Inference and Annotation of the Sugarcane Pan-Transcriptome

Felipe Vaz Peres[1], Diego Mauricio Riaño-Pachón[1] , Jorge Mario Muñoz-Pérez[1]

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

SUGARCANE

(Saccharum spp.)

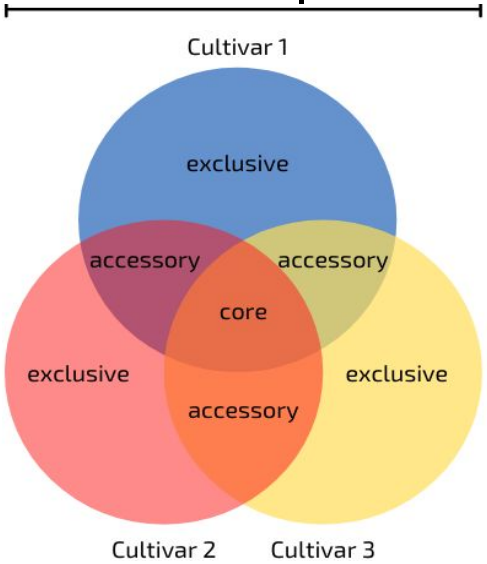
Agriculture - 2022/2023 harvest - 596.066 millions of tons[1]
Economy - 2% Brazilian GDP[2]

[1] - CONAB, 2022. Estimativa realizada em Abril de 2022 - safra 2022/2023.
[2] - Cepea, CNA e Fealq. **PIB DO AGRONEGÓCIO**, São Paulo, v.4, p. 1-17. out./2020

TRANSCRIPTOME

A transcriptome, by definition, is a complete set of transcripts from an organism, tissue, or cell lineage. Being the direct reflection of the expression of genes.

