Inference and Annotation of the Sugarcane Pan-Transcriptome

Felipe Vaz Peres, Diego Mauricio Riaño-Pachón, Jorge Mario Muñoz-Pérez

Laboratory of Computational, Evolutionary and Systems Biology, Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, SP, Brazil

Abstract

Sugarcane breeding programs are increasingly exploring molecular approaches for the development of new varieties. However, due to the sugarcane genome complexity, there is not a single version of the genome that represents the multiple allelic copies (i.e., ploidy >10) and achieves chromosome scale, difficulting the identification of genes responsible for phenotypes of interest.

An alternative to identify target genes responsible for the phenotypes of interest is the large-scale sequencing of mRNA. Many of the RNA-Seq datasets generated so far are available in public repositories and can be used to generate new insights. Here we show the inference of the sugarcane Pan-transcriptome generated from 48 automatically assembled sugarcane genotypes.

Transcriptome Assembly

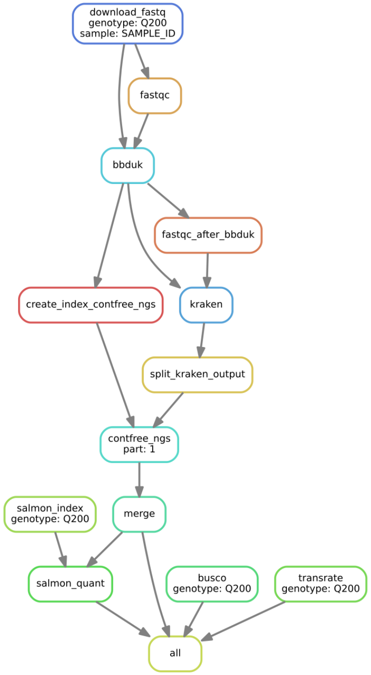

Figure 1. Directed Acyclic Graph (DAG) showing the steps of our automated transcriptome assembly pipeline.

Our pipeline is available at: github.com/labbces/SCPT

Sugarcane Pan-Transcriptome Inference

Number of genotypes	48
Number of total transcripts	16,237,098
Number of transcripts with CDS	**5,240,794**
Percentage of transcripts with CDS in orthogroups	96.9
Total groups	153,841
Core groups	12,738
Genotype-specific groups	653

Figure 2. Pan-Transcriptome inference generated by OrthoFinder2. For each transcriptome, transcripts containing CDS were extracted and these CDS were translated into proteins. Then we ran OrthoFinder2 to generate groups of related proteins from the 48 sugarcane transcriptomes.

Sugarcane Pan- and Core-Transcriptome

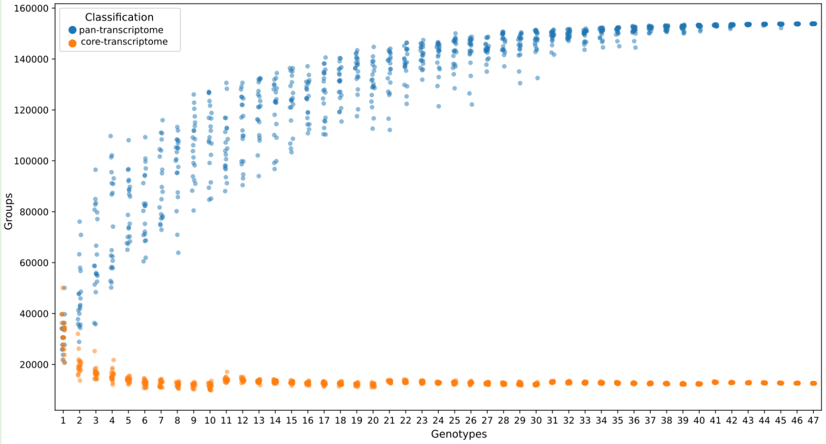

Figure 3. Pan- and Core-Transcriptome. The total number of transcript groups increased as additional transcriptomes were added and started to reach a plateau when n >= 24 genotypes were included (143,290 groups and 5,077,629 transcripts). Similarly, the core-transcriptome size also reaches a plateau, even faster than the pan-transcriptome, when n >= 11 genotypes (13,978 groups and 2,853,218 transcripts).

Genotype-specific Assembly Quality

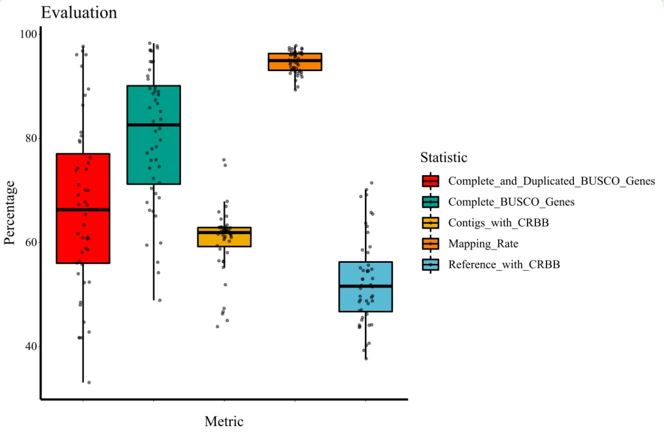

Figure 4. BUSCO Genes, Transrate (Contigs with CRBB and Reference with CRBB) and Salmon (Mapping Rate) metrics generated for each genotype-specific transcriptome assembled by our automated pipeline.

Discussion

- We assembled 48 sugarcane genotype-specific transcriptomes that contains 16,237,098 assembled transcripts (5,240,794 of these have CDS). Clustering based on sequence similarity classified all transcripts with CDS into 153,841 groups.
- The sugarcane pan-transcriptome generated in this project has some interesting features. One of our doubts was whether, with the amount of transcripts and genotypes available, it would be possible to represent the set of all sugarcane transcripts, this would be observed as a plateau of the curve of the number of groups per genotype (pan-size transcriptome), which is what we actually observed (blue curve) in Figure 3, where around 24 genotypes it is possible to recover most groups of transcripts (average of 143,290 groups). Likewise, we can see in Figure 3 that the size of the central transcriptome (defined here as those groups that have transcripts from at least 90% of the genotypes) also stabilizes, even faster than for the pan-transcriptome, at about 11 genotypes (average of 13,978 groups).

Financial Support

This research has been supported by São Paulo Research Foundation (FAPESP), project number: 2019/24796-5 to FVP, by CAPES, project number: 88887.597556/2021-00 to JMMP, and by National Council for Scientific Technological Development (CNPq), project number: 310080/2018-5 and 311558/2021-6 to DMRP;