DEVELOPMENT OF AN ALGORITHM FOR THE ASSEMBLY OF THE SUGARCANE POLYPLOID GENOME

Glaucia Mendes Souza
Chemistry Institute / University of São Paulo (IQ/USP)
FAPESP Process 2012/51062-3 | Term: Feb 2013 to Jan 2016 | PITE – Business partner: Microsoft

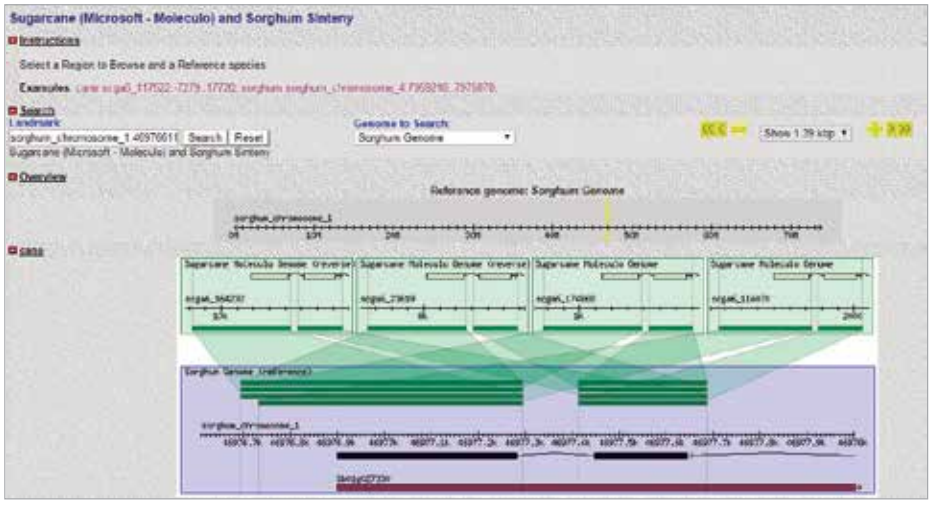

Figure 1. Sorghum and Sugarcane sinteny for one selected sorghum gene showing multiple sugarcane contigs

The challenges in sequencing the sugarcane genome relies in the assembly and analysis of a highly complex genome that is polyploid and aneuploid, with a complete set of homeologous genes predicted to range from 10 to 12 copies (alleles). The present project aims to improve, through the development of new algorithms, the assembly from shotgun reads of datasets generated by the BIOEN research groups. We intend to combine shotgun sequencing data obtained using 454 and Illumina using different protocols to assemble a reference genome for cultivar SP80-3280. To accomplish assembly the group will develop algorithms to enable the recognition of multiple copies of any given region of the reference genome. The group will perform comparative analysis with other genomes, especially sorghum that is very synthenic to sugarcane, to evaluate the quality of the assembled sequence. Existing data on the sugarcane transcriptome will be integrated to validate gene models and infer gene function. Microsoft will develop algorithms for assembly of the sugarcane polyploid genome to strengthen allele identification and improve assembly. The results will contribute to the advancement of knowledge in plant genomics, establishment of computational infrastructure and the formation of highly qualified human resources in bioinformatics. In the long term, we expect to create computational biology tools to resolve polyploid genomes first with the sugarcane genomics but that may also be applied to other cultivated polyploid crops.

SUMMARY OF RESULTS TO DATE AND PERSPECTIVES

Several sequencing platforms (Roche 454; Illumina mate-pair, paired-end and Moleculo) with different characteristics were used to obtain sequences from the sugarcane genome. Approximately 98% of the CEGMA genes were identified in our assembly. Single-copy sorghum genes were aligned to the assembly and the identified copy number varied between one and 15. This shows that it was possible to separate the sugarcane allelic copies in the assembly obtained. Using a custom gene prediction pipeline it was possible to identify approximately 375,000 genes. Orthologues comparisons and transcription active regions identification will be performed to characterize the sugarcane genome.

Glaucia Mendes Souza

Instituto de Química
Universidade de São Paulo (USP)
Departamento de Bioquímica
Av. Prof. Lineu Prestes 748, sala 954
CEP 05508-900 – São Paulo, SP – Brasil

+55-11-3091-8511
glmsouza@iq.usp.br