XIII Simpósio dos Pós-graduandos no CENA

Sustentabilidade na Agricultura e Meio Ambiente

Inferência e anotação do pan-transcriptoma da cana-de-açúcar

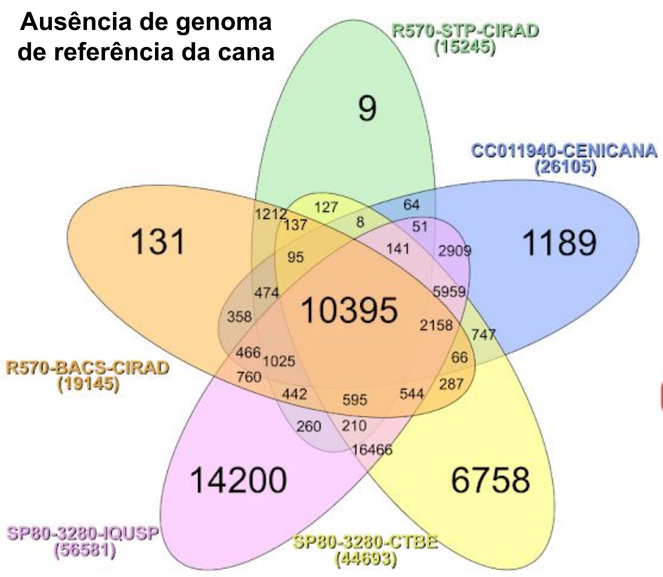
Felipe Vaz Peres[a], Jorge Mario Muñoz Pérez[a], **Diego Mauricio Riaño-Pachón**[a]*

[a]Laboratório de Biologia Computacional, Evolutiva e de Sistemas - Centro de Energia Nuclear na Agricultura - Universidade de São Paulo, Piracicaba, Brasil

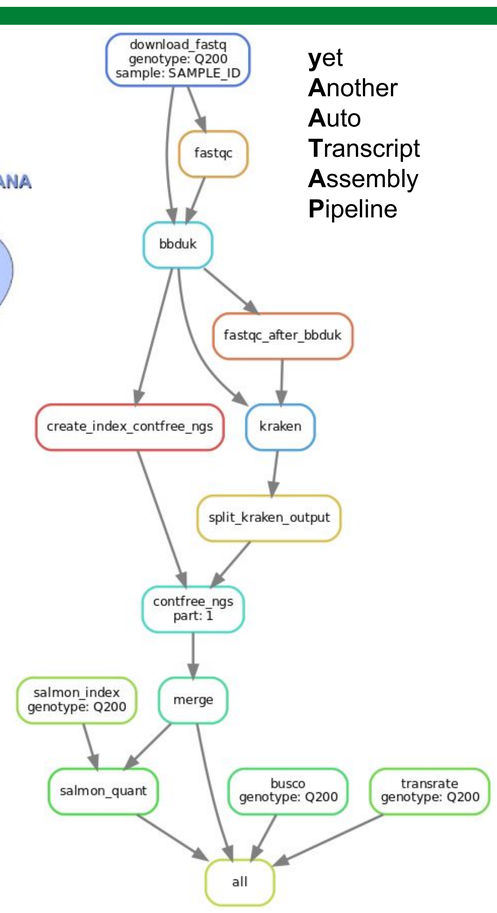
contato: diego.riano@cena.usp.br

INTRODUÇÃO

Os programas de melhoramento da cana-de-açúcar estão explorando cada vez mais abordagens moleculares para o desenvolvimento de novas variedades. No entanto, devido à complexidade de seu genoma, não existe uma versão única que represente as múltiplas cópias alélicas (ploidia >10) e alcance a escala cromossômica, dificultando a descoberta dos genes responsáveis por fenótipos de interesse. Na era genômica uma das descobertas mais interessantes é que nem todos os indivíduos de uma espécie compartilham a totalidade de informação genética. A união da informação genética de "todos" os indivíduos de uma espécie é denominada pan-genoma, e de maneira semelhante, esse conceito pode ser aplicado ao estudo de RNA-Seq, para avaliar transcritos interessantes. Neste trabalho realizamos a inferência e anotação do pan-transcriptoma da cana-de-açúcar a partir de dados brutos do RNA-Seq de 48 genótipos de híbridos comerciais da cana, com o objetivo de identificar quantos genes/famílias de genes estão presentes em todos os indivíduos, o tamanho do pan-transcriptoma (quantos genes/famílias de genes estão presentes dentro da espécie?) e quais genes estão presentes só em alguns indivíduos da espécie.



