FUNCTIONAL PLANT BREEDING JOURNAL
Scientific Journal. ISSN 2595-9433 - D.O.I. http://dx.doi.org/10.35418/2526-4117/v7a01
Volume 7, Article n. 01, Jan/December - Received: 06/09/2024 - Accepted: 12/01/2025

Identification and analysis of SNP markers associated with traits of interest in rice using Machine Learning methodology

Agnes Cardoso da Cruz[1], Marcelo Gonçalves Narciso[1], Ricardo Cerri[2], Paula Arielle Valdisser[1], Lucas Matias Gomes Messias[1], Breno Osvaldo Funicheli[2], Rosana Pereira Vianello[1], Claudio Brondani[1]

Abstract: The use of molecular markers to select superior individuals for traits of interest is essential to accelerate the development of rice cultivars. Quantitative traits are challenging to work with in marker-assisted selection, and new methodologies must be continually evaluated. This study aimed to identify SNP markers associated with five quantitative traits through the Machine Learning (ML) methodology, which used genotyping (4,709 SNPs) and grain yield data from 541 accessions from Embrapa's Rice Core Collection evaluated in nine locations. Fifteen TaqMan® hydrolysis probe-based assays were developed from SNPs associated with key traits, and 31 rice varieties were both genotyped and phenotyped for validation. Using simple linear regression analysis, four SNPs were significantly associated with panicle number and grain yield, while three were linked to the percentage of filled grains. The application of machine learning methods, coupled with the evaluation of selected SNPs and the development of TaqMan® assays, provided an effective approach for identifying markers to support routine marker-assisted selection in rice breeding programs.

Keywords: Genome, molecular markers, *Oryza sativa*, marker assisted selection.

Introduction

Rice (*Oryza sativa*) has hundreds of thousands of unique varieties stored in genebanks worldwide, representing one of the largest collections of genetic resources among plant species of economic interest (Pathirana and Carimi, 2022). The exclusive use of elite parents in breeding has resulted in an increase in annual production of about 1% per year, which is a lower rate than that needed to meet projected consumption demand for 2050, estimated at 2.4% per year (Ray et al., 2013). Over time, increases in productivity gains through selection are only possible if additional genetic variability is introduced into breeding populations. Much of the genetic diversity present in germplasm banks has proven to be extremely important for breeding (Jamora and Ramaiah, 2022). In

[1] *Embrapa Arroz e Feijão, Santo Antônio de Goiás, Goiás State, Brazil.*
[2] *Universidade Federal de São Carlos, São Carlos, São Paulo State, Brazil.*
* *Corresponding author: claudio.brondani@embrapa.br*

the search for better and more efficient alternatives to develop new cultivars that result in greater production sustainability, rice lines with the best gene and allelic combinations must be identified, and molecular markers can monitor this variability over generations during the development of new cultivars.

Single nucleotide polymorphism (SNP) markers have been widely used in genetic analyses due to their wide genomic distribution, increasing the ease of access to genotyping and reducing analysis costs (Voss-Fels et al., 2016). DNA sequencing and phenotyping platforms can identify genes related to quantitative traits, such as productivity, through QTL (Quantitative Trait Loci) and GWAS (Genome Wide Association Studies) analyses (Mochida et al., 2018). Zhang et al. (2021) resequenced 450 rice accessions and found an average of 25 SNPs associated with each trait evaluated. This wide availability of markers allows the development of marker-assisted selection strategies, increasing the chance that these markers will identify lines with a favorable phenotype (Gentzbittel et al., 2019). Phenotype:SNPs association by GWAS was also reported as the first step toward the cloning of genes/QTLs associated with the grain size trait, allowing inferences about the regulatory mechanism with the aim of providing the theoretical basis for improving the productivity of rice cultivars (Jiang et al., 2022).

