Characterizing the Pan-genome of Maize with PacBio SMRT Sequencing

Michelle Vierra, Greg Concepcion, Aaron Wenger, David Rank, Paul Peluso
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Abstract

Maize is an amazingly diverse crop. A study in 2005[1] demonstrated that half of the genome sequence and one-third of the gene content between two inbred lines of maize were not shared. This diversity, which is more than two orders of magnitude larger than the diversity found between humans and chimpanzees, highlights the inability of a single reference genome to represent the full pan-genome of maize and all its variants. Here we present and review several efforts to characterize the complete diversity within maize using the highly accurate long reads of PacBio Single Molecule, Real-Time (SMRT) Sequencing. These methods provide a framework for a pan-genomic approach that can be applied to studies of a wide variety of important crop species.

The Maize Reference Genome

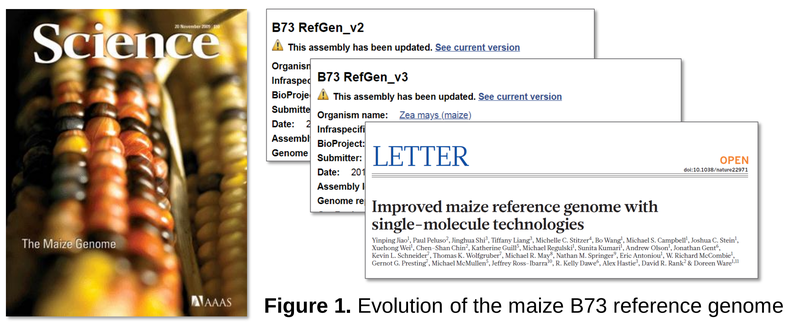

Figure 1. Evolution of the maize B73 reference genome

Published in 2009[2], the original maize B73 version 1 reference genome was the product of a monumental effort to sequence BAC by BAC for a high-quality genome assembly. Incremental improvements to this first version were made with the addition of new data until 2017, when a completely new assembly, version 4, was published[3] with a 30-fold improvement in contiguity for one-thousandth of the cost using PacBio SMRT Sequencing.

	B73 RefGen_v1	B73 RefGen_v4
Method	BAC tiling	PacBio WGS
Cost	$30 million	$30 thousand
Contig N50	40 kb	1.2 Mb

Table 1. Comparison between versions 1 and 4 of the maize B73 reference

One reason for the difficulty in sequencing and assembling the 2.3 Gb maize genome is the high abundance of long terminal repeat (LTR) retrotransposons, some of which are 10-15 kb in length. While less than 1% of the LTR retrotransposons were intact in the annotation of the v3 genome, almost 70% of them were assembled in the v4 assembly[3], largely due to the long, unbiased reads generated by PacBio systems.

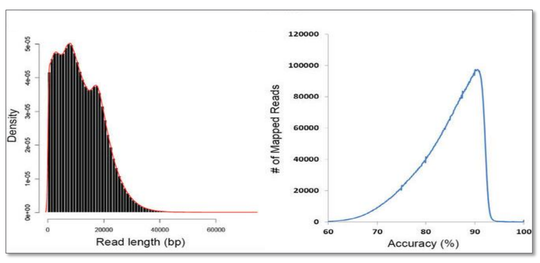

Figure 2. Read length (left) and raw accuracy (right) from SMRT Cells used to generate the v4 maize reference genome[3].

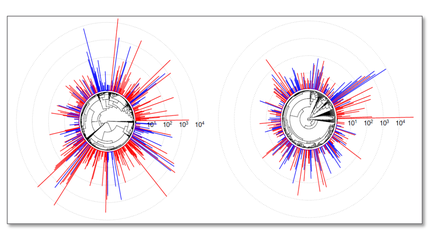

Figure 3. Phylogeny of LTR retrotransposon families. Ty3/Gypsy (left) and Ty1/Copia (right) superfamilies have higher copy numbers in maize (red) than in sorghum (blue)[3].

Improved Genome Annotation

In addition to improving the maize reference genome for a more comprehensive view of genetic diversity, an effort to fully characterize the transcriptome with long reads was undertaken. Using the PacBio Iso-Seq method for RNA sequencing, researchers were able to produce >100,000 non-redundant full-length isoforms, 57% of which were novel. These isoform sequences were able to correct gene models and identify previously missed genes, boosting the quality of the genome annotation[4]. These benefits were largely due to the fact that sequencing cDNA with PacBio generates full-length transcripts with no assembly required.

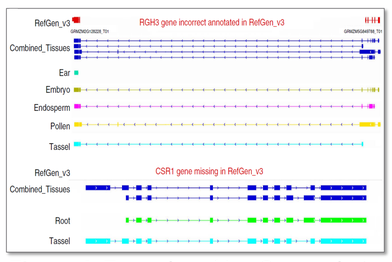

Figure 4. Examples of Iso-Seq analysis correcting gene models (top) and identifying previously missed genes (bottom)[4].

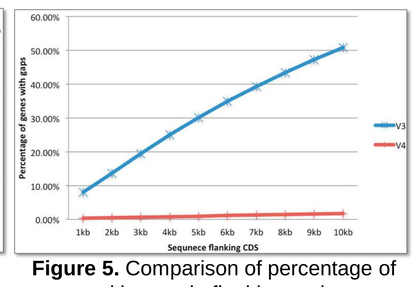

Figure 5. Comparison of percentage of genes with gaps in flanking regions between versions 3 and 4 of the maize B73 reference annotations[3].