Pan-Transcriptome



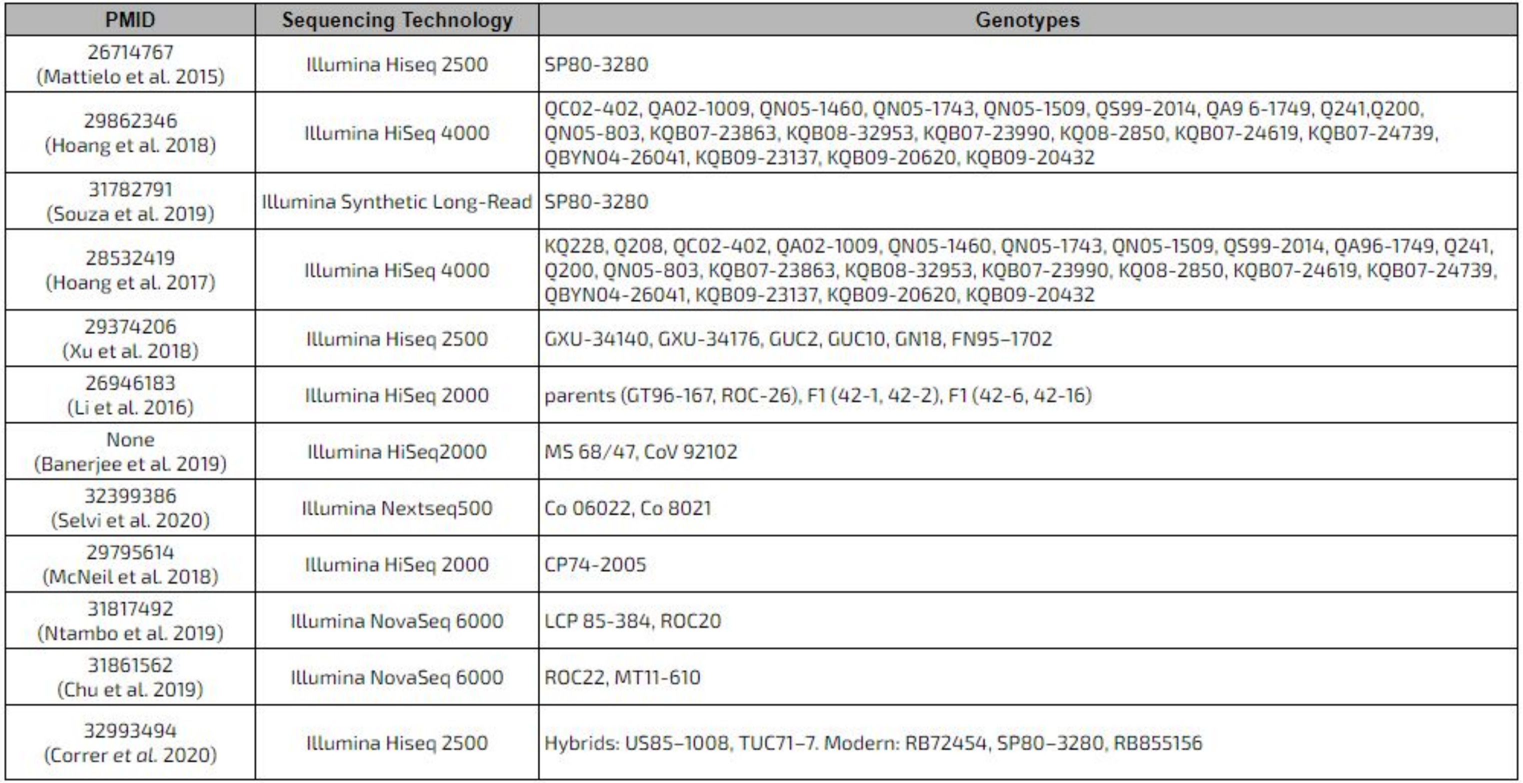
PUBLIC DATA

PAPERS

12

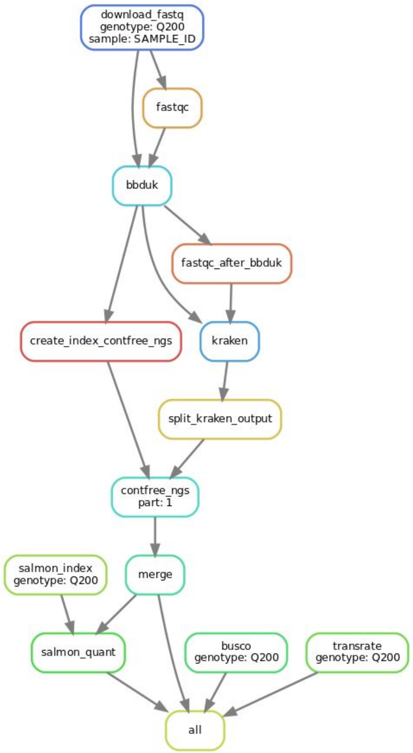
GENOTYPES

48



PMID	Sequencing Technology	Genotypes
26714767 (Mattielo et al. 2015)	Illumina Hiseq 2500	SP80-3280
29862346 (Hoang et al. 2018)	Illumina HiSeq 4000	QC02-402, QA02-1009, QN05-1460, QN05-1743, QN05-1509, QS99-2014, QA9 6-1749, Q241,Q200, QN05-803, KQB07-23863, KQB08-32953, KQB07-23990, KQ08-2850, KQB07-24619, KQB07-24739, QBYN04-26041, KQB09-23137, KQB09-20620, KQB09-20432
31782791 (Souza et al. 2019)	Illumina Synthetic Long-Read	SP80-3280
28532419 (Hoang et al. 2017)	Illumina HiSeq 4000	KQ228, Q208, QC02-402, QA02-1009, QN05-1460, QN05-1743, QN05-1509, QS99-2014, QA96-1749, Q241, Q200, QN05-803, KQB07-23863, KQB08-32953, KQB07-23990, KQ08-2850, KQB07-24619, KQB07-24739, QBYN04-26041, KQB09-23137, KQB09-20620, KQB09-20432
29374206 (Xu et al. 2018)	Illumina Hiseq 2500	GXU-34140, GXU-34176, GUC2, GUC10, GN18, FN95–1702
26946183 (Li et al. 2016)	Illumina HiSeq 2000	parents (GT96-167, ROC-26), F1 (42-1, 42-2), F1 (42-6, 42-16)
None (Banerjee et al. 2019)	Illumina HiSeq2000	MS 68/47, CoV 92102
32399386 (Selvi et al. 2020)	Illumina Nextseq500	Co 06022, Co 8021
29795614 (McNeil et al. 2018)	Illumina HiSeq 2000	CP74-2005
31817492 (Ntambo et al. 2019)	Illumina NovaSeq 6000	LCP 85-384, ROC20
31861562 (Chu et al. 2019)	Illumina NovaSeq 6000	ROC22, MT11-610
32993494 (Correr *et al.* 2020)	Illumina Hiseq 2500	Hybrids: US85–1008, TUC71–7. Modern: RB72454, SP80–3280, RB855156