METODOLOGIA



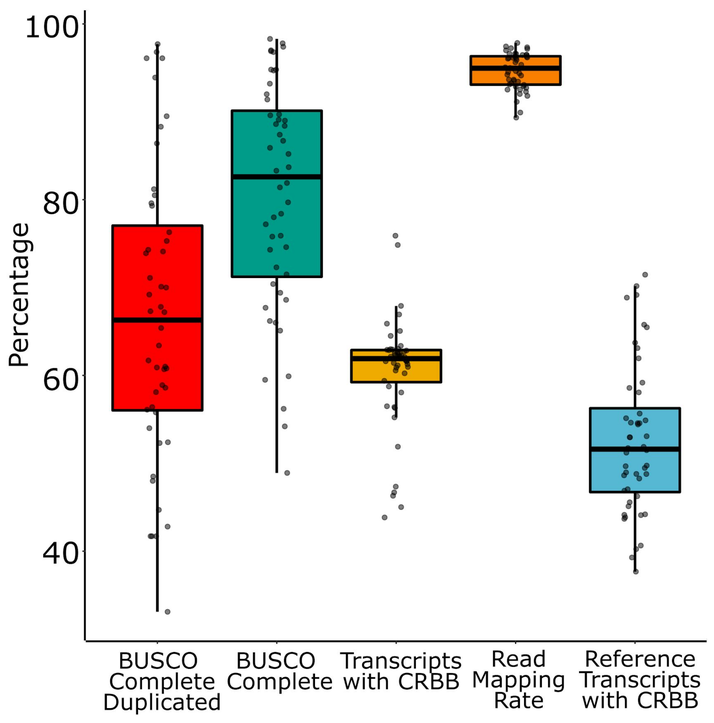
Muitos dados de RNA-Seq gerados até agora estão disponíveis em repositórios públicos e podem ser usados para gerar novos insights. Dados brutos de RNA-Seq de 48 genótipos da cana-de-açúcar foram baixados do NCBI e seus transcriptomas foram montados com o *YAATAP*, um pipeline automatizado para a montagem de transcriptomas desenvolvido pelo nosso grupo. O *YAATAP* é responsável por realizar o download dos dados brutos, controle de qualidade, remover sequências contaminantes e executar a montagem do transcriptoma, avaliando sua qualidade ao final do processo.

A anotação foi realizada individualmente para cada genótipo seguindo o protocolo Trinotate (Bryant *et al*., 2017), onde **5,240,794 proteínas totais** foram identificadas e utilizadas na inferência do pan-transcriptoma com o OrthoFinder2 (Emms and Kelly, 2019), através da criação de uma rede de similaridade entre as sequências e identificação de grupos com alta densidade de conexões utilizando a estratégia de Markov Chain Clustering (inflação = 1.5).