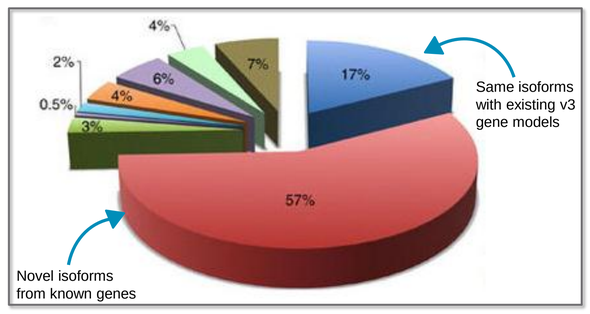

Figure 6. Classification of PacBio and v3 isoforms[4].

Discovering New Variation

The version 4 reference genome, with it's highly contiguous sequence, allowed for a genome-wide evaluation of structural variants in two lines of maize, Ki11 and W22. The optical maps showed only 32% of the Ki11 and 39% of the W22 maps were alignable to the B73 reference, highlighting the tremendous genetic diversity found within *Zea mays*[3].

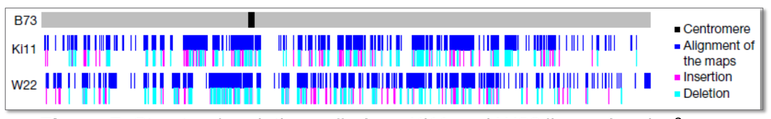

Figure 7. Structural variation calls from Ki11 and W22 lines of maize[3]

Long- and short-read structural variation detection were also performed by PacBio[5] using two parallel pipelines appropriate for the respective technologies. Despite being at a coverage disadvantage, more than 5 times the number of structural variants were detected with PacBio data than with short reads.

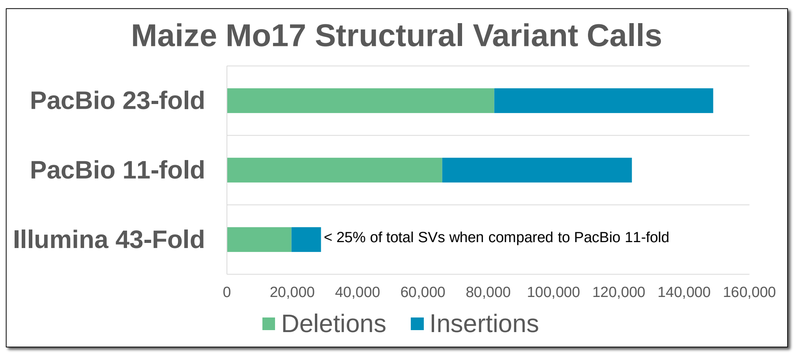

Figure 8. *Zea mays* data mapped to AGPv4 reference for structural variant detection.

The Era of the Pan-genome[7]

Now that high-quality genome assemblies are affordably being generated for many species of plants and animals, it is becoming abundantly clear that the era of the single reference genome is ending. From surveying natural diversity to crop improvement to answering basic questions about biology, generating and analyzing multiple reference genomes allows for the most comprehensive view of variation in plant genomes.

DuPont Pioneer is sequencing the genomes of 100+ maize lines to characterize single nucleotide and structural variation to improve their commercial offerings in maize.

Dario Cantu at UC Davis is sequencing all of the wine grape varieties and their pathogens to identify variants for disease resistance, drought tolerance, and to characterize flavor genes.

The International Oryza Map Alignment Project is generating genome assemblies for all sub-populations of domestic and wild rice to scout for drought tolerance and disease resistance genes[6].

Jeremy Schmutz, of HudsonAlpha and JGI, is generating multiple reference genomes for sorghum, cotton, and peanut to enable functional studies into genotype-phenotype association.

Steve Knapp at UC Davis is sequencing relatives of the commercially cultivated octoploid strawberry to enable family-wide comparative studies.

Figure 9. Multiple reference genome projects currently being undertaken with PacBio SMRT Sequencing technology

Conclusion

- A single reference genome is ***NOT*** enough to represent the genetic diversity within a species or clade
- PacBio SMRT Sequencing provides a workflow for characterizing the pan-genome of crop species, including *de novo* assembly, isoform sequencing for exploring transcriptome diversity, and structural variation detection

References

1. Brunner S, et al. (2005). Evolution of DNA Sequence Nonhomologies among Maize Inbreds. *The Plant Cell*, 17, 343-360.
2. Schnable P, et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science, 326 (5956), 1112-1115.
3. Jiao Y, et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, 546, 524-527.
4. Wang B, et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, 7.
5. Concepcion G, et al. (2017). Structural Variation Detection in Crops Using PacBio SMRT Sequencing. PAG 2018 Poster.
6. Stein J, et al. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nature Genetics*, 50, 285-296.
7. Shackford, S. When A Single Reference is Not Enough. PacBio Blog, 15 Jan 2018.

Acknowledgements

The authors would like to thank everyone associated with the collaborations and publications cited for their contributions.