PIONEERING A PAN-GENOME REFERENCE COLLECTION

THE LEADER IN LONG READ SEQUENCING

PACBIO®

At DuPont Pioneer, DNA sequencing is paramount for R&D to reveal the genetic basis for traits of interest in commercial crops such as maize, soybean, sorghum, sunflower, alfalfa, canola, wheat, rice, and others. They cannot afford to wait the years it has historically taken for high-quality reference genomes to be produced. Nor can they rely on a single reference to represent the genetic diversity in its germplasm.

So an ambitious project has begun: A pan-genome reference collection based on high-speed, high throughput sequencing and assembly, starting with its own elite maize breeding germplasm and other important maize varieties harboring traits of interest.

Maize has a diploid genome with an estimated size of 2.3 Gb to 2.7 Gb and consisting of 10 chromosomes. These are made up of repetitive fractions of transposable elements, ribosomal DNA (rDNA), and high-copy short-tandem repeats (mostly at the telomeres, centromeres, and heterochromatin knob), punctuated by islands of unique DNA that harbor single genes or small groups of genes.

Maize is an incredibly diverse species. In 2005, researchers at DuPont Pioneer published a study which analyzed allelic genome segments in maize inbreds B73 and Mo17 and demonstrated extensive DNA sequence non-homologies. On average, 50 percent of the sequence and one-third of the gene content was not shared between the compared loci. In contrast, while humans and chimpanzees are different species, they exhibit more than 98 percent sequence similarity. In other words, one reference genome is not enough, especially for researchers who are scouring both wild and domesticated varieties for traits like disease resistance.

Many genome assemblies for crop plants have been generated over the past decade with next-generation sequencing (NGS), but these assemblies are often highly fragmented with limited value. By relying on short-read sequencing, these efforts have not adequately re-constituted the complexity of these genomes, where structural variations, in addition to nucleotide polymorphism, play important roles in phenotypic variation.

Kevin Fengler works to develop high-quality reference genomes that underlie hundreds of in-house bioinformatics systems and processes at DuPont Pioneer.

The creation of the first high-quality maize reference genome (B73) by Doreen Ware's team at Cold Spring Harbor Laboratory and the USDA Agricultural Research Service using Single Molecule, Real-Time (SMRT®) Sequencing has benefitted the maize community. But it became apparent to the Pioneer group that they needed reference genomes for additional strains, both commercial and wild.

They also hoped to improve on the time it took to optimize the assembly process. Their efforts have led to the creation of a high-throughput method that whittled the process down to less than one month: two weeks of sequencing and 10 days of assembly, polishing, and chromosome-scale scaffolding.

"The power comes from doing it again and again, refining the process each time" says Research Scientist Kevin Fengler, of the Data Science and Informatics group at DuPont Pioneer. "We've done several now. We know exactly what it takes to make a high-quality reference genome. But we had to develop a workflow to do it in a high-throughput way in order to enable a pan-genome view."

Scientists at DuPont Pioneer are generating a maize pan-genome reference collection based on SMRT Sequencing and *de novo* assembly, including Pioneer's own elite lines and other important maize varieties.

Recipe for Success

So, what does it take? Speedy long-read sequencing, which is achieved at DuPont Pioneer via in-house Sequel® Systems.

A group of sequencing experts, led by Greg May, Research Director of Genomics Technologies, generates the PacBio® data, which is passed on to Fengler to be assembled into contigs using tools such as FALCON for outbred, heterozygous strains or Canu for inbred lines, and polished.

Large, robust contigs are the cornerstone of the reference genome assembly project. Achieving an assembly with a contig N50 greater than 1 Mb is viewed as an accomplishment, recognized among online social networks as entry into the virtual "1 Mb Contig Club". Recently, PacBio long-read SMRT Sequencing has enabled many plant genome assemblies to readily reach this status.

"Our new benchmark for maize is currently a contig N50 of 2.8 Mb. We aim to meet or exceed this standard with each new maize genome sequenced," Fengler says.

Fengler then adds other data types to rapidly elevate contig assemblies into more complete reference genomes. He further polishes the contigs with short read clouds, links contigs into scaffolds with optical mapping, and places scaffolds into chromosome pseudomolecules using Hi-C sequencing data.

Reference genome projects often get bogged down during the transition from contigs to chromosomes, but layering other complementary genomics technologies in-house to the assembly recipe overcomes this traditional hurdle, Fengler says.

The accessibility and utility of contig assemblies and scaffolds are maximized when elevated to the status of a finished reference genome, Fengler says. "We're making reference genomes, not just assemblies. Reference genomes touch hundreds of in-house bioinformatics systems and processes. It's easier to do if all the assemblies are done to the same high level."

> "Now with high-quality PacBio genomes, we can provide researchers with actionable sequence information for gene discovery and product development."

Previous NGS approaches often yielded sequencing information that was incomplete or required validation. "Now with high-quality PacBio genomes, we can provide researchers with actionable sequence information for gene discovery and product development," he adds.

Having multiple genome assemblies of the same high standard for several genotypes will be increasingly important as researchers try to achieve a greater understanding of the impacts of structural variation on plant genomes. Otherwise, mis-assemblies and mis-placements of contigs will be interpreted as variants.

"We want to focus on true structural variation and have confidence in the new discoveries we find in these genomes," says Fengler.

Streamlining Whole Genome Sequencing

SMRT Sequencing does not require an army of people, nor does it take a ton of money. A handful of researchers can generate high-quality reference genomes in a cost-effective way, and Fengler says the process can actually conserve resources by simplifying workflows.

Some of the tried-and-true methods are becoming obsolete due to the speed and cost efficiency of high-quality whole genome sequencing. Even if Pioneer scientists are interested in just one gene, sometimes the most straightforward approach is to sequence and assemble the entire genome.

"In the past, we may have considered several targeted genomics approaches, such as BAC sequencing or target capture, to obtain the sequence information that we need. But now often the quickest, easiest, and most economical way to go is whole-genome. In the time it would take us to make a BAC library, screen it, and sequence the target region, we can have the whole-genome assembly complete," Fengler says.

Whole-genome assembly coupled with long-read sequencing is particularly powerful and useful when going after disease resistance genes that are clustered in complex regions, for example. Long reads are needed to tease them apart.

DuPont Pioneer has already seen enormous benefit from its deeper exploration of maize genomics.

"For the first time, we are seeing aspects of the maize pan-genome we never knew existed," Fengler says. "Until now, focusing on B73 reference genome has limited our view. We are just beginning to explore what we have been missing all along."

PLANT AND ANIMAL SCIENCES

PN: CS121-010518