

FOR CASHEW TREE, CHROMOSOME-SCALE GENOME ASSEMBLY GENERATED TO IMPROVE BREEDING



Scientists in Brazil paired PacBio long-read sequencing with Dovetail Genomics chromatin proximity ligation to generate a highly contiguous genome assembly for the cashew tree. With this resource, they are on their way to improving breeding programs to protect the plant from disease and boost yield.

Beyond being a beloved snack and protein source, cashews are a big business. Cashew nuts, apples, and oil are consumed around the world, comprising a multi-billion-dollar diverse market. For such an important crop, breeding programs need to evolve to optimize health and production. To help improve the cashew tree's natural ability to resist disease and boost yield, scientists in Brazil generated a reference-grade genome assembly that was used to develop a large-scale SNP genotyping platform for accelerated breeding.

Research scientist Dario Grattapaglia at EMBRAPA, the Brazilian Agricultural Research Corporation, launched the cashew genome project in 2014. A key motivation was the almost complete absence of genomic resources, including a reference genome, for a tree that is native to the country and is used in many products: juices, ice cream, bioactive compounds for pharmaceutical use, and even chemicals and intermediates for chemical industries.

Breeding programs for trees can take much longer to achieve success than for other plants. "Whenever you try to breed a tree, the generation times are much longer than they would be for crop plants," Grattapaglia says. "It takes several years before you have your fruit and can measure whatever needs to be measured for selection." With a high-quality genome assembly, he and his team aim to develop marker-based prediction models that can be used early in the tree's development to accelerate breeding selection decisions. "We have a large interest in speeding up breeding to help produce new varieties that are more resistant to diseases, are more productive, and have increased concentrations of different compounds like sugars, oils, and bioactive molecules," Grattapaglia adds.

He and his team turned to long-read sequencing from PacBio to generate an impressive assembly, following that with proximity ligation from Dovetail Genomics to further improve the assembly. The results are expected to deliver new insight into the dwarf cashew tree *Anacardium occidentale*, a highly productive variety developed by EMBRAPA breeders and the most widely planted in Brazil.

Sequencing Through Repeat Regions

This project was made possible by the rapidly falling cost of sequencing. Five or ten years ago, Grattapaglia says, sequencing a plant genome would have required millions of dollars and an army of scientists. Now, technology and throughput improvements made the cashew genome project feasible for Grattapaglia's three-person lab team.

The scientists had enough sequencing experience to know that short reads would not produce the kind of quality they needed in a genome assembly. "Plant genomes are complex: they are often highly heterozygous, have lots of paralogous sequences, large gene families, and a lot of repetitive DNA," Grattapaglia says. "When you just use short reads, you have major problems in trying to assemble them. There are lots of published plant genomes that are not very good quality because they were done exclusively with short-read sequencing, so there are lots of 'holes' in these Swiss-cheese-like genomes."

"There are lots of published plant genomes that are not very good quality because they were done exclusively with short-read sequencing, so there are lots of 'holes' in these Swiss-cheese-like genomes."

The scientists had previously worked with Single Molecule, Real-Time (SMRT®) Sequencing from PacBio and knew that its extraordinarily long reads could span large repeats and provide information on haplotype phasing. Since repetitive elements may comprise over 40% of the cashew genome, and accurate haplotype phasing is crucial in imputation for genomic prediction and association studies, the team bet that long reads would be crucial to making sense of the genome.

Grattapaglia chose the McGill University and Génome Québec Innovation Centre as his SMRT Sequencing service provider. Research scientist Orzenil Silva-Junior at EMBRAPA managed the bioinformatics analysis of the sequence data. By using the FALCON genome assembly algorithm and Quiver polishing tool, Silva-Junior generated an assembly with a contig N50 that broke the megabase mark at 1.05 Mb. That's an order of magnitude better than the best contig N50s seen in typical short-read assemblies, he notes. The genome assembly covered more than 450 Mb in fewer than 2,600 contigs. "It was a very good starting result for a complex plant genome," Silva-Junior says.