Both selection and population history have important influences on the amount and patterns of genetic variability, and consequently, populations with different genetic histories can have differences in allele frequencies for many genome-wide polymorphisms (Flood and Hancock, 2017). If these populations have different values for the phenotype, any polymorphisms that differ in frequency between two populations will be associated with the phenotype, even if they are neither causal nor in strong linkage equilibrium with the causal polymorphism (Gentzbittel et al., 2019).

Methods that aim to identify causal loci are therefore highly influenced by population structure. For this reason, new strategies based on artificial intelligence, such as Machine Learning (ML), are being applied, aiming to explain variation in complex traits using populations with admixture proportions and capable of predicting quantitative phenotypes. By including genotypes with admixture or selection, Machine Learning can explain more of the phenotypic variation than methods such as GWAS or QTL mapping (Gentzbittel et al., 2019). ML methodology involves areas of computer science, artificial intelligence, computational statistics and information theory to build algorithms that can learn from existing datasets and make predictions on new datasets (Wang et al., 2018). ML allows the exploration of Big Data concepts in plant genomics, which involves analyzing a huge amount of data and extracting knowledge to understand cellular mechanisms or the expression of complex traits. Some steps are common to several ML studies, such as (i) choosing the database, which must be accurate and with minimal redundancy; (ii) extracting information from the selected data; (iii) evaluating and selecting the main attributes to be investigated; (iv) choosing algorithms; v) creating prediction/classification models; and (iv) evaluating prediction/classification performance (Silva et al., 2019).

ML is well suited to infer nonlinear relationships in biological systems due to the large volume and variety of data from different categories that must be included in a complete analysis of a given dataset. Several algorithms have been used in ML, such as artificial neural networks, random forest, LogitBoost, and support vector machines (Tong et al.,

2024). The aim of this work was to identify SNPs associated with traits of interest in rice using ML methodology and then convert them into a viable genotyping system for use in assisted selection in rice breeding programs.

Material and Methods

Identification of SNPs associated with traits of interest by Machine Learning (ML)

The grain yield data, previously described in Bueno et al. (2012), were obtained in nine experiments (Boa Vista, Santo Antônio de Goiás, Goianira Year 1, Goianira Year 2, Vilhena, Teresina, Sinop, Uruguaiana and Pelotas) carried out in the Federer Augmented Block design. The data from the 541 accessions of these experiments were previously analyzed using mixed linear models and the estimates of the variance components were obtained using the residual maximum likelihood (REML) method, with application of the best linear unbiased prediction (BLUP) procedure to predict the random effects genetic values (eBLUP) associated with each of the accessions (Bueno et al., 2012). These phenotyping data were associated, for each accession, with data from 4,709 SNPs obtained by the Genotyping by Sequencing (GBS) methodology, regularly spaced in the genome, selected from among 445,589 SNPs (Pantalião et al., 2016).

Using the phenotypic and genotypic data, methods based on the random forest algorithm (Breiman, 2001) were developed, and two strategies were used to classify the SNPs related to the nine characteristics of interest, based on the "Increased Mean Square Error" (iMSE) and the "Mean Decrease of Gini index" (MDGI). In the first strategy, for each trait of interest, a random forest algorithm was run 1000 times on the entire phenotypic and genotypic dataset. In each of these 1000 runs, the iMSE was calculated for each of the SNPs in the dataset. For each decision tree that makes up the forest of the random forest algorithm, the iMSE was calculated using the portion of the data that was not used to train the tree. Using this strategy, the SNPs that were most frequently in the top ranking positions were selected. The second method employed uses the MDGI calculation to rank the SNPs used in the random forest. In decision trees, the Gini Index defines the purity of a tree node and how much an attribute managed to divide a data set into pure subsets, in which the examples of each subset belong to the same class.

