INTEGRAÇÃO DE GWAS E REDES BIOLÓGICAS COMPLEXAS PARA COMPREENSÃO DOS MECANISMOS MOLECULARES ENVOLVIDOS NO CRESCIMENTO DA SERINGUEIRA

Felipe Roberto Francisco [1]*; Alexandre Hild Aono[1]; Livia Moura Souza[1]; Carla Cristina da Silva[1]; Roberto Fritsche Neto[2] and Anete Pereira de Souza[1]

[1]Institute of Biology, University of Campinas (UNICAMP), [2]Luiz de Queiroz College of Agriculture (ESALQ). *E-mail do autor para correspondência: felipe.roberto.francisco@gmail.com

Identificação do evento: VII Congresso Brasileiro de Heveicultura 10 a 12 novembro; Piracicaba (SP).

Resumo

A seringueira (*Hevea brasiliensis*) é de grande importância no cenário econômico mundial, pois é a única capaz de produzir borracha natural em quantidade e qualidade para suprir o mercado desse produto. Devido ao longo ciclo melhoramento da espécie (~ 30 anos) a *H. brasiliensis* ainda está em domesticação. Neste contexto, a biologia molecular pode contribuir para programas de melhoramento genético da seringueira por meio da seleção assistida por marcadores (MAS), que possibilita a redução desse tempo. Tal abordagem requer o entendimento e a caracterização da arquitetura genética relacionada às características agronômicas de interesse. Apesar disso, os estudos genéticos em seringueira são bastante limitados, pois a espécie possui um genoma altamente complexo. Neste trabalho, identificamos genes causais para diâmetro do caule usando um modelo GWAS, e as interações biológicas desses genes foram acessadas por meio de RNA-Seq. Encontramos quatro SNPs associados ao SD (snpsGWAS), que foram correlacionados por desequilíbrio de ligação com outros 181 SNPs (snpsLD). Dados de RNA-Seq de diferentes cultivares de *H. brasiliensis* foram acoplados ao trabalho, uma rede de coexpressão de genes foi modelada e cinco módulos funcionais que continham os SNPs foram selecionados. Esses grupos foram relacionados a categorias biológicas de para resistência a estresses abióticos e processos biológicos relacionados ao crescimento da planta. Além disso, uma rede metabólica foi construída para melhor compreender as interações enzimáticas envolvidas no crescimento da seringueira. Concluímos que a integração multiômicas possibilitou uma compreensão profunda dos mecanismos moleculares associados à configuração genética do SD, uma característica complexa amplamente utilizada em programas de melhoramento genético de seringueira.

Palavras-chave: GWAS; rede de coexpressão; rede enzimática.

Introdução

A Hevea brasiliensis é uma espécie de grande importância no cenário econômico mundial, pois é a única capaz de produzir borracha natural em quantidade e qualidade capaz para suprir as necessidades do mercado desse produto que é matéria prima para mais de 40 mil produtos (DE FAŸ e JACOB, 1989; POOTAKHAM et al., 2017). Embora exista essa grande importância econômica da seringueira, essa espécie ainda é considera em domesticação, por conta do curto período de tempo do seu cultivo, com início em 1876, e do longo ciclo de melhoramento (~30 anos) (PRIYADARSHAN e CLÉMENT-DEMANGE, 2004). O uso de marcadores moleculares tem sido proposto para se fazer uma seleção precoce dos genótipos com características de interesse em dada população a partir da seleção assistida por marcadores (SAM), o que faz necessário a descoberta de marcadores moleculares em regiões de *quantitative trait loci* (QTLs) (POOTAKHAM et al., 2017; PRIYADARSHAN, 2017).

O GWAS é uma ferramenta bastante utilizada para identificar QTLs em diversas espécies, porém apresenta limitações relacionadas a pequena proporção da variação fenotípica explicada por essas marcas (MANOLIO et al., 2009; TAM et al., 2019). Este fato combinado com a grande complexidade do genoma de *H. brasiliensis* (TANG et al., 2016; LIU et al., 2020), torna o uso de muitos QTLs descobertos por essa técnica inviável. Neste contexto a integração do GWAS com outras ômicas como redes de coexpressão e redes metabólicas pode ajudar a superar tais limitações.

