Towards Deciphering the Sugarcane Pan-Genome

Gustavo Carvalho do Nascimento[1] ; Diego Mauricio Riaño-Pachón[1]
1 Computational, Evolutionary and Systems Biology Laboratory - LabBCES - CENA/USP, Brazil
(gustavocn@usp.br; diego.riano@cena.usp.br;)

INTRODUCTION

Sugarcane (*Saccharum* spp.) is one of the most important tropical crops. It is a polyploid organism, and modern comercial varieties are interspecific hybrids that have different number of copies of each chromosome. Due to this complexity, assembling the sugarcane genome has been a challenge, and only very recently polyploid assemblies have been generated. To better understand and represent the genomic diversity in sugarcane, we are constructing a pan-genome graph, a way to study the genomic variation present in the *Saccharum* complex and enable a better understanding of the genomic makeup of the hybrids.

METHODS

We are using the Cactus-Minigraph pipeline, which employs Minigraph for the construction of the SV-only graph, Giraffe for mapping the reads on the graph, Cactus to construct the actual graph that contains variants of all sizes, vg to filter the variants, and ODGI to create a 1D visualization of each chromosome. We start with the genomes of the two parental species *S. officinarum* cultivar LA-Purple and *S. spontaneum* cultivar Np-X. Then we add the genomes of two comercial hybrids (Fig. 1).

Genome/Variety	A (%)	B (%)	C (%)	D (Gbp)	E (#)	F (Kbp)
LA-Purple SOFFI	99.82	1.21	98.61	6.8	9617	82862
Np-X SSPON	99.86	1.31	98.55	2.7	1033	68640
CC-01-1940	97.55	64.05	33.50	0.9	35089	34980
R570 JGI	99.84	0.47	99.37	5.0	143	79221

Table 1. Statistics of *Saccharum* complex genomes available in this project, including metrics of gene content and contiguity. Gene content metrics were calculated using COMPLEASM with the Poales order database, which has 4896 conserved single-copy genes. A) Complete genes, B) Complete genes in single copy, C) Duplicated complete genes, D) Assembly size, E) Number of contigs, F) N50

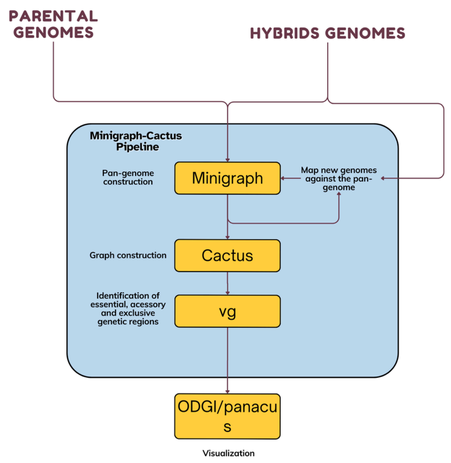

Figure 1. Illustrative figure of the proposed work plan in this project.

RESULTS

Our current pan-genome graph for sugarcane includes so far 4 genomes, representing 14.744.06.390 bp of Genetic information. The graph has 59.644.996 nodes and 80.593.237 edges, that represent 34.940 structural variants.
Our graph is stored in Graphical Fragment Assembly (GFA) format, in which each segment is a piece of contiguous sequence and also a node in the graph. Figure 2 shows the number of segments present as a function of the number of genomes they appear in.

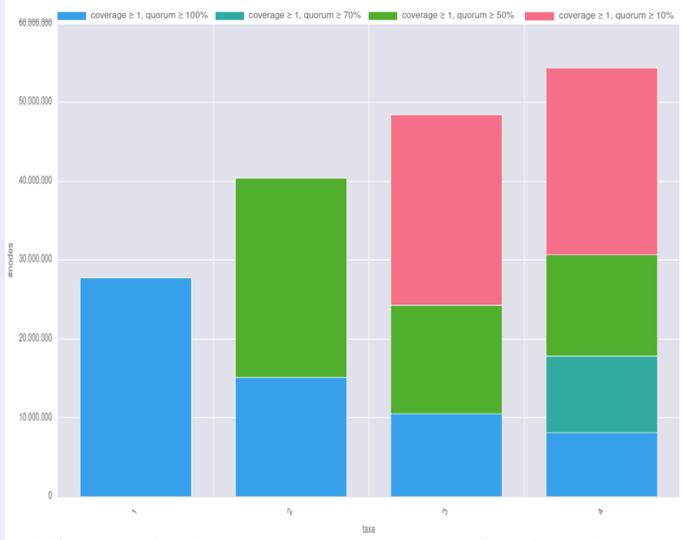

Figure 2. Pan-genome growth plot showing the number of nodes that are core (blue), or dispensable (any other color).

In the figure, we see a highlighted block (red circle) where the Brul locus is located. This locus confers durable resistance against brown rust in sugarcane. It is evident that this is a single-copy locus, present only in the haplotype 3D of R570. Additionally, it appears to be absent in all other haplotypes analyzed here.

It can be used assembled genomes by mapping them and can be more precise than using a reference genome.
We are able to identify specific regions of interest, and see in which haplotype or cultivar they are presento shown in Figure 3:

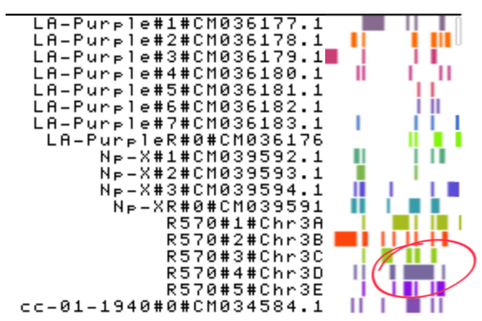

Figure 3. The region in the pangenome that carries the *Bru1* locus, note that it is largely absent in all haplotypes except in R570 3D

NEXT STEPS

We will continue adding further sugarcane genome to our pan-genome as well as two genomes of the genus *Miscanthus* and *Erianthus*, to help us better understand the variation within the *Saccharum* complex. One of the cultivars soon to be added is the Brazilian SP80-3280 sequenced in our laboratory, with very high contiguity.