Identificação e Validação de SNPs por *Machine Learning* Relacionados a Caracteres de Interesse em Arroz[1]

Agnes Cardoso da Cruz[2], Ricardo Cerri[3], Marcelo Gonçalves Narciso[4], Paula Arielle Mendes Ribeiro Valdisser[5], Rosana Pereira Vianello[6] e Claudio Brondani[7]

[1] Pesquisa financiada pela Embrapa, CNPq e pela Fapesp.
[2] Engenheira-agrônoma, doutoranda em Genética e Melhoramento de Plantas, estagiária da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO
[3] Bacharel em Ciência da Computação, professor assistente da Universidade Federal de São Carlos, São Carlos, SP
[4] Engenheiro eletrônico, doutor em Computação Aplicada, pesquisador da Embrapa Arroz Feijão, Santo Antônio de Goiás, GO
[5] Farmacêutica, mestre em Genética e Biologia Molecular, analista da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO
[6] Bióloga, doutora em Biologia Molecular, pesquisadora da Embrapa Arroz Feijão, Santo Antônio de Goiás, GO
[7] Engenheiro-agrônomo, doutor em Biologia Molecular, pesquisador da Embrapa Arroz Feijão, Santo Antônio de Goiás, GO

Resumo - O arroz é o cereal mais consumido no mundo, e o aumento da produtividade é fundamental para atender a demanda crescente pelo grão. O melhoramento de arroz necessita explorar de modo mais eficiente a variabilidade genética de seus bancos de germoplasma, e os marcadores SNPs, com ampla distribuição no genoma e reduzido custo de genotipagem, são fundamentais para estimar a diversidade genética e selecionar genótipos úteis. Este trabalho objetivou identificar e validar SNPs associados a caracteres de interesse para uso no melhoramento de arroz. Para a seleção de SNPs fortemente relacionados aos caracteres, a técnica de machine learning foi aplicada a dados de fenotipagem (nove caracteres, obtidos em nove experimentos de campo) e genotipagem (4.709 SNPs espaçados a cada 200 Kpb) de 541 acessos de arroz. Para isso, foram desenvolvidos algoritmos inéditos pelas metodologias de Random Forest e XGBoost. Validando os 15 SNPs identificados conduziu-se experimento em blocos ao acaso com quatro repetições, avaliando 29 acessos do banco de germoplasma e duas testemunhas. Os SNPs foram convertidos em ensaios TaqMan, possibilitando a análise em aparelho de PCR quantitativo. Dos 15 ensaios TaqMan, seis discriminaram os acessos avaliados no experimento, com destaque para o SNP_S6_30 (A/G), cujo padrão G/G (presente em 23 acessos) apresentou média superior de produtividade ($p < 0,05$) e número de panículas ($p < 0,01$) e, portanto, está validado para a seleção de genótipos de arroz mais produtivos. Os 15 ensaios TaqMan serão avaliados em novo experimento envolvendo acessos da coleção de arroz oriunda dos Estados Unidos.