

Sorghum Pan-transcriptome: A new tool for bioenergy crop improvement

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

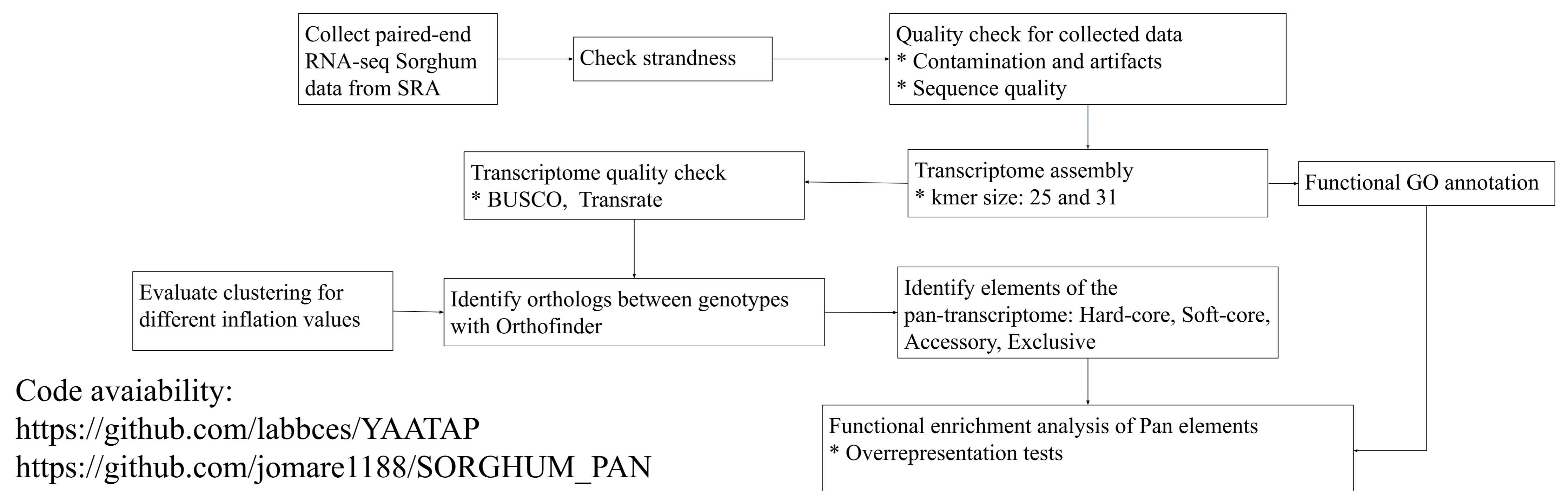
*Laboratório de Biologia Computacional, Evolutiva e de Sistemas, Divisão de Produtividade Agroindustrial e Alimentos, Centro de Energia Nuclear na Agricultura, Piracicaba/SP, Brasil.