Scientists in Brazil plan to use the high-quality genome assembly to accelerate breeding of the cashew tree whose nuts, fruits, and oil are used for food and industrial applications around the world.

Proximity Ligation

With the PacBio assembly, the team already had a draft genome that surpassed the quality of many published plant genomes. With their mandate to make the information useful for breeding programs, the scientists decided to add orthogonal data to improve the results even more.

They considered optical mapping, which is frequently paired with SMRT Sequencing to boost scaffold length, but getting enough high molecular weight DNA was a major obstacle. While participating in a research scholar program as an affiliate at the U.S. Department of Energy Joint Genome Institute (DOE JGI), Silva-Junior heard about Chicago + HiRise, a service from Dovetail Genomics that also leads to extremely long scaffolds, but didn't have the same burdensome DNA requirements. Dovetail's all-in service model was an added bonus. "This commitment from Dovetail to get the project working from the DNA sample extraction was important for us," Silva-Junior says.

The proximity ligation approach generates extremely long-range information that can be used to order and orient contigs into long scaffolds. Dovetail's team produced in

vitro proximity ligation Chicago™ libraries, sequenced them on an Illumina HiSeq, and assembled the data into scaffolds using its HiRise™ software pipeline. Initially, the contigs of the PacBio assembly were assembled with a 4.3 Mb scaffold N50, with most of the genome in fewer than 400 scaffolds. In addition, Silva-Junior and Grattapaglia took advantage of a new service offered by Dovetail to produce an *in vivo* proximity ligation Hi-C library that was sequenced on an Illumina HiSeq4000 and run through the HiRise pipeline. Using the Hi-C data on top of the PacBio+Chicago assembly, Dovetail's team achieved a final scaffold N50 of 17.05 Mb and L50 of 10 scaffolds. The final assembly includes 90% of the sequences in just 20 scaffolds greater than 10 Mb each (scaffold N90 of 11 Mb and L90 of 20 scaffolds), representing nearly 75% of the expected genome size of 490 Mb and almost the totality of the number of expected chromosomes for the species ($2n=42$).

Next Steps

Going forward, Grattapaglia's team will perform additional quality assurance on the assembly, verifying the joins that Dovetail made and checking other measures of integrity. They are also using the PacBio

data, Dovetail's Chicago and Hi-C data, and various bioinformatics algorithms to fully phase the genome. Silva-Junior plans to take advantage of the high-quality SNP set, which was generated for developing the genotyping platform and was based on both short- and long-read sequencing, to continue improving the assembly's contiguity. The scientists intend to publicly release a first version of the reference-grade assembly in the Phytozome comparative genomics platform at the DOE JGI. This release will include gene predictions from the JGI annotation pipeline, which combines homology-based and *ab initio* methods with cashew RNA-Seq data, as well as sequence variant data derived from 25 whole-genome resequenced cashew accessions. This genomic resource will be very useful for scientists and breeders around the world working with cashew and other plants of the *Anacardiaceae* family, such as mango and pistachio.

At EMBRAPA the priority is to apply this valuable new resource to cashew breeding programs in Brazil. "Having a good reference genome will let us leverage the SNP genotyping platform capabilities to increase the quality and speed of the breeding process," Grattapaglia says.

The scientists were so impressed by the results that they already have new projects in the works that will make the most of PacBio long-read sequencing and Dovetail's proximity ligation service. "At the end of the day, a combination is really very powerful," he says. "We hope to repeat the same success with several tropical fruit genomes."

Grattapaglia also hopes that the cashew project can inspire other small teams to conduct their own genome projects. Thanks to great service providers, three EMBRAPA scientists were able to produce this high-quality assembly affordably and quickly. "It's an interesting example of how you start from scratch – from zero information on the genome – use PacBio sequencing and then use these new proximity ligation methods to really boost the quality of the assembly to the chromosome scale," he says.