Material e Métodos

Neste trabalho foi utilizada uma população contendo 438 indivíduos genotipados pela metodologia de genotipagem por sequenciamento (GBS) (ELSHIRE ET AL., 2011) e fenotipada para diâmetro de planta (SD). O SD foi escolhido por apresentar correlação com produção e vigor, possibilitando ser mensurado nos primeiros anos de plantio. A variância causada pelo efeito genotípico foi estumada usando *best linear unbiased predictor* (BLUP) com o pacote breedR (MUNÕZ e SANCHEZ, 2017) no R. A chamada de SNPs (*single nucleotide polymorphism*) foi realizada no programa TASSAL GBS 5 (GLAUBITZ et al., 2014) utilizando o genoma de referência proposto por Liu et al. (2020). A filtragem dos SNPs descobertos foi realizada no pacote do R snpReady (GRANATO e FRITSCHE-NETO, 2018).

O GWAS foi realizado a partir do programa FarmCPU (LIU et al., 2016) implementado no R. Foram incluídas as matrizes de componente principal (PC1 e PC2), com um *threshold* de significância para associação calculada a partir dos dados. Além dos SNPs identificados pelo GWAS (snpsGWAS), nós também selecionamos SNPs que a presentavam uma correlação de Pearson de $R^2 > 0.7$ ao qual chamaremos de snpsLD.

Foi utilizado um RNA-Seq obtidos dos genótipos RRIM600 e GT1 submetidos a estresse a frio (MANTELLO et al., 2019). Esse transcriptoma foi anotado utilizando o programa Trinotate (HAAS, 2015) e o banco de dados SwissProt (BOECKMANN et al., 2003). As posições dos transcritos foram estimadas se fazendo um alinhamento com BASTn (JOHNSON et al., 2008) no genoma de referência proposto por Liu et al. (2020)

Todos os SNPs identificados pelo GWAS (snpsGWAS e snpsLD) foram anotados tomando o primeiro gene mais próximo da marca nas posições *upstream* e *downstream.* Os Termos GO associados a essas anotações foram resumidas no REVIGO (SUPEK et al., 2011). Esses conjuntos de SNPs foram associados com QTLs descoberto em mapa de ligação (CONSON et al., 2018).

Uma rede de coexpressão ponderada (WGCNA) foi modelada com os dados de RNA-Seq, usando o WGCNA no R (LANGFELDER E HORVATH, 2008) utilizando a correlação de Pearson. Medida de sobreposição topológica (TOM) foi utilizada para a construção da matriz de dissimilaridade e o método *unweighted pair-group method using arithmetic avarages* (UPGMA) foi usado para definir os grupos. Os grupos contendo os SNPs (snpsGWAS e snpsLD) foram selecionados.

As enzimas presentes na rede de coexpresão foram anotadas a partir do bando de dados *kyoto encyclopedia of genes and genomes* (KEGG) (KANEHISA e GOTO, 2000). Todas as vias metabólicas da *H. brasiliensis* relacionadas aos snpsLD e snpsGWAS foram acessadas e usadas para construir uma rede metabólica usando o programa BioPython (COCK et al., 2009).

Resultados

A herdabilidade estimada na população (H^2) foi de 0,55 e identificamos 30.266 SNPs de alta qualidade utilizados para a realização do GWAS.

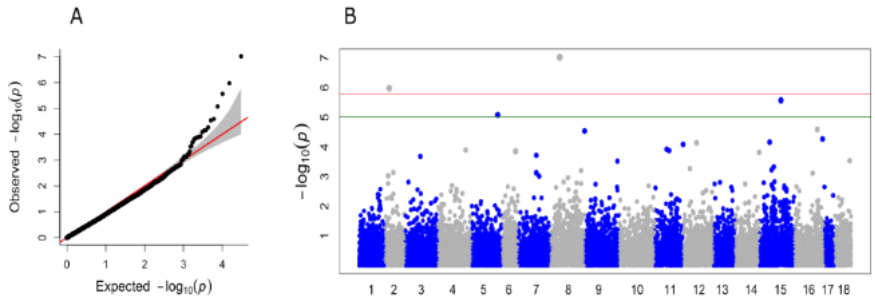

Figura 1: (A) Gráfico quantil-quantil para o modelo de associação genômica ampla (GWAS), com a inclusão do primeiro componente principal (PC1 e PC2) como uma covariável. (B) Parcela de Manhattan para o GWAS. O eixo X mostra os cromossomos contendo os marcadores descobertos em suas respectivas posições. O eixo Y mostra o log (valor p) da associação. A linha vermelha representa o *threshold* obtido com base nos dados e a linha verde representa o *threshold* corrigido de Bonferroni de 0,05.

