Software locates sugarcane genes of interest

15 May 2019, by Peter Moon

Brazilian researchers develop a program for high-performance computers to map specific portions of plant DNA faster and less expensively for use in breeding more productive and stress-resistant varieties. Credit: Léo Ramos Chaves/Pesquisa FAPESP magazine

Plants have larger and more complex genomes than all animals, be they mammals, birds, reptiles or amphibians. Fish are the exception to the rule.

Human DNA consists of some 3.2 billion base pairs spread out over 23 pairs of chromosomes, for a total of 46 chromosomes. The genome of wheat (Triticum aestivum), however, comprises 17 billion base pairs divided into 21 pairs of chromosomes (a total of 42). The genome of sugarcane (Saccharum spp.) contains 10 billion base pairs in 100-130 chromosomes.

The sugarcane grown today is a hybrid (S. hybridum) cross-bred from two species, S. officinarum—the original sugarcane domesticated in India 3,000 years ago—and S. spontaneum.

"The sugarcane genome has become a giant. It's very hard to work with it using current genomic methods. Deciphering it requires a huge amount of computing power. It's difficult even with state-of-the art computers in processing terms, and they're expensive. In sum, this is a challenge for bioinformatics," said Marcelo Falsarella Carazzolle, bioinformatics coordinator in the Genomics and Bioenergy Laboratory (LGE) at the University of Campinas's Biology Institute (IB-UNICAMP) in São Paulo State, Brazil.

"For years, laboratories in various parts of the world have tried and failed to map the sugarcane genome. The first successful endeavor was completed only a few months ago by a consortium of researchers in several countries, including Brazil," Carazzolle said.

The strategy deployed by the consortium involved massive large-scale computing and heavy investment to sequence the whole genome, i.e., all 10 billion base pairs.

In an article published in the journal *DNA Research*, Carazzolle and colleagues present a different strategy that is much less costly and time consuming. This technique is designed to map specific portions of the genomes of polyploid plants.

Some of the research underpinning this innovation was performed for a Ph.D. thesis by Karina Yanagui de Almeida and for a postdoctoral project by Juliana José. Both are biologists at IB-UNICAMP and were supervised by Professor Gonçalo Amarante Guimarães Pereira. Brazil's National Council for Scientific and Technological Development (CNPq) also provided funding.

"We developed the software necessary to reconstruct these complex genomes and applied it to sugarcane. We weren't trying to assemble the whole genome. Previous studies set out to reconstruct the plant's entire DNA, but our strategy consisted of focusing on small portions corresponding to about 1%-2%, exactly where the genes of interest for plant breeders are located,"

Carazzolle explained.

This strategy saved at least two orders of magnitude compared with the tens of millions of dollars it would cost to map the whole genome. When the project was completed, the consortium had not yet published their results, so the Brazilian geneticists had to use publicly available data—such as the genomes of sorghum, rice and corn, which are related to sugarcane to a greater or lesser extent—to locate the areas they wanted to decipher in the analogous regions of the sugarcane genome.

Selection by analogy was possible because all grasses have a common ancestor that existed more than 50 million years ago. In other words, after all this time, the DNA of any grass today—sugarcane, wheat, sorghum, rice, corn, etc.—still preserves the original core structure, alongside the billions of mutations that have occurred over the eons.

Gene assembler

The outcome of the research conducted at IB-UNICAMP was a software package called Polyploid Gene Assembler (PGA). "PGA represents a novel strategy for assembly of a genetic space based on complex genomes using low-coverage DNA sequencing," Carazzolle said.

Although PGA requires less computer power than the massive processing of a polyploid's whole genome, a very large system is still required to run the program in a timely manner. In this case, the researchers used the computer cluster belonging to the Center for Computing in Engineering & Science (CCES), one of the Research, Innovation and Dissemination Centers (RIDCs funded by São Paulo Research Foundation—FAPESP. Carazzolle is the principal investigator for bioinformatics at CCES.

"The project required the use of CCES's high-performance computers with plenty of memory," Carazzolle said.

They loaded PGA with known gene loci from public genome databases, deploying assembly strategies to construct high-quality genome sequences for the species investigated, and validated the procedure with wheat (Triticum aestivum), a hexaploid species, using barley (Hordeum vulgare) as a reference. More than 90% of the genes were identified, as well as several new genes.

In addition, they used PGA to assemble the genes from grass species S. spontaneum—grouped in the same genus as traditional sugarcane (S. officinarum), S. spontaneum is used in the parental lineage of the hybrid sugarcane varieties widely grown today (S. hybridum).

"We identified a total of 39,234 genes, 60.4% of which were clustered into known grass gene families. Thirty-seven gene families were expanded when compared with other grasses. Three stood out for the number of gene copies potentially involved in initial development and stress response," Carazzolle said.

"Our findings for the genome of S. spontaneum highlighted for the first time the molecular basis of certain significant characteristics, such as high productivity and resistance to biotic and abiotic stress. These results can be used in future functional and genetic studies. They will also support the development of new sugarcane varieties.

"Using PGA, we provided a high-quality assembly of gene regions in T. aestivum and S. spontaneum, proving that PGA can be more efficient than conventional strategies applied to complex genomes and using low-coverage DNA sequencing. PGA's low memory requirement in comparison with the conventional assembly strategy is also an advantage."

Carazzolle stressed that even with significant advances in sequencing technology, the assembly of complex genomes still represents a bottleneck, owing mainly to polyploidy and high heterozygosity. The development of new bioinformatics efforts, he added, can help overcome these constraints, especially in the case of the whole genomes of closely related organisms, for which reference-guided assembly methods can be used.

More information: Leandro Costa Nascimento et

al, Unraveling the complex genome of Saccharum spontaneum using Polyploid Gene Assembler, *DNA Research* (2019). DOI: 10.1093/dnares/dsz001

Provided by FAPESP