XVIII BRAZILIAN CONGRESS OF PLANT PHYSIOLOGY
I Ibero-latinamerican Congress of Plant Biology


DMRP

PLANT SYSTEMS BIOLOGY: INTEGRATING SCALES AND TIMES

PLANT SYSTEMS BIOLOGY: INTEGRATING SCALES AND TIMES

Sugarcane genomics and transcriptomics resources

Felipe Vaz Peres[1], Verusca Semmler Rossi[1], Jorge Mario Muñoz Perez[1], **Diego Mauricio Riaño-Pachón**[1,2,3]

1. Center for Nuclear Energy in Agriculture, University of São Paulo, 2. National Institute for Science and Technology of Bioethanol,
3. Research Center for Greenhoouse Innovation

diego.riano@cena.usp.br

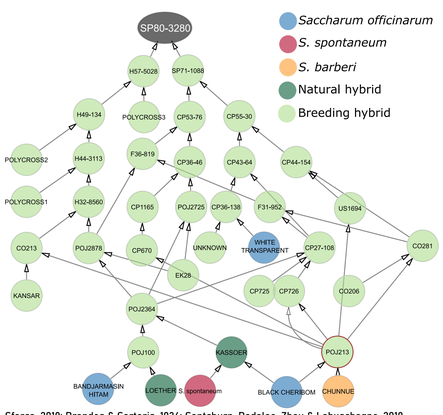
BACKGROUND

Complex Genome


Grivet & Arruda, 2002

The genome of modern sugarcane cultivars is a mosaic, with chromosomes originating from **S. officinarum** (yellow, 70%-90%), and **S. spontaneum** (green, 10%-20%), some recombinant chromosomes (5%-10%) and genetic information from other species within the **Saccharum** complex. Genetic information from **S. spontaneum** brings traits related to biotic and abiotic stress resistance, and **S. officinarum**'s brings sugar content related traits. The monoploid genome size of the modern hybrids is around 1Gbp, with ploidy levels 10-12, highly repetitive, aneuploid and polymorphic.

Multi-species, polyploid breeding



Sforca, 2019; Brandes & Sartoris, 1936; Santchurn, Badaloo, Zhou & Labuschagne, 2019

Pedigree of cultivar SP80-3280. With contributions of three different species within the *Saccharum* complex.

Available genome assemblies

Assembly	Contigs	N50 (Kbp)	Size (Mbp)	Genes	Technologies	Publication
SP80-3280-CTBE	199,028	8.4	1,167	153,078	TruSeq Synthetic Long Reads	Riaño-Pachón & Mattiello, 2017
SP80-3280-IQ/USP	450,608	13.2	4,259	374,774	TruSeq Synthetic Long Reads	Souza, et al., 2019
R570-MTP-CIRAD	5,708	116.7	530	41,223	BAC, PacBio RSII	Garsmeur, et al., 2018
R570-STP-CIRAD	211	45,576.6	427	25,316	BAC, PacBio RSII	Garsmeur, et al., 2018
CC-01-1940-CENICANA	35,089	34,980.0	904	68,260	PacBio RSII, HiC, genetic map	Trujillo-Montenegro, et al., 2021

There are 5 genome assemblies for three sugarcane cultivars, using modern DNA sequencing technologies. The SP80-3280 assemblies are an attempt to represent the poliploid genome, but they are still highly fragmented. The assemblies for the other two cultivars are a reconstruction of a mosaic haploid genome. **How do they compare to each other?**

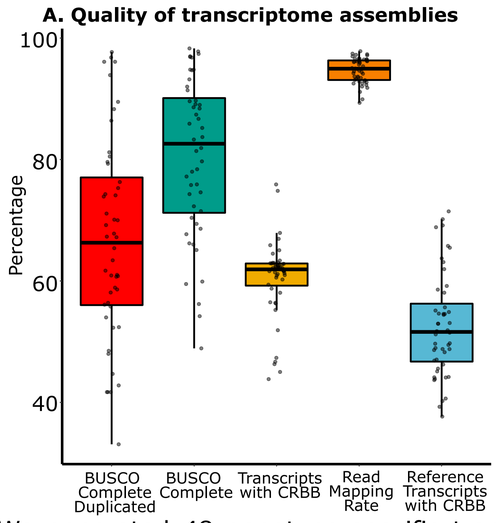
RESULTS AND DISCUSSION

Do we have a complete sugarcane genome yet?