The Gini index value lies in the range between 0 and 1, where 0 represents a pure classification, that is, all examples in the set belong to the same class. A value of 1 indicates a random distribution of examples within the set across several classes. A value of 0.5 indicates that there is an equal distribution of examples across classes. When constructing the decision tree, the attributes (in this case, SNPs) that have the lowest Gini index value are considered the most important. The random forest algorithm was run 1000 times, as in the previous strategy. With each run of the algorithm, the MDGI measure was used to select the best SNPs from the dataset. At the end of the runs, the SNPs that appeared most frequently among the best ranked were selected.

Development of TaqMan® assays for the identified SNPs

After the selection of SNPs by ML, an in silico analysis was performed to select those with the greatest potential for association with the grain yield trait. The following criteria were used: 1) Presence of the SNP in the RiceVarMap rice sequence database (*http://ricevarmap.ncpgr.cn/vars_in_region/*), to infer that the SNP is real, without the need to sequence the genomic DNA fragment that includes the SNP;

2) SNP with one of the genotypic classes with a minimum allele frequency of 0.05 (5%), to avoid selecting a SNP locus that has a pattern (G/G, for example) with high frequency in a given set of rice genotypes; 3) Effects of predicted variations (obtained from the website RiceVarMap) and classified as: a) modifier (the SNP is located in an intergenic region, 3' UTR, 5' UTR or in an intron); b) moderate (located in an exon, where the SNP changes the amino acid); and c) low (in an exon, where the SNP does not change the amino acid, due to the degeneration of the genetic code); 4) Functional annotation of the gene modified by the SNP, i.e., assignment of the gene function, obtained from the Rice Genome Annotation Project website (*http://rice.uga.edu/analyses_search_locus.shtml*); and 5) Regarding the number of genes regulated by the gene modified by the SNP, obtained from the RiceNETDB website (*http://bis.zju.edu.cn/ricenetdb/*).

The SNPs identified by ML and selected by in silico analysis were converted to TaqMan® assays (ThermoFisher Scientific, USA), based on hydrolysis probes. For this, the position of the SNP in the genome was identified by the Genome Browse tool (*http://rice.uga.edu/cgi-bin/gbrowse/rice/#search*), and sequences 250 base pairs (bp) above and 250 bp below this position were selected. The TaqMan® assay development process requires a high-quality target sequence for oligonucleotide design. Therefore, it is essential to identify the presence of repetitive DNA in the DNA fragment adjacent to the SNP using the RepeatMasker program (*https://repeatmasker.org/cgi-bin/WEBRepeatMasker*), which automatically inserts "N" at each nucleotide of the repeat, which prevents the primer design program from using this part of the genome sequence.

Furthermore, the 501 bp sequence was evaluated for the presence of non-target SNPs using the SNPseek program (*https://snp-seek.irri.org/_snp.zul*), which are also replaced by "N". This "masked" sequence was then submitted to the "Custom TaqMan® Assay Design Tool" program (*https://www.thermofisher.com/order/custom-genomic-products/tools/ga/ga_index.php?defaultApp=cadt*), which automatically designs the primers and probe that hybridizes at the point where the SNP is located. Each SNP allele is labeled with fluorescence (VIC for the reference nucleotide allele and 6-FAM for the alternative allele), thus allowing their distinction in the analysis performed on an RT-qPCR device.

Greenhouse experiment for SNP validation

To validate the SNPs associated with grain production by ML, 29 contrasting rice varieties were selected regarding productivity, height and cycle, in addition to two cultivars as checks (Table 1). The experiment was conducted in a greenhouse located at Embrapa Arroz e Feijão (Santo Antônio de Goiás, GO, 16º28'S; 49º17'W, 779 m altitude).

The design was in randomized blocks with four replications, and the evaluated characteristics were production (acronym GY, in grams, g), plant height (PH, measurement from the soil to the point of insertion of the panicle in the primary tiller, in centimeters, cm), number of panicles (PAN) and percentage of filled grains (PFG). Analysis of variance and the test of means were performed using the R program (R Core Team, 2023).