Fonte: github.com/labbces/YAATAP

RESULTADOS E DISCUSSÃO

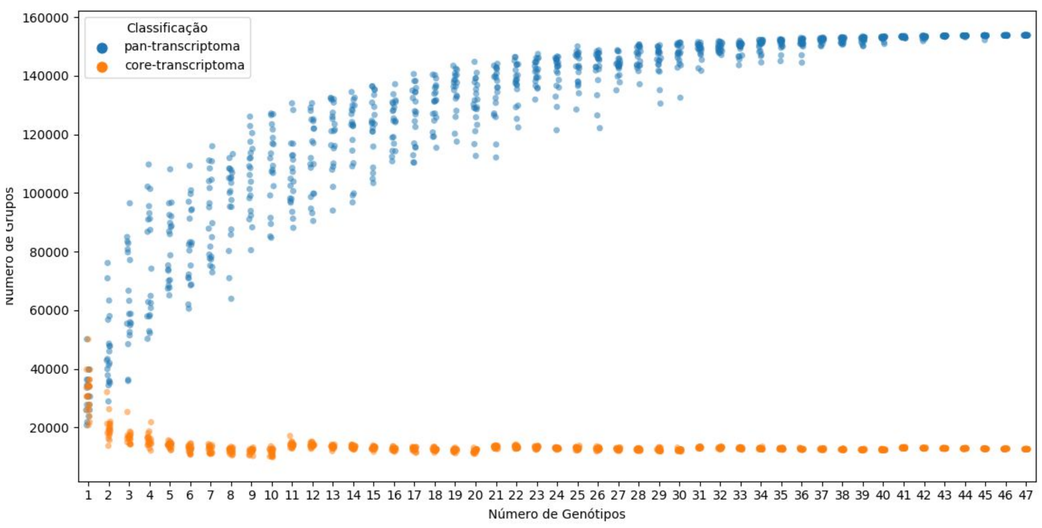
A. Qualidade das montagens de transcriptoma



B. Estatísticas do pan-transcriptoma

Métrica	Valor
Genótipos	48
Transcritos	16,237,098
Transcritos codificantes de proteínas	5,240,794
Transcritos em grupos	5,077,629
Grupos **(pan)**	153,841
Grupos com 90% dos genótipos **(soft-core)**	12,738
Grupos com todos genótipos **(hard-core)**	8,142
Grupos genótipo-específicos **(exclusive)**	653
Transcritos em grupos genótipo-específico	1,578
Média de transcritos por grupo	33
Mediana de transcritos por grupo	6

C. Pan-transcriptoma



(A) 48 transcriptomas de genótipos específicos da cana foram montados de forma automática com *YAATAP*, a partir de 4,075,446,623 reads cruas, das quais foram obtidas 2,464,540,068 leituras limpas e sem a presença de contaminantes. Estas foram utilizadas nas montagens *de novo* de seus respectivos transcriptomas com o Trinity (Grabherr *et al.,* 2011) usando kmer=[25,31], totalizando 16,237,098 transcritos nos 48 transcriptomas com ótimas métricas de qualidade. **(B)** 96.9% dos 5,240,794 transcritos codificantes de proteínas dos 48 genótipos foram atribuídos em 153,841 grupos (pan-transcriptoma) pelo OrthoFinder2, enquanto o core-transcriptoma é constituído por 8,142. **(C)** Uma das nossas principais perguntas era se com a quantidade de transcritos e genótipos disponíveis seria possível representar o conjunto de todos os transcritos da cana-de-açúcar (pan-transcriptoma), isto se observaria como um plateau da curva de número de grupos por genótipo, o qual é alcançado com apenas 24 genótipos. O core-transcriptoma é alcançado ainda mais rápido, com apenas 11 genótipos. Na anotação do pan-transcriptoma foram identificadas 4,074,811 proteínas contendo domínios proteicos conservados, 190,189 peptídeos sinalizadores e 846,378 proteínas transmembrana foram identificadas.

FINANCIAMENTO

CNPq DMRP 310080/2018-5 e 311558/2021-6
FAPESP FVP 2019/24796-5
CAPES JMMP 88887.597556/2021-00

DISPONIBILIZAÇÃO DOS DADOS

Acesse neste QR Code o link para o repositório do nosso laboratório com as montagens, anotação e conformação dos clusters do pan-transcriptoma

l1nq.com/8k3CH

REFERÊNCIAS

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A.. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52. doi: 10.1038/nbt.1883. PubMed PMID: 21572440.

Bryant D. M., K. Johnson, T. DiTommaso, T. Tickle, M. B. Couger, et al., 2017 A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 18: 762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Emms, D.M., Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20, 238 (2019). https://doi.org/10.1186/s13059-019-1832-y