Scientists at the USDA and Cold Spring Harbor Laboratory know that better breeding of maize to feed a growing population will depend on an accurate reference assembly. They tackled the previously intractable crop with a combination of PacBio® Sequencing and BioNano Genomics® genome maps, leading to the first-ever high-quality reference assembly.

When Doreen Ware and her team’s latest genome project is complete, the plant sciences community will have a critical new tool that once seemed virtually impossible: a robust reference assembly for the maize genome. This resource will support breeding of hardier, higher-yielding lines of maize, the number one crop plant in both the U.S. and China.

Ware, a computational biologist with the USDA at Cold Spring Harbor Laboratory, says climate change and protecting the environment are major challenges facing agriculture. “We know that we must increase yield in order to meet the expected 9 billion people in less than 25 years,” she says. “Having the genome and the genome content allows us to accelerate the improvement of maize varieties and the germplasm.”

Maize, with about 2.3 Gb of sequence, seems almost designed to evade genomic characterization. Previous sequencing efforts with Sanger and other technologies were stymied by the plant’s complex universe of transposable elements and highly repetitive regions. The genome differs significantly between plants: the gene complement of any two maize plants, even of the exact same variety, can differ as much as five percent. For population-level crop improvement, Ware says, this field needs not one reference assembly but many.

Her team is well on its way to delivering the first truly high-quality reference assembly for maize, with plans in the works to produce several more in the near future. This impressive achievement was based on pairing two technologies: Single Molecule, Real-Time (SMRT®) Sequencing from PacBio, and Next-Generation Mapping from BioNano Genomics’ Irys® System. Together, these approaches have enabled Ware and her team to produce an assembly with greater contiguity and accuracy than has ever been possible for this challenging genome, providing the first-ever look at important regulatory and structural elements that could influence breeding approaches.

The Maize Enigma

Many plant genomes have been difficult to sequence, but maize has a number of features that confounded previous efforts. The vast majority of its genome – a staggering 85 percent – is made up of transposable elements. Its repeats were copied more recently than those in many other organisms, resulting in higher sequence identity between them. Genome architecture is also different in plants than in humans, Ware says: maize introns are fairly compact, at around 500 bases, while exons average about 300 bases.

The first reference assembly for maize was released in 2009. Using a BAC-based approach with Sanger sequencing, this assembly missed about 10 percent of the genome, according to Ware. “Many of the BACs did not have the order and orientation of the contigs within them,” she adds. “Because of the challenges of repeats, the genome was not tractable with Sanger technology.”

Sequence data from a 454 de novo assembly was added later in an attempt to represent the gene space missed by the BACs, but this did not address the issues of contig order and orientation, nor did it accurately cover the repetitive regions. Full-length transposons in maize can stretch across 10-15 kb, Ware says. With any sequence data that does not span the full regions of that length, “you can’t tell where the appropriate mapping of these repeats is,” she adds.

Despite these pitfalls, characterizing the maize genome has become increasingly important for the advanced crop breeding needed to feed the global population. “Having the baseline of the genes and their organization is critical,” Ware says. This information will make it possible for scientists to identify better genetic biomarkers for traits associated with improved outcomes.

Combining PacBio and BioNano Technologies

Ware and her team knew the only thing that would reveal the secrets of the maize genome was sequence information spanning very long repeats. She brought together two technologies that offered to do just that: PacBio SMRT Sequencing and BioNano Next-Generation Mapping. The PacBio System routinely generates multi-kilobase reads, with a high fraction of the reads being tens of kilobases in length. By producing a de novo
assembly from PacBio data, Ware’s team was able to get nucleotide-level resolution, high consensus accuracy, and reads spanning even the toughest transposons. Meanwhile, the BioNano technology stretches across hundreds of kilobases to provide a genome-wide view essential for correctly scaffolding the sequence data.

The team began with de novo PacBio sequencing, generating an assembly of stunning quality. The previous version of the publicly available assembly for maize (B73 RefGen_v3), which was based on Sanger-sequenced BACs and 454 sequencing data, was broken up into about 140,000 contigs; the PacBio assembly has just 3,300 contigs. Contig N50 lengths rose from about 19 kb in the previous assembly to more than 1 Mb in the PacBio assembly.

“The contigs that we’re looking at now have almost no unknown bases in them,” Ware says. In every chromosome, existing sequence length was extended, filling in gaps that had previously peppered the assembly. The new assembly includes most of the centromeric sequence across all 10 chromosomes and even represents portions of the telomeres.

“In the gene space that we’ve looked at where there were underlying inaccuracies that didn’t allow us to build a correct gene model, almost all of those have been corrected,” Ware says. “Where we were missing portions of sequence, we filled it in. It’s a huge difference.”

The next step was layering BioNano Mapping into the new assembly. Again, the results represented a significant improvement made possible only by pairing the technologies: The total contig count was reduced to just 768, while the contig N50 length increased to more than 9.5 Mb.

Contig length means more to Ware than just a shiny metric: it tells her about the relatedness of genes contained in each section of the assembly. With the previous assembly, contigs only connected two to five genes. “Now we’ve gone from five genes to knowing the order and orientations of hundreds and maybe thousands of genes. Now these have context,” Ware says. She notes that this information is critical for crop breeding, because the ability to use genetics to introduce new traits or influence existing ones may be influenced by where the gene lies on the chromosome.

“What the PacBio assembly and the BioNano map allowed us to do is to achieve improved contigs and scaffolding,” Ware says. She says that each tool offers value – nucleotide-level resolution and long reads with PacBio sequencing, and massive scaffolds with BioNano Mapping – but that together they’re even more effective. “The combination gives you a better product than either technology does alone,” she adds.

Exploring the Maize Genome

The new assembly has allowed Ware’s team to dig into the regulatory mechanisms and structural variants in the maize genome. “We couldn’t do that before,” she says. Structural variation accounts for a lot of the phenotypic plasticity seen in other organisms, Ware adds, and her team anticipates the same will hold in maize.

Being able to view repeat content gives the team its first clear look at gene regulation. “Which transposable element is near a gene may determine the methylation status surrounding that gene, and the methylation status can directly impact expression,” Ware notes. “With the complete reference, we can start to understand how these genes are regulated.”

Now, the team is plowing ahead with plans to sequence three or four additional maize genomes in the coming year. “We don’t want just one reference, we want hundreds of references,” Ware says. “Having more genomes will let us see the full complement of genes within the species, how much of the regulatory regions are conserved and how diverse they are.”

The combination of PacBio and BioNano technologies will serve the team well as they start this next phase of analyzing maize. Ware and her collaborators aim to achieve high quality for each assembly, using long-read sequencing with physical maps to provide optimal resolution. As they generate more assemblies, “we’ll have for the first time a really robust set of resources to start looking at changes between individuals in the population,” Ware says.