The validation of the SNP converted into TaqMan® assay was performed by genotyping, together with the phenotyping data of the 31 materials evaluated in the experiment under controlled conditions. Initially, the genomic DNA of each rice variety was extracted according to the procedure described in Pantalião et al. (2016). Subsequently, the Custom TaqMan® SNP Genotyping Assays 40 X solution and the TaqMan® GTXpress

master mix 2 X were added to the DNA of each genotype, according to the manufacturer's recommendations, for a final volume of 5 µl.

Table 1. Materials evaluated in the greenhouse experiment to validate the TaqMan® assays. BAG: Active Germplasm Bank (Embrapa Rice and Beans); C.S.: cultivation system, U: upland; L: lowland.

	BAG Code	Name	C.S.
1	BGA011943	Caqui	U
2	BGA000003	Amarelão	U
3	BGA000009	Agulhinha	L
4	BGA000941	Ciwini	L
5	BGA000994	Esav X Matão	U
6	BGA001416	IPSL 0574	L
7	BGA002482	M 44	L
8	BGA002524	Moroberekan	U
9	BGA003196	GZ 809-4-1-2	L
10	BGA004206	CNA 108-B	U
11	BGA004463	N7441	U
12	BGA004566	Metica-1	L
13	BGA004697	N2583	U
14	BGA005342	Rio Verde	U
15	BGA005461	CO 18	U
16	BGA006030	Padi Senemok	U
17	BGA006170	LS 85-125	U
18	BGA006574	Irat 112	U
19	BGA006666	A12-286	U
20	BGA008412	Bluebonnet	U
21	BGA008545	CT 11216	U
22	BGA008711	CNAx 4914	U
23	BGA009364	CT13581	U
24	BGA010692	Oryzica Lhanos 4	L
25	BGA020067	Cambará	U
26	BGA003490	Mearin	U
27	-	AB172678	U
28	-	AB172748	U
29	-	BRS A503	U
Check 1	**BGA015465**	**BRS Esmeralda**	**U**
Check 2	**BGA008070**	**Primavera**	**U**

Reactions were performed in the QuantStudio 7 Flex RealTime PCR System (Applied Biosystems, USA) under the following amplification conditions: 30 s at 60°C, 20 s at 95°C followed by 50 cycles of 3 s at 95°C and 1 min at 60°C, with a final step of 30 s at 60°C. SNP genotyping was performed using TaqMan® Genotyper Software® (ThermoFisher Scientific, USA). From the genotypes of the SNP markers integrated with the phenotypic values, a simple linear regression analysis (p-value ≤ 0.10) based on the nonparametric Wilcoxon test was performed using R software, version 4.1.0 (R Core Team, 2023).

Results

Of the 74 SNPs identified by the ML approach as associated with grain yield, 61 (82%) were identified in the RiceVarMap database and proceeded to the next steps of SNP selection. Based on filtering using MAF ≤ 5%, all 61 SNPs identified via ML were retained. The next step was to determine the position of the 61 SNPs in the genome, which was distributed as follows: intergenics, 34 occurrences; in the introns, 9; in the 3' UTR regions, 6; and in the exons, 12. A SNP in an exon can result in an amino acid change (missense variant), or without amino acid modification (synonymous variant). To complete this SNP selection step, it is important to determine the effect of the SNP (modifier, moderate or low). The vast majority of SNPs were categorized as having a modifier effect (49), followed by a moderate effect (6) and a low effect (6).

Of the 61 SNPs evaluated, 46 had an effect on genes with a putative gene product already described, while the remaining 15 were classified as expressed protein, that is, their gene product has yet to be determined. Another criterion for evaluating SNPs was to find the interaction of genes that are under the effect of the SNP with other genes and, when possible, their participation in metabolic pathways. According to these criteria, 15 SNPs were chosen for the development of the TaqMan® assays (Table 2). Chromosomes 1, 2, 3, 6, 7, 8 and 9 were represented. The selected SNPs are located, in decreasing order of occurrence, in introns (6), intergenic regions (5), exons (2), 3' UTR region (1) and 5' UTR region (1).