Table 1. SNPs identifications através do GWAS.

SNP	Chrom	Position	P-value	MAF	Effect	Va	PVE	Gene
SNP6421	chrom02	14,565,718	1.06E+08	0.10	-1.00	0.18	0.05	SBT4.6
SNP30209	chrom05	75,998,329	8.38E+08	0.17	0.54	0.08	0.02	GEK1
SNP43760	chrom08	26,946,649	9.61E+06	0.45	0.84	0.35	0.09	-
SNP92152	chrom15	50,878,458	2.71E+08	0.29	0.43	0.08	0.02	IQM2

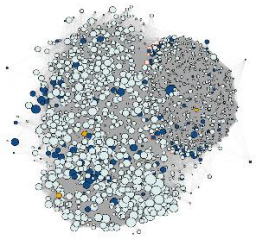

Figura 2: Rede de coexpressão contendo os módulos do gene SNP descobertos por GWAS. Amarelo mostra os genes anotados para o snpsGWAS, azul mostra os genes anotados para o snpsLD e cinza mostra os genes identificados nos módulos. Os genes destacados com uma borda vermelha representam os 10 hubs com maior conectividade, enquanto o tamanho dos nós mostra o número de genes conectados.

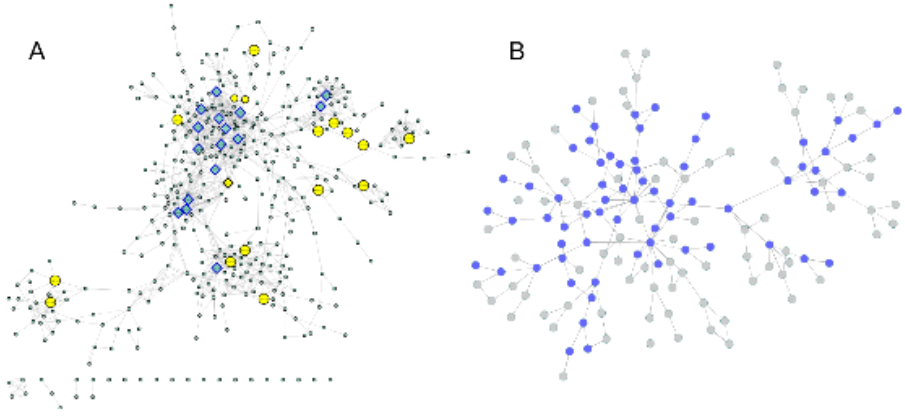

Figura 3: (A) Rede de enzimas. Os nós amarelos representam as enzimas descobertas nos módulos de coexpressão, e os nós retangulares indicam as enzimas com os maiores valores de centralidade. (B) Comunidades. Os nós azuis são representados por comunidades contendo enzimas descobertas nos módulos de coexpressão.

Discussão

Utilizamos metodologia de GWAS acoplada a outras ômicas para desvendar as bases moleculares do crescimento da seringueira. Este trabalho é a primeira iniciativa que integra a multiômica no estudo de QTLs em *H. brasiliensis*. Usando essa abordagem, fomos capazes de acessar todos os níveis moleculares importantes para a definição de SD. Apesar da grande importância econômica da espécie, por ser a única capaz de produzir borracha natural em quantidade e qualidade suficientes para abastecer o mercado mundial desse produto (DING et al., 2020), seus estudos genéticos ainda são bastante limitados devido à complexidade de seu genoma (TANG et al., 2016), sua grande variabilidade genética (DE SOUZA et al., 2018) e às grandes áreas necessárias para seu plantio. Apesar de todas essas limitações, este trabalho supera essas dificuldades, produzindo dados, resultados e novas perspectivas metodológicas para futuros estudos genômicos desta espécie e identificando marcadores e genes úteis para o melhoramento genético.

