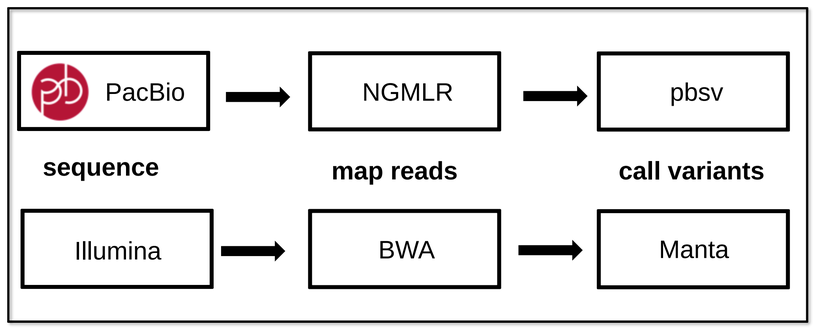

Figure 3. Overall workflow of parallel SV calling pipelines for both PacBio and Illumina sequencing data

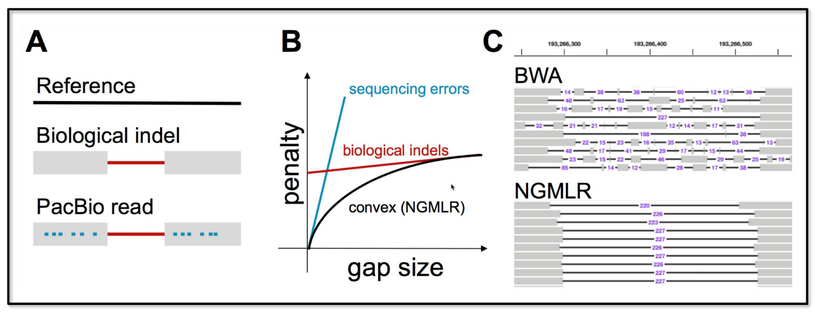

Figure 4. NGMLR correctly aligns PacBio reads around structural variants (A) PacBio reads have indels both from biological variation and sequencing errors. (B) NGMLR uses a convex gap penalty to effectively model the statistics of both types. (C) The same reads aligned with BWA and NGMLR illustrate how NGMLR produces sharp alignment gaps.

Results

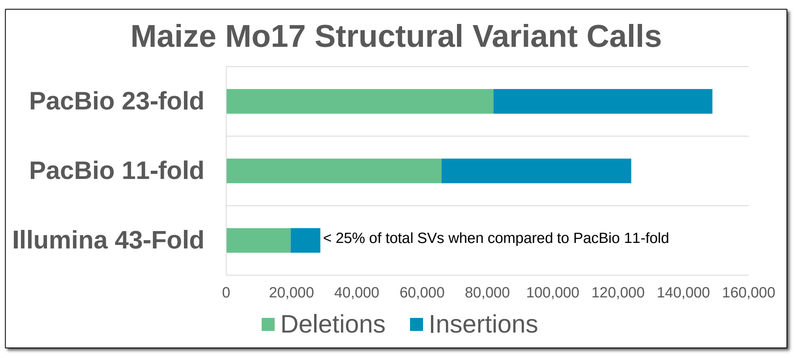

Figure 5. *Zea mays* data mapped to AGPv4 reference DNA from the same strain of *Zea mays* (mo17) was sequenced in parallel with both PacBio and Illumina and subsequently analyzed for structural variant detection. Despite being at a coverage disadvantage, more than 5 times the number of structural variants were detected with PacBio long read technology.