⁺diego.riano@cena.usp.br

INTRODUCTION

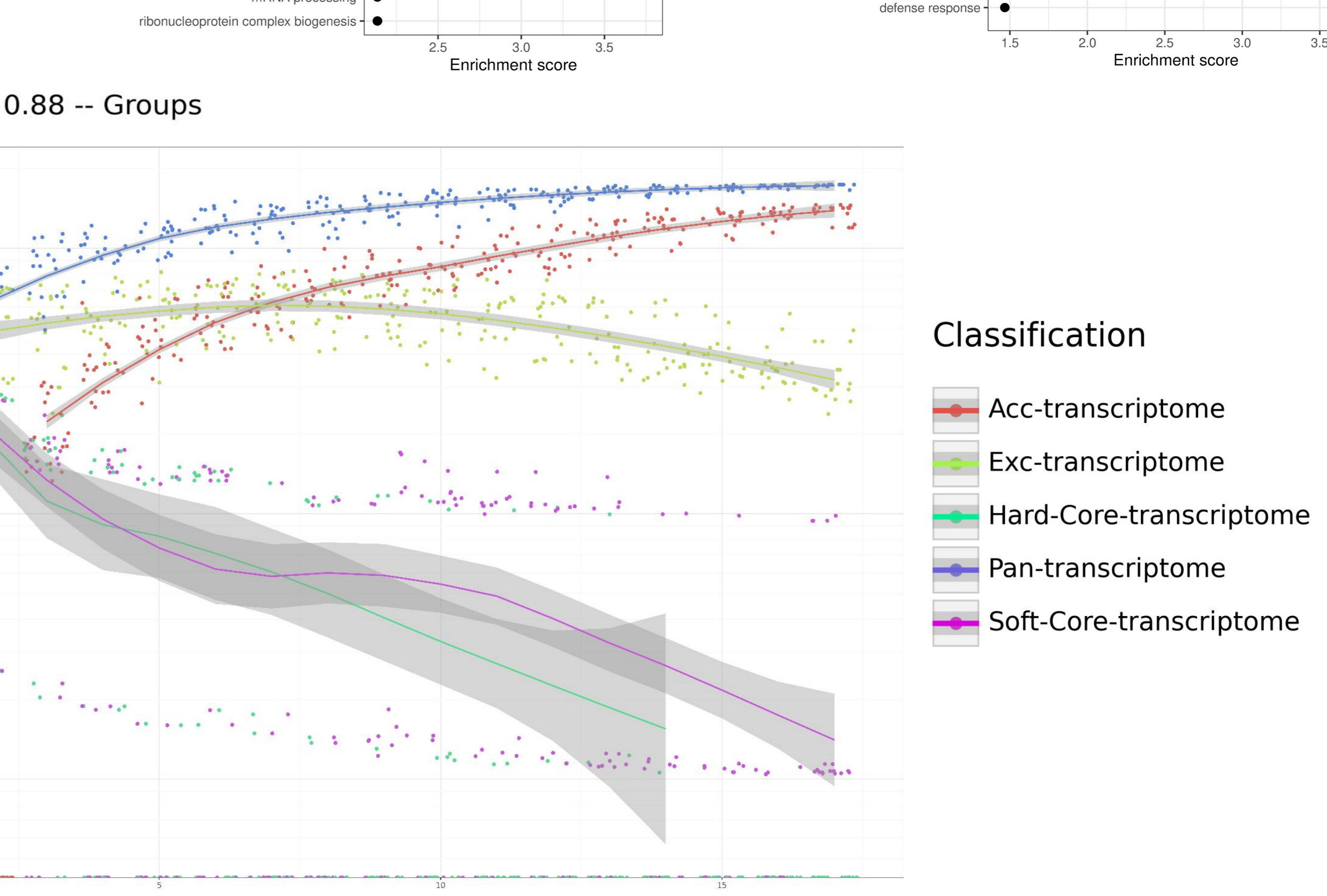
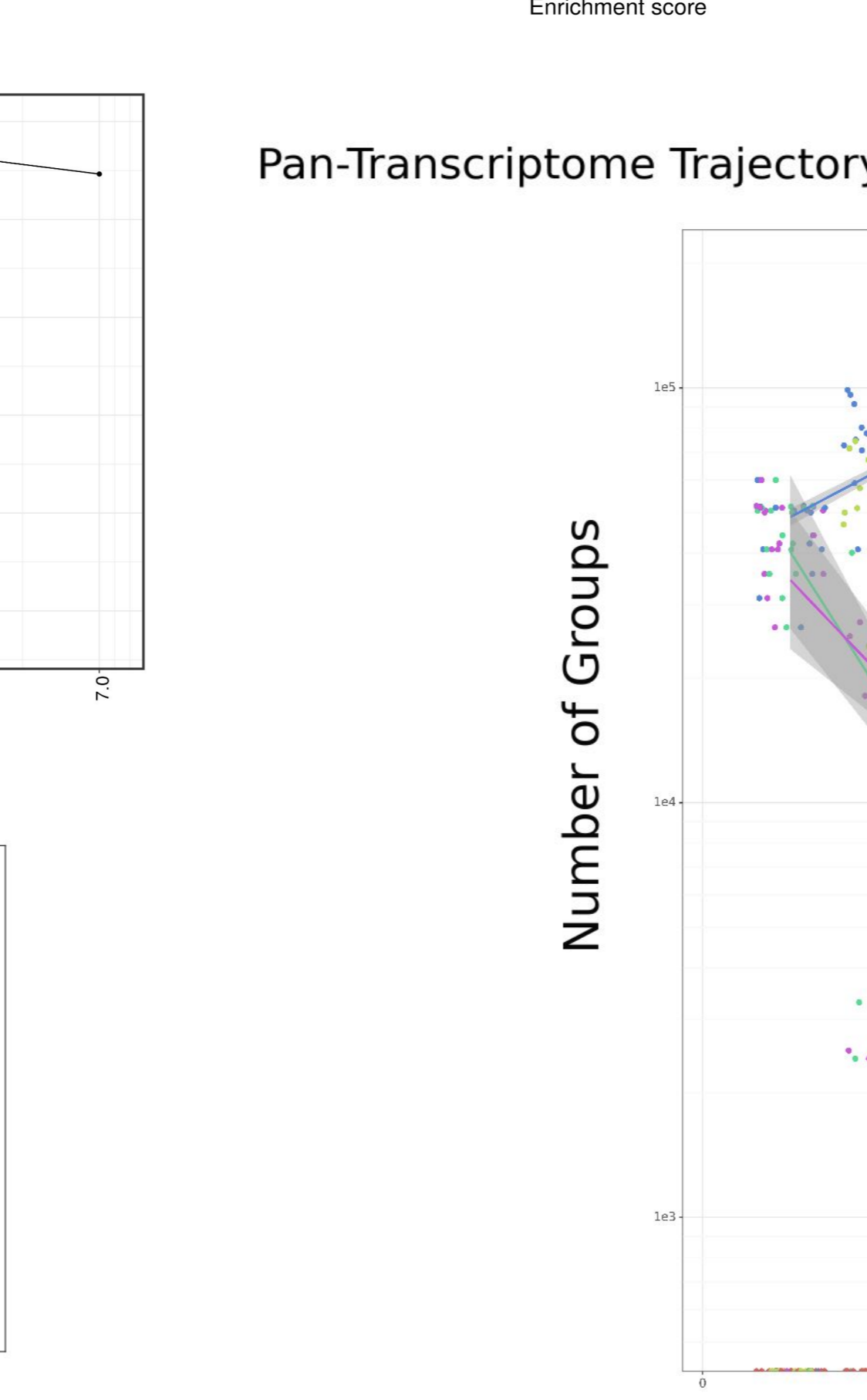
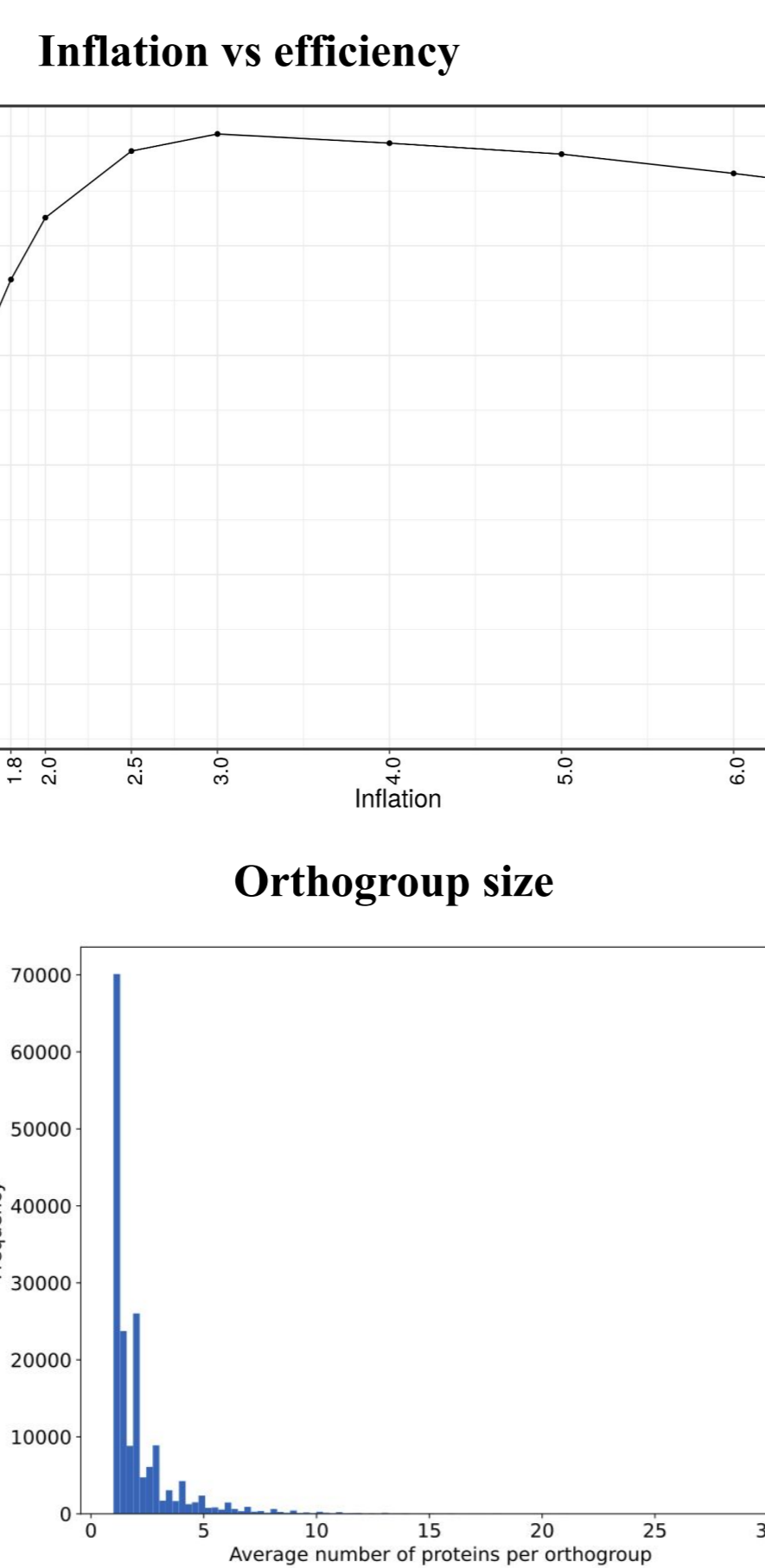