Referencias

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365-370. doi: 10.1093/nar/gkg095

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422-1423. doi: 10.1093/bioinformatics/btp163

Conson, A.R.O., Taniguti, C.H., Amadeu, R.R., Andreotti, I.A.A., De Souza, L.M., Dos Santos, L.H.B., et al. (2018). High-resolution genetic map and QTL analysis of growth-related traits of *Hevea brasiliensis* cultivated under suboptimal temperature and humidity conditions. Front. Plant Sci. 9, 1255. doi: 10.3389/fpls.2018.01255

De Faÿ, E., and Jacob, J.L. (1989). "Anatomical organization of the laticiferous system in the bark," in Physiology of Rubber Tree Latex, eds. J. D'Auzac, J. Jacob and H. Chrestin (Boca Raton, FL: CRC Press), 3-14.

De Souza, L.M., Dos Santos, L.H.B., Rosa, J., Da Silva, C.C., Mantello, C.C., Conson, A.R.O., et al. (2018). Linkage disequilibrium and population structure in wild and cultivated populations of rubber tree (*Hevea brasiliensis*). Front. Plant Sci. 9, 815. doi: 10.3389/fpls.2018.00815

Ding, Z., Fu, L., Tan, D., Sun, X., and Zhang, J. (2020). An integrative transcriptomic and genomic analysis reveals novel insights into the hub genes and regulatory networks associated with rubber synthesis in *H. brasiliensis*. Ind. Crops Prod. 153, 112562. doi: 10.1016/j.indcrop.2020.112562

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379. doi: 10.1371/journal.pone.0019379

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9, e90346. doi: 10.1371/journal.pone.0090346

Granato, I., Fritsche-Neto, R. (2018). snpReady: Preparing genotypic datasets in order to run genomic; analysis. R package version 0.9.6. Available: https://CRAN.R-project.org/package=snpReady. [Acessed July 24, 2020]

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31, 5654-5666. doi: 10.1093/nar/gkg770

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. 36, W5-W9. doi: 10.1093/nar/gkn201

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27-30. doi: 10.1093/nar/28.1.27

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. doi: 10.1186/1471-2105-9-559

Liu, J., Shi, C., Shi, C.C., Li, W., Zhang, Q.J., Zhang, Y., et al. (2020). The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis. Mol. Plant 13, 336-350. doi: 10.1016/j.molp.2019.10.017

Liu, X., Huang, M., Fan, B., Buckler, E.S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet. 12, e1005767. doi: 10.1371/journal.pgen.1005767

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753. doi: 10.1038/nature08494

Mantello, C.C., Cardoso-Silva, C.B., Da Silva, C.C., De Souza, L.M., Scaloppi Junior, E.J., De Souza Gonçalves, P., et al. (2014). De novo assembly and transcriptome analysis of the rubber tree (Hevea brasiliensis) and SNP markers development for rubber biosynthesis pathways. PLoS One 9, e102665. doi: 10.1371/journal.pone.0102665

Munõz, F., Sanchez, L. (2017). breedR: statistical methods for forest genetic; resources analysts. R package version 0.12-2. Available: https://github.com/famuvie/breedR.

Pootakham, W., Sonthirod, C., Naktang, C., Ruang-Areerate, P., Yoocha, T., Sangsrakru, D., et al. (2017). De novo hybrid assembly of the rubber tree genome reveals evidence of paleotetraploidy in Hevea species. Sci. Rep. 7, 41457. doi: 10.1038/srep41457

Priyadarshan, P.M. (2017). Refinements to Hevea rubber breeding. Tree Genet. Genomes 13, 20. doi: 10.1007/s11295-017-1101-8

Priyadarshan, P.M., and Clément-Demange, A. (2004). Breeding Hevea rubber: formal and molecular genetics. Adv. Genet. 52, 51-115. doi: 10.1016/s0065-2660(04)52003-5

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6, e21800. doi: 10.1371/journal.pone.0021800

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. 20, 467-484. doi: 10.1038/s41576-019-0127-1

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., et al. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. Nat. Plants 2, 16073. doi: 10.1038/nplants.2016.73