TRANSCRIPTOME ASSEMBLY


download_fastq
genotype: Q200
sample: SAMPLE_ID

fastqc

bbduk

fastqc_after_bbduk

create_index_contfree_ngs

kraken

split_kraken_output

contfree_ngs
part: 1

salmon_index
genotype: Q200

merge

salmon_quant

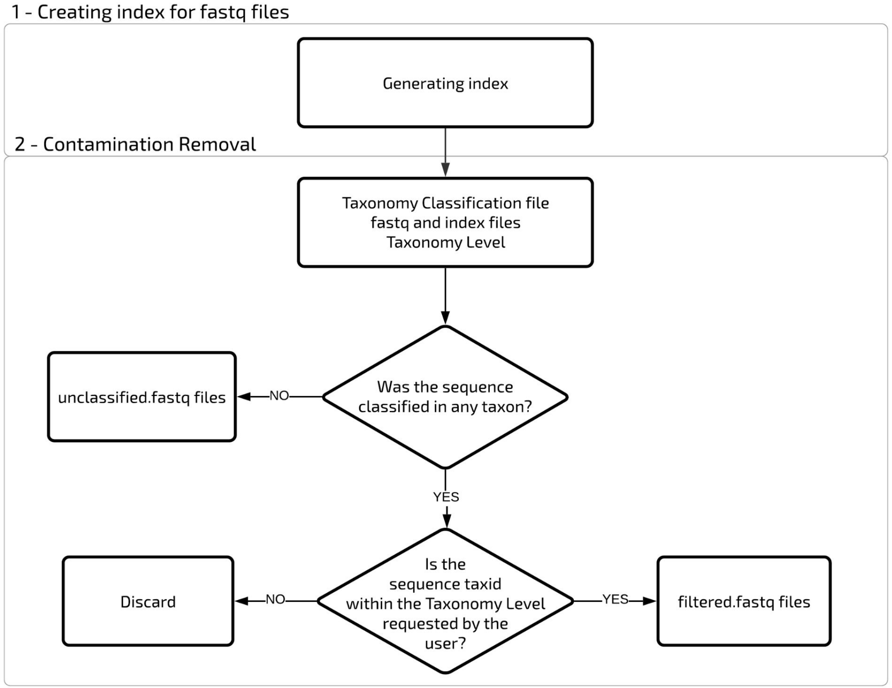
busco
genotype: Q200

transrate
genotype: Q200

all

DAG - Directed Acyclic Graph generated by Snakemake[5]

CONTAMINATION REMOVAL


1 - Creating index for fastq files

Generating index

2 - Contamination Removal

Taxonomy Classification file
fastq and index files
Taxonomy Level

Was the sequence classified in any taxon?

NO

unclassified.fastq files

YES

Is the sequence taxid within the Taxonomy Level requested by the user?