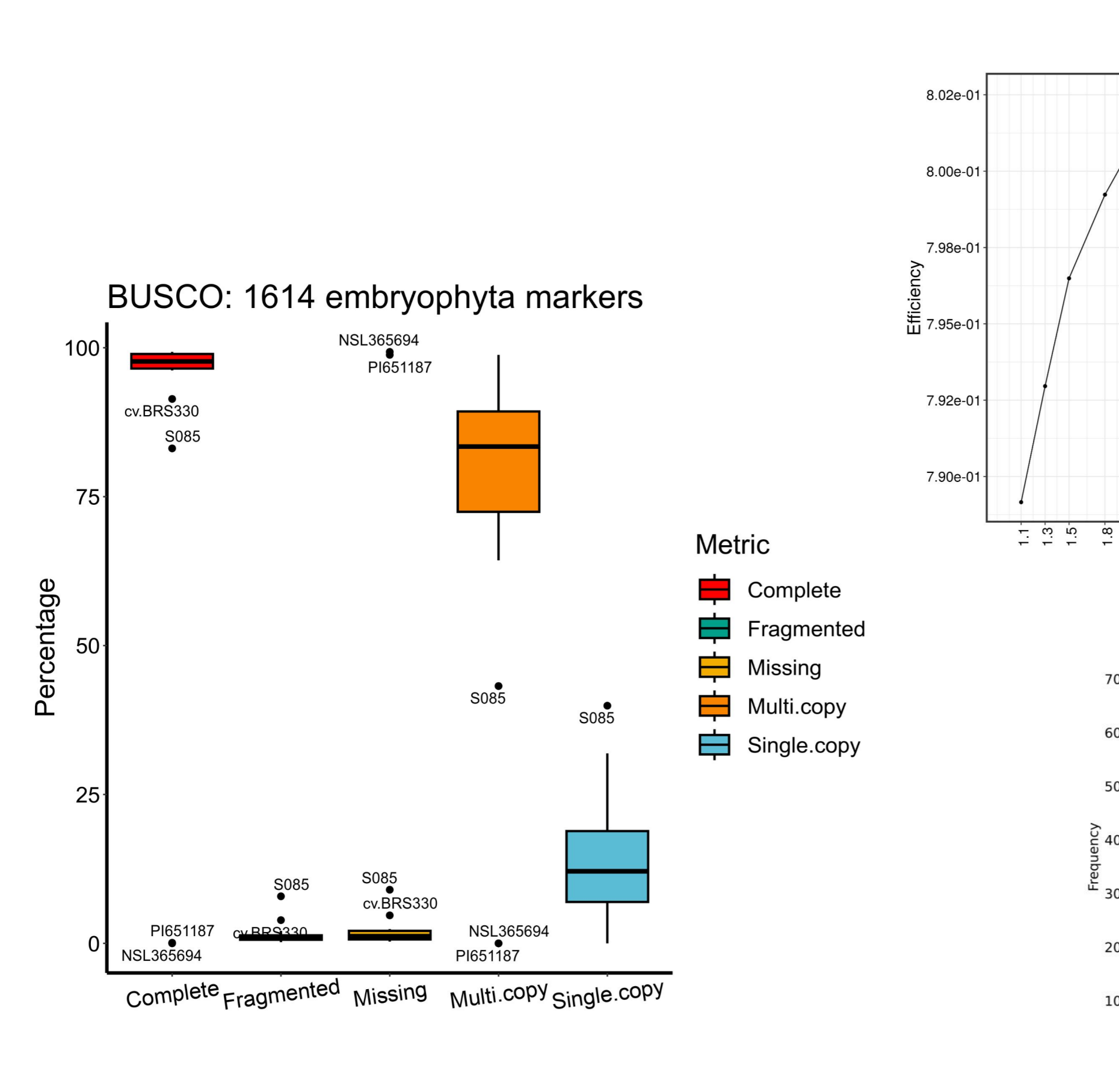
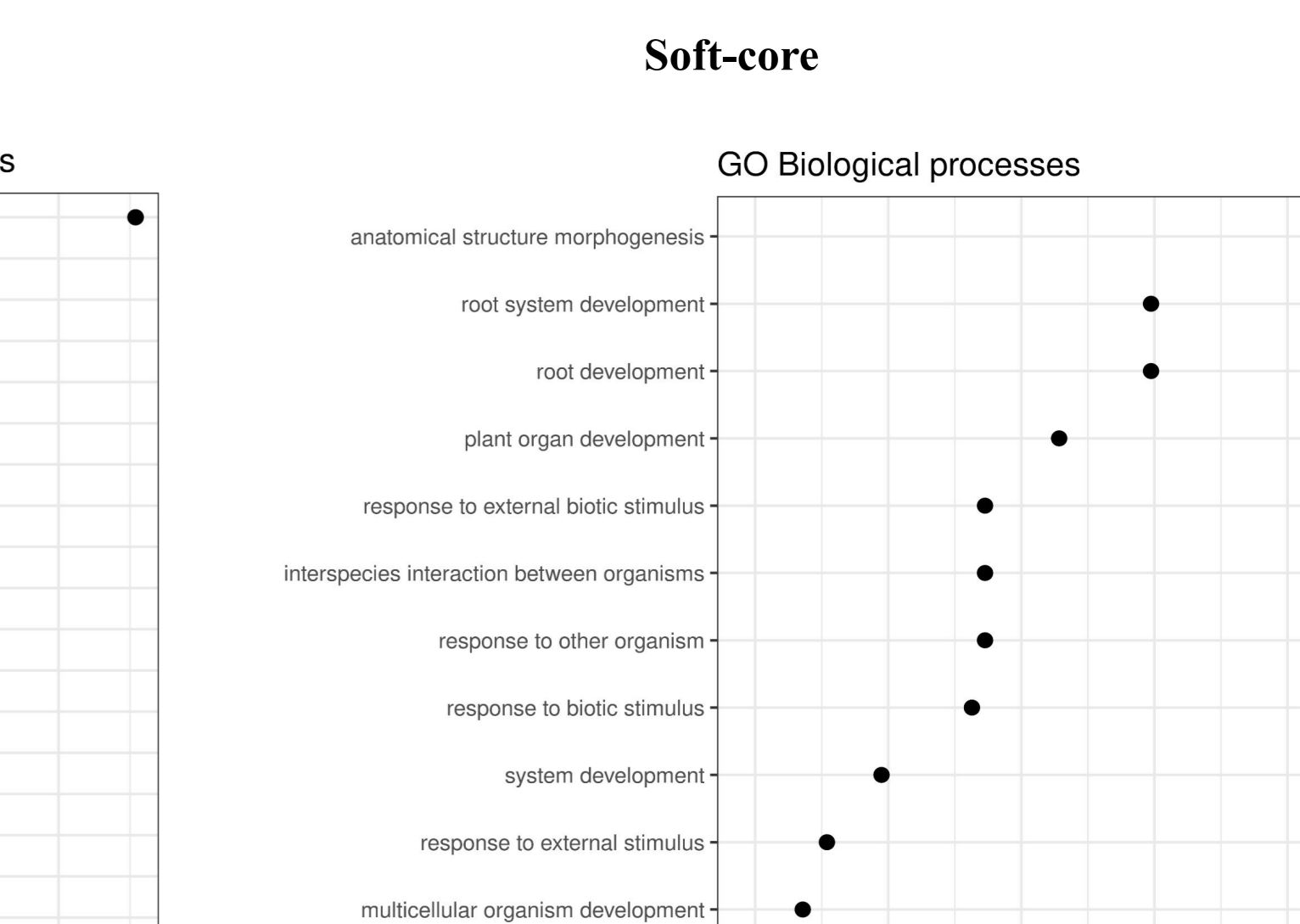
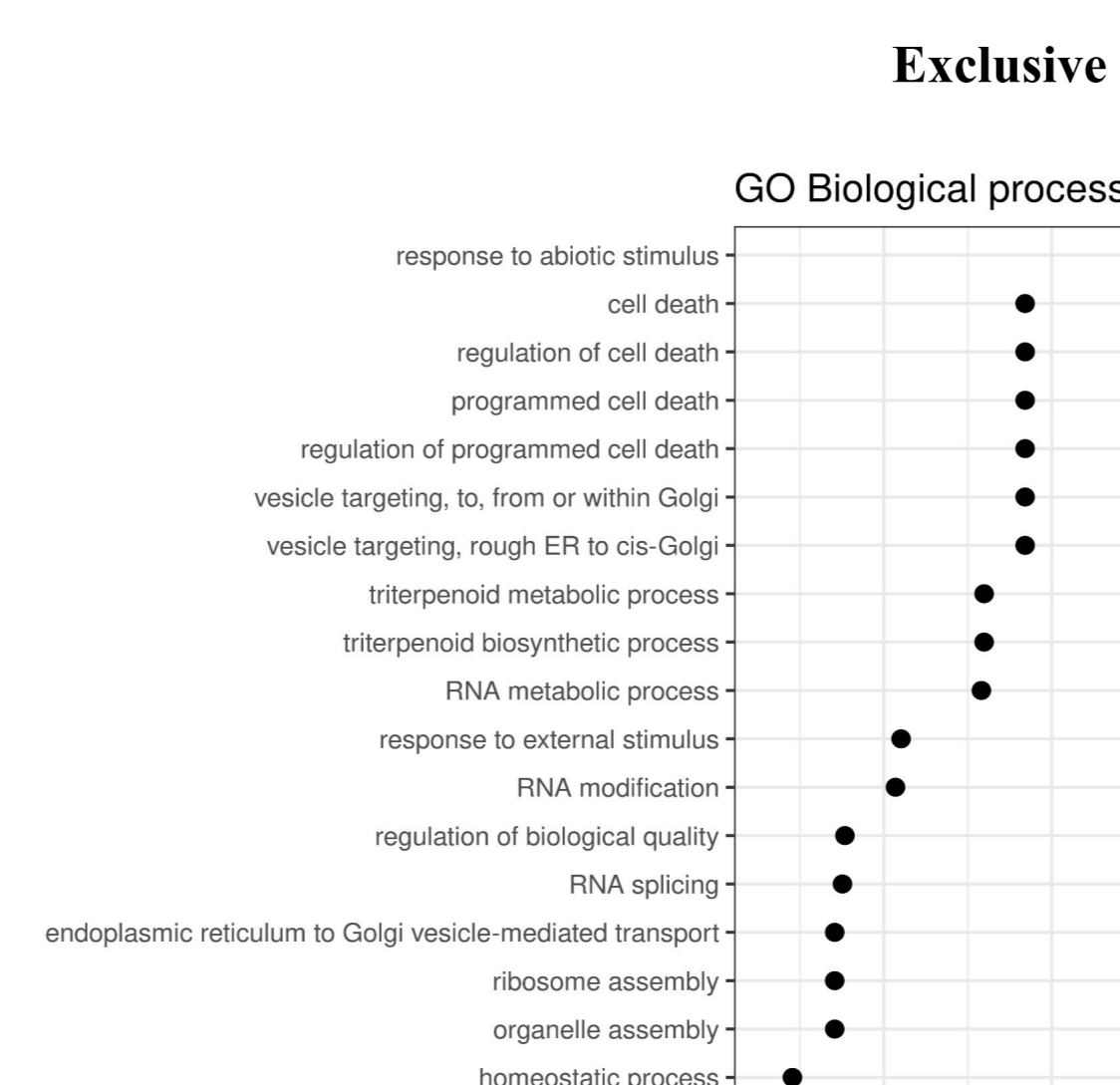
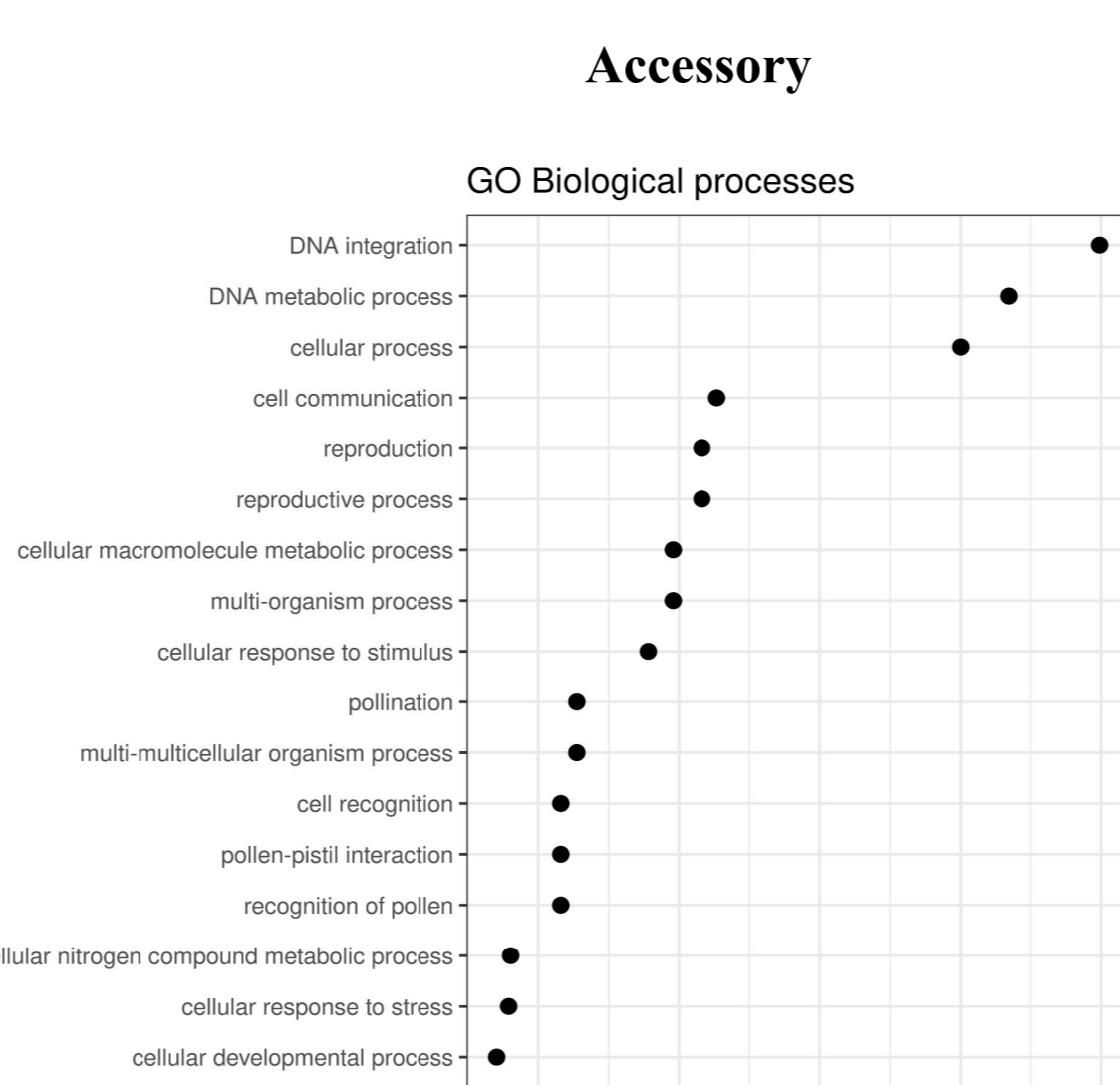