Table 2. Description of the 15 SNPs that originated the TaqMan® assays. In the variation, the first allele is the reference allele and the second, the alternative allele. Chrom.: chromosome number. (a.t.): associated with evaluated traits.

SNP	Chrom.	Variation	Location
S1_32807862 (a.t.)	1	A/G	Intergenic (LOC_Os01g56810-LOC_Os01g56820)
S1_33836751 (a.t.)	1	G/A	Intergenic (LOC_Os01g58540-LOC_Os01g58550)
S1_39254110 (a.t.)	1	G/C	Intron (LOC_Os01g67530)
S2_17491825 (a.t.)	2	A/G	Intron (LOC_Os02g29464)
S3_2150128 (a.t.)	3	T/C	5' UTR (LOC_Os03g04600)
S6_26027620 (a.t.)	6	T/C	Intron (LOC_Os06g43304)
S6_30346911 (a.t.)	6	G/A	3' UTR (LOC_Os06g50100)
S1_22936476	1	T/C	Intron (LOC_Os01g40610)
S1_41924060	1	G/A	Intron (LOC_Os01g72310)
S2_35383271	2	G/C	Intron (LOC_Os02g57780)
S3_14416044	3	A/G	Intergenic (LOC_Os03g25220-LOC_Os03g25240)
S6_5720734	6	T/A	Intergenic (LOC_Os06g10950-LOC_Os06g10960)
S7_1055849	7	A/C	Exon (LOC_Os07g02810); missense (Asn/Thr)
S8_20351530	8	C/T	Exon (LOC_Os08g32840); synonymous (Asp/Asp)
S9_1420365	9	A/T	Intergenic (LOC_Os09g03010-LOC_Os09g03030)

The analysis of variance of the experiment conducted in a greenhouse showed that there was a significant difference between the treatments for all traits, except for the percentage of filled grains (Table 3). The traditional variety Caqui stood out in terms of productive performance (244.01 g/plant) and number of panicles (53), which is favorable, but it also presented undesirable traits for the breeding program, such as late cycle (119 days to flowering) and greater height.

Table 3. Analysis of variance of the greenhouse experiment. SV: source of variation; DF: degrees of freedom; SS: sum of squares; MS: mean sum of squares; CV%: coefficient of variation.

Source		DF	SS	MS	F-statistic	P-value	CV%
Yield	Genotypes	30	211949	7065	7.3168	0**	42.87
	Block	3	3553	1184.3	1.2265	0.3048ns	
	Residual	90	86903	965.6			
	Total	123	302405				
Flowering	Genotypes	30	22654.2	755.14	35.054	0**	4.82
	Block	3	84.7	28.24	1.311	0.2758ns	
	Residual	90	1938.8	21.54			
	Total	123	24677.7				
Panicle number	Genotypes	30	17148.4	571.61	7.7218	0**	34.27
	Block	3	172.9	57.64	0.7787	0.5089ns	
	Residual	90	6662.3	74.03			
	Total	123	23983.6				
Height	Genotypes	30	53083	1769.42	21.1599	0**	10.77
	Block	3	891	297.14	3.5535	0.0175ns	
	Residual	90	7526	83.62			
	Total	123	61500				
% Filled grains	Genotypes	30	0.54875	0.018292	1.17483	0.276ns	13.73
	Block	3	0.03168	0.010561	0.67834	0.5675ns	
	Residual	90	1.40127	0.01557			
	Total	123	1.98171				

**: significant (p< 0,01); ns: not significant

The accession that stood out the most, in the set of traits evaluated, was the cultivar Metica-1, with good productivity (127.22 g/plant), medium cycle (98 days), good number of panicles (41) and low height (69 cm) (Table 4).

Table 4. Test of means of the evaluated traits. Means followed by the same letter, in the column, do not differ from each other at 5% probability by the Scott-Knott test. C: check cultivar.

Name	Yield