NO

Discard

YES

filtered.fastq files

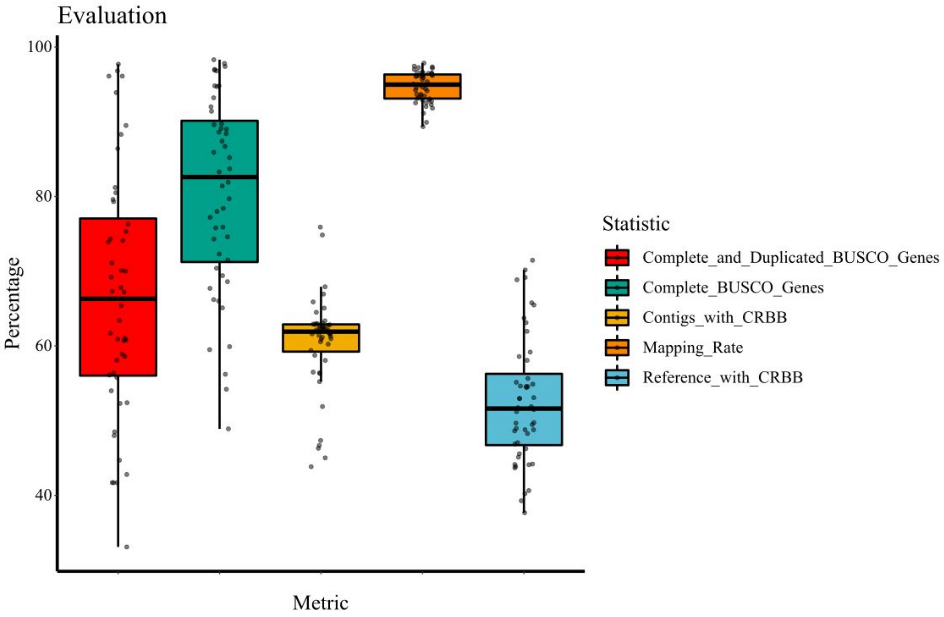
ContFree-NGS - Removing contaminants from reads

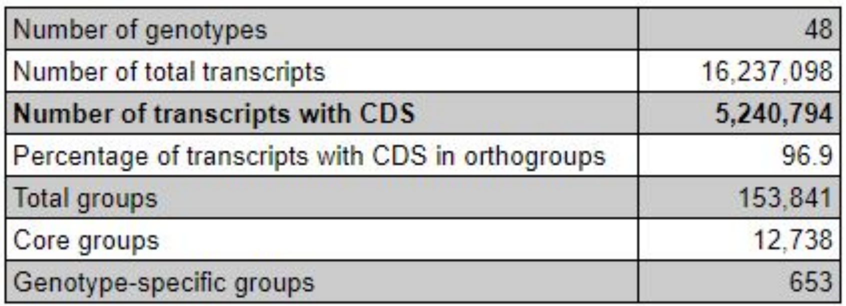
[4] - Our automated pipeline is available at: https://github.com/labbces/YAATAP

[5] - Köster, J., Rahmann, S. (2012) Snakemake - a scalable bioinformatics workflow engine, Bioinformatics, Volume 28, Issue 19, 1 October 2012, Pág 2520–2522 - https://doi.org/10.1093/bioinformatics/bts480