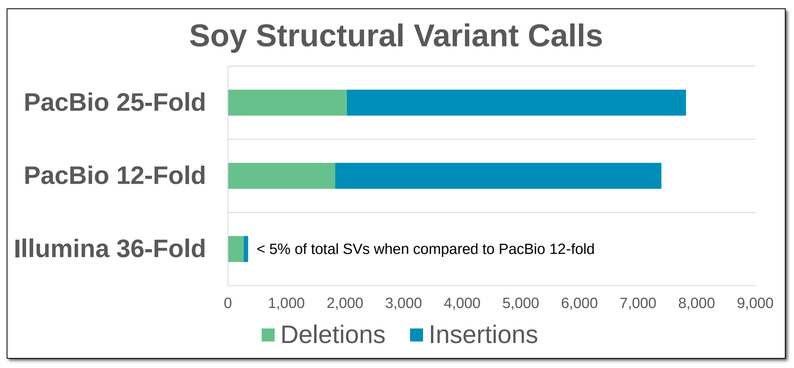

Figure 6. *Glycine max* var Williams data mapped to (Wm82a2) reference DNA from both PacBio and Illumina were mapped to the reference. Despite the PacBio coverage being half that of short reads, 22 times more structural variants were detected with PacBio at 25-fold coverage, and 20 times more at 12-fold coverage.

Aligned reads in IGV

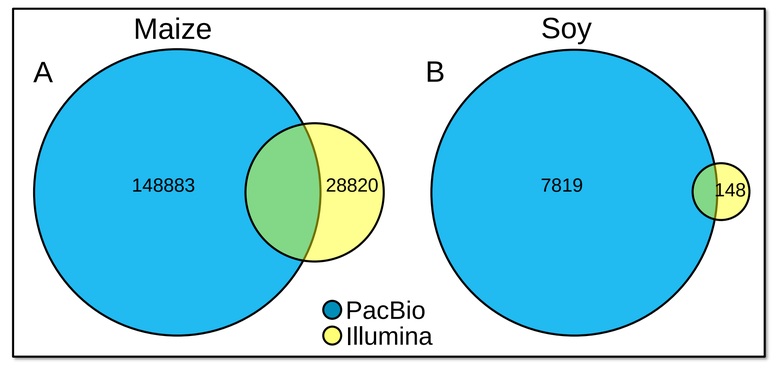

Figure 7. Venn diagrams showing overlap between A) Maize B) Soy structural variation call sets

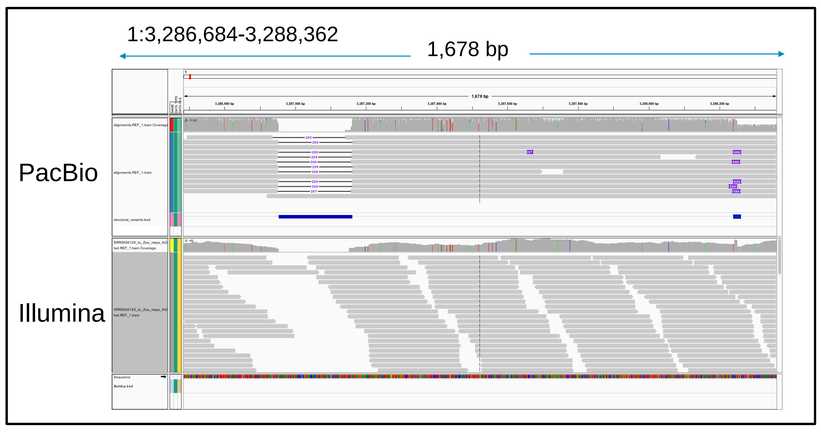

Figure 8. Structural variation in *Zea mays* visualized with IGV IGV 2.4 makes it easy to visualize structural variants in haplotypes. On *Zea Mays* AGPv4 Chromosome 1, SNP locations between Illumina and PacBio alignments are in agreement. In addition, at low-fold coverage PacBio alignments also highlight one large insertion and one large deletion in that are not detected in Illumina alignments.

Conclusions

- Structural variant annotation performed with PacBio long reads detects many more variants than short reads in both maize and soy.
- In soy, SVs account for ~6.4 Mb of sequence while for maize the number is much higher at ~492 Mb. Part of this large number is likely due to strain differences from the reference.
- SV detection with low-fold coverage PacBio data is a viable approach for genomic characterization of crops.

References

1. Chaisson, MJ, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* (2017).
2. Sedlazeck, FJ, et al. Accurate detection of complex structural variations using single molecule sequencing. *bioRxiv* (2017).
3. Li, H, et al. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60 (2009).
4. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32, 1220-1222 (2016).