We clustered all the protein coding genes of the 5 genome assemblies using OrthoFinder. The number below the assembly name is the total number of clusters. Although most of the clusters are shared with all assemblies, each assembly has exclusive clusters, this is a consequence of the high genetic variability of the crop, and highlights the importance of pan-genomics approaches.

Pan-transcriptome: 48 genotype-specific transcriptomes

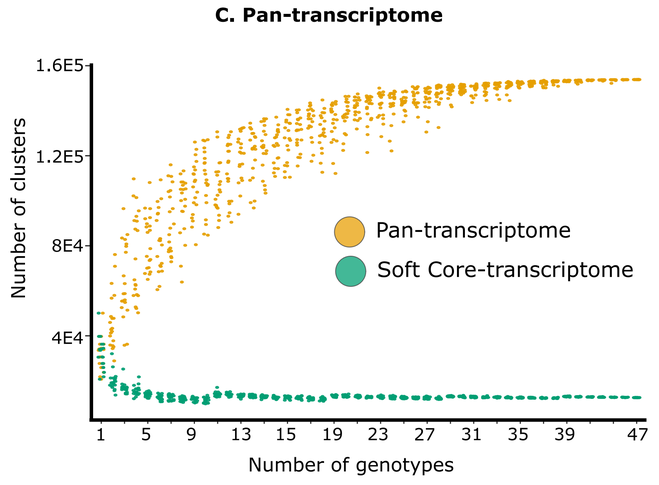
A. Quality of transcriptome assemblies



B. Pan-transcritome statistics

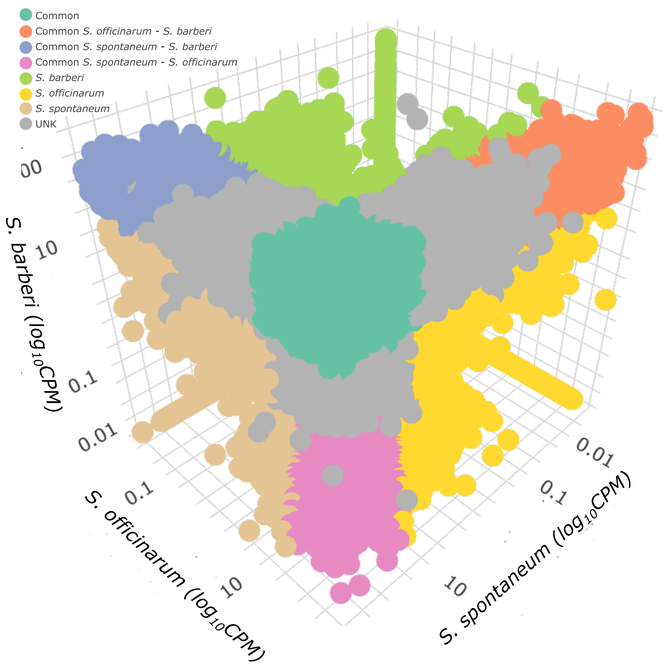
Metric	Value
Number of genotypes	48
Sum of the number of transcripts	16,237,098
Protein coding transcripts	5,240,794
Transcripts in clusters	5,077,629
Number of clusters	153,841
Genotype-specific clusters	653
Number of transcripts in genotype-specific clusters	1,578
Average number of transcripts per cluster	33
Median cluster size	6
Number of cluster with all genotypes	8,142
Number of cluster with 90% genotypes	12,738

C. Pan-transcriptome



We generated 48 genotype-specific transcriptome assemblies exploiting public data, using Trinity with kmer={25,31}. **(A)** The resulting assembly metrics are very good for most assemblies. The average BUSCO was over 80%, and over 60% of the sugarcane transcripts had homologues in *Sorghum bicolor*. **(B)** There are over 5.2E6 protein-coding transcripts, which is 32% of all assembled transcripts. Most protein-coding transcripts can be assigned into one of 153,841 clusters. This clustering was obtained with OrthoFinder with inflation 1.5 and similarity search with default values, more stringent values (inflation ={1,5, 2.0, 4.0, 6.0} and sensitive similarity searches, increase the number of clusters. The hard core-transcritpome is constituted of 8,142 clusters and the soft core-transcriptome of 12,738 clusters. **(C)** Most of the clusters can be recovered already with only 24 genotypes (pan-transcriptome), while the soft-core (groups with 90% of the genotypes) is recovered with 11 genotypes already.

Species of origin in SP80-3280



As shown above in the pedigree for the cultivar SP80-3280 at least three *Saccharum* species have contributed to its genetic background, i.e., *S. officinarum*, *S. spontaneum* and *S. barberi*. Using the transcriptome assembly for this cultivar we tried to assing the probable species of origin for each transcript, exploiting publicly available genomics datasets for the three species, and computing "Counts per Million" (CPM), for each transcript and species. For this analyses we have excluded transcripts with CPM in the top 1%, as they likely represent repetitive regions, from the remainder we only kept these that have a CPM of at least 1 for at least one of the species. At least 81.8% of the transcripts appear to be present in the three species (Common and UNK). A surprisingly small fraction of transcripts can unambiguously be identified as originating from only one of the three species (2.9% from *S. barberi*, 2.1% from *S. officinarum*, and 4.7% from *S. spontaneum*). The remaining 8.5% are transcripts that are common between pairs of the three founding species.

DATA AVAILABILITY

Transcriptome assembllies, their annotation and cluster conformation are publicly available via https://figshare.com/projects/Sugarcane_Pan-transcriptome/130586

We have set up a BLAST server to query our transcriptome assemblies, temporarily available at: http://200.144.245.42:4567/

For more information use the QR code at the top and bottom of the poster.
Interested in working with this (and more of this), please get in contact!

REFERENCES

Brandes, E. W. ; Sartoris, G. B. 1936. Yearbook of the United States Department of Agriculture pp.561-624

Garsmeur, O.; Droc, G.; et al. 2018. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nature communications; 9(1):2638

Grivet L., and Arruda P. 2002. Sugarcane genomics: depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology 5(2):122-127

Riaño-Pachón, D.M.; and Mattiello, L. 2017. Draft genome sequencing of the sugarcane hybrid SP80-3280. F1000Res; 6:861

Santchurn, D.; Badaloo, M.G.H.; Zhou, M and Labuschagne, M.T. 2019. Contribution of sugarcane crop wild relatives in the creation of improved varieties in Mauritius. Plant Genetic Resources; 1–13

Sforca, DA. Variacão genética em poliploides complexos: desvendando a dinámica alélica em cana-de-acúcar. 2019. PhD Thesis, UNICAMP.

Souza G.M.;Van Sluys M.A.; et al. 2019. Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. Gigascience 8(12):giz129

Trujillo-Montenegro, J.H.; Rodríguez-Cubillos, M.J.; et al. 2021. Unraveling the genome of a high yielding colombian sugarcane hybrid. Frontiers in Plant Sciences; 12:694859

FUNDING

FAPESP INCT Bioetanol 14/50884-5
FAPESP RCGI 2020/15230-5
FAPESP FVP IC 2019/24796-5
CNPq DMRP PQ 1D 311558/2021-6
CNPq DMRP PQ 2 310080/2018-5
CNPq FVP Master 130748/2022-6
CNPq VSR Master133214/2019-2
CAPES JMMP Finance Code 001