RESULTS

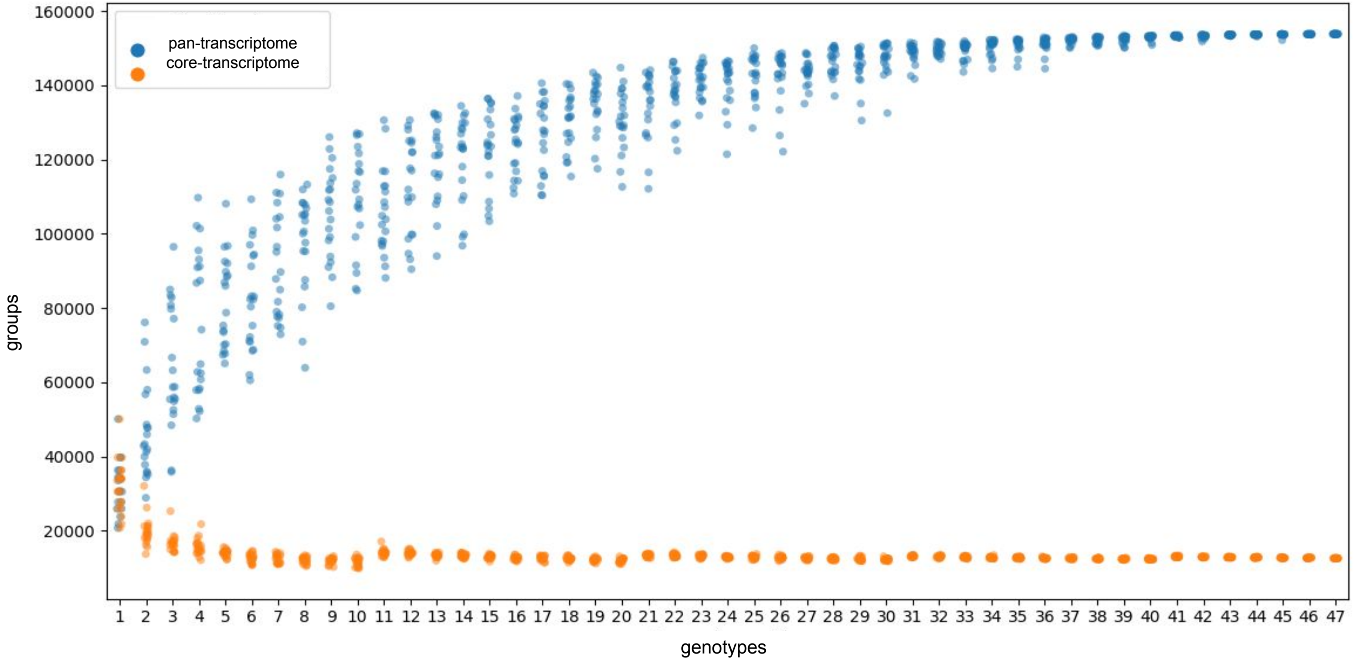
PAN-TRANSCRIPTOME INFERENCE



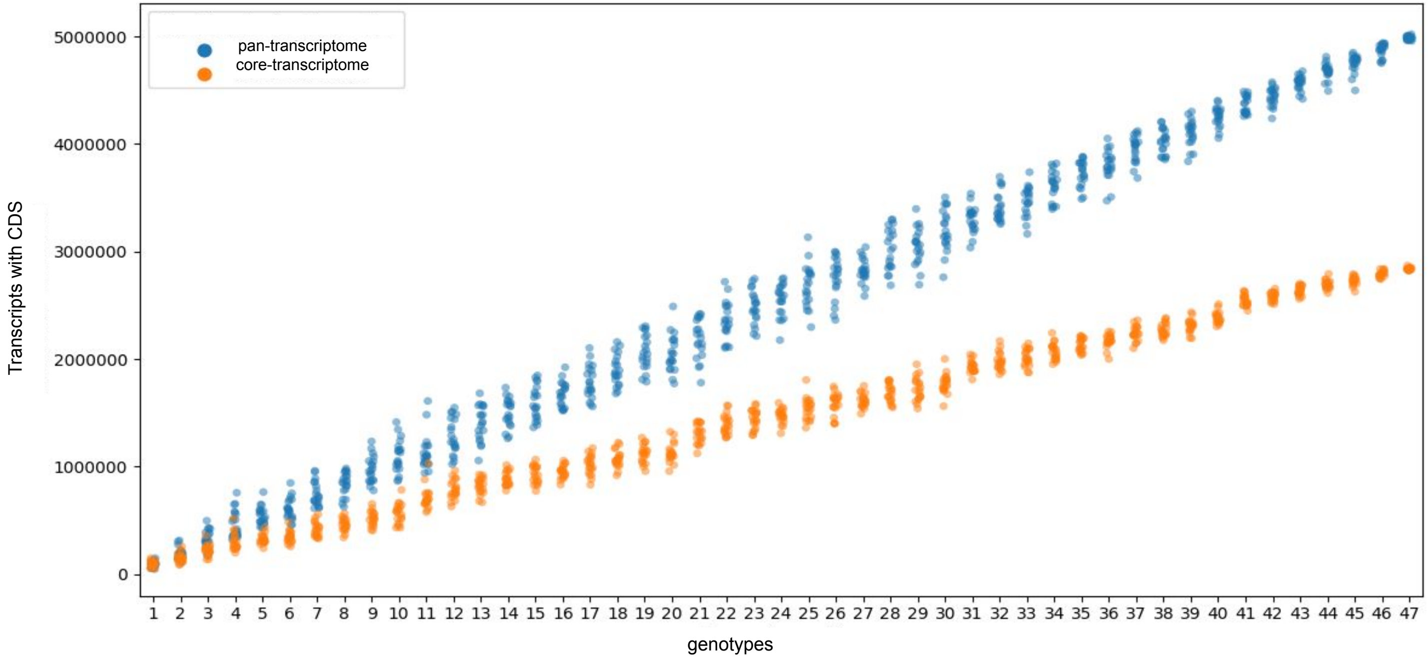


Number of genotypes	48
Number of total transcripts	16,237,098
Number of transcripts with CDS	**5,240,794**
Percentage of transcripts with CDS in orthogroups	96.9
Total groups	153,841
Core groups	12,738
Genotype-specific groups	653

SUGARCANE PAN-TRANSCRIPTOME

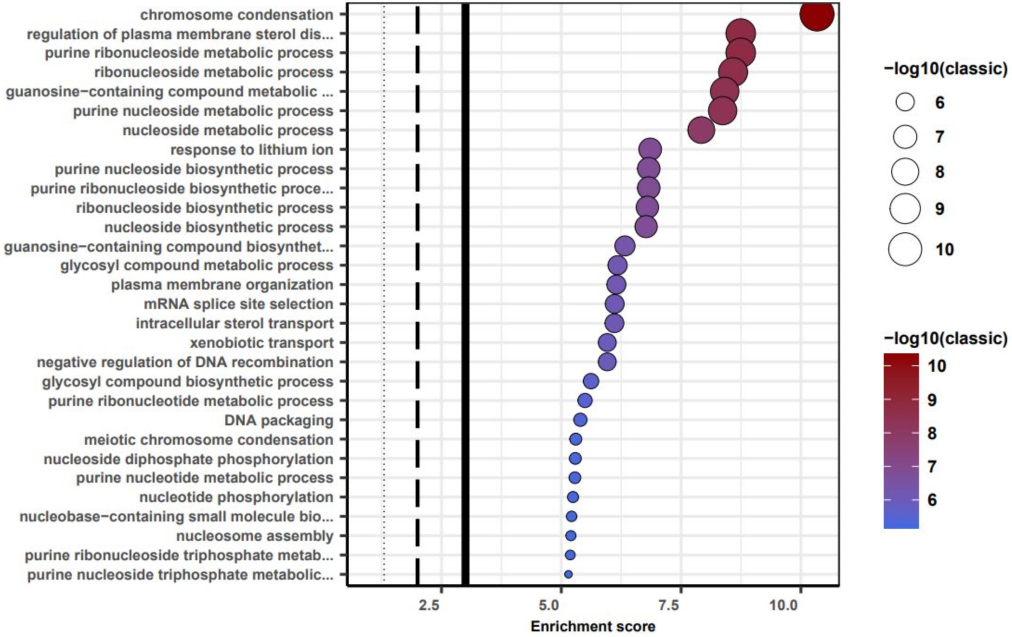


SUGARCANE PAN-TRANSCRIPTOME



SUGARCANE PAN-TRANSCRIPTOME

Analysis of enriched GO terms in exclusive groups



Cut−off lines drawn at equivalents of p=0.05, p=0.01, p=0.001

CONCLUSIONS

- ➢ 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.

- ➢ Total number of transcript groups increased as additional transcriptomes were added and approached 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).

- ➢ hard-core, soft-core, accessory, and exclusive groups are composed of 301,937; 817,355; 3,711,778; and 117,189 transcripts, respectively."

ACKNOWLEDGMENTS

FAPESP

CNPq

CAPES

Thanks!
Have any questions or suggestions?
Contact: **felipe.vzps@gmail.com**

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. DMRP is a level 2 CNPq research fellow.