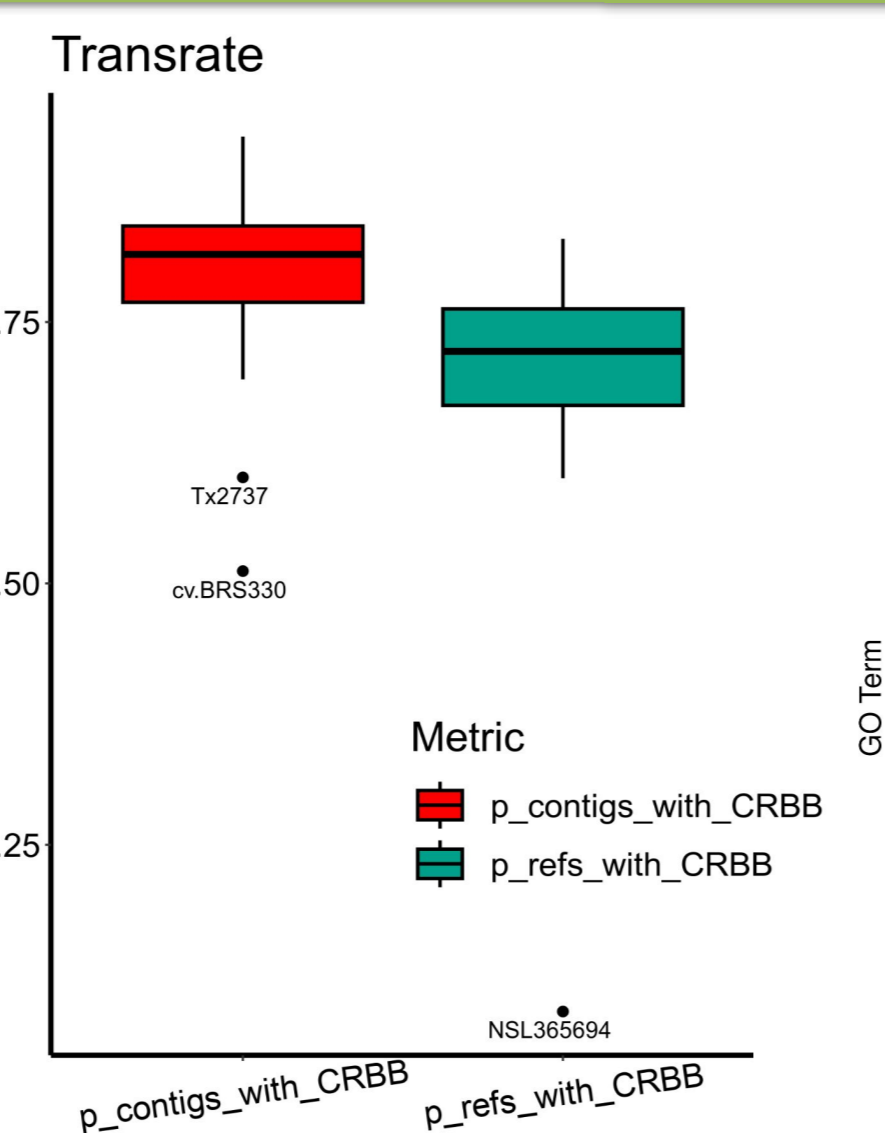
We collected and processed about $1,23 \times 10^{10}$ publicly available raw RNAseq reads from 18 different sorghum genotypes. Assembled transcriptomes sizes range from $1,51 \times 10^5$ to $4,32 \times 10^5$ transcripts. We have created, and we are making available, genotype-specific transcriptome assemblies (15 *de-novo* and three genome-guided), Gene Ontology annotation, and a general description of the Sorghum pan-transcriptome. We identified the accessory, exclusive, soft-core, and hardcore elements of the Sorghum pan-transcriptome and even more functional enrichment analyses revealed functions associated with specific elements of the Pan-transcriptome.

METHODS



RESULTS AND DISCUSSION

Genotype	Raw reads	nscqs	largest	mean len	proteins	proteins/n scqs	GOhits/proteins
BAZ9504	4.25E+08	2.39E+05	1.66E+04	1.74E+03	1.81E+05	0.76	4.32%
BTx623	1.21E+09	4.36E+05	4.37E+04	4.37E+03	2.49E+05	1.42	35.74%
cv-BRS330	2.17E+08	2.17E+05	1.64E+04	1.21E+03	1.17E+05	0.54	0.36%
Della	2.03E+09	2.80E+05	2.01E+04	2.18E+03	2.26E+05	0.81	16.50%
DKS-1707	7.73E+08	1.72E+05	1.67E+04	1.83E+03	1.25E+05	0.73	0.47%
DKS-4420	6.43E+08	1.66E+05	1.67E+04	1.79E+03	1.18E+05	0.71	15.85%
keller	1.39E+09	1.94E+05	1.92E+04	1.44E+03	1.26E+05	0.65	0.70
M35-1	9.91E+08	2.40E+05	4.54E+04	1.95E+03	1.67E+05	0.70	51.25%
Mota	3.78E+08	1.99E+05	2.32E+04	1.57E+03	1.42E+05	0.72	-
NSL365694	7.89E+06	4.91E+03	1.21E+03	3.90E+02	2.86E+02	0.06	-
PK61187	2.84E+07	1.54E+04	1.31E+03	4.25E+02	8.78E+02	0.06	-
R9188	6.17E+08	1.93E+05	1.66E+04	1.71E+03	1.46E+05	0.78	14.39%
Riceef	1.02E+09	2.62E+05	2.83E+04	2.26E+03	4.05E+05	1.55	43.45%
Rio	1.01E+09	2.70E+05	1.67E+04	1.68E+03	1.84E+05	0.68	-
RTx430	2.02E+09	3.28E+05	2.61E+04	2.80E+03	5.80E+05	1.78	16.18%
S085	8.31E+07	1.06E+05	1.47E+04	1.18E+03	1.19E+05	1.13	22.86%
SC187	9.28E+08	2.74E+05	1.73E+04	1.49E+03	1.66E+05	0.60	0.96%
TAM425	7.69E+08	1.87E+05	1.59E+04	1.29E+03	1.04E+05	0.55	15.12%
Tx2737	7.69E+08	2.51E+05	1.64E+04	1.54E+03	1.83E+05	0.73	2.65%
Tx3362	1.26E+09	2.20E+05	1.68E+04	1.51E+03	1.83E+05	0.83	0.35%
Tx378	1.88E+08	1.54E+05	1.61E+04	1.51E+03	1.27E+05	0.83	10.93%
Tx7000	7.90E+08	1.42E+05	1.84E+04	1.83E+03	1.07E+05	0.76	23.40%



REFERENCES

Emms, D.M. and Kelly, S. (2019) 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome Biology*, 20(1), p. 238. Available at: <https://doi.org/10.1186/s13059-019-1832-y>.

Henschel, R. et al. (2012) 'Trinity RNA-Seq assembler performance optimization', in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond*. New York, NY, USA: Association for Computing Machinery (XSEDE '12), pp. 1-8. Available at: <https://doi.org/10.1145/2335755.2335842>.

Manni, M. et al. (2021) 'BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes', *Molecular Biology and Evolution*, 38(10), pp. 4647-4654. Available at: <https://doi.org/10.1093/molbev/msab199>.

Petek, M. et al. (2020) 'Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato', *Scientific Data*, 7(1), p. 249. Available at: <https://doi.org/10.1038/s41597-020-00581-4>.

Smith-Unna, R. et al. (2016) 'TransRate: reference-free quality assessment of de novo transcriptome assemblies', *Genome research*, 26(8), pp. 1134-1144.

ACKNOWLEDGMENTS

- CAPES Programa de Excelência Acadêmica 88887.597556/2021-00, CNPq Bolsa de Produtividade em Pesquisa (310080/2018-5 e 311558/2021-6)
- Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP)
- CNPq 311558/2021-6
- FAPESP 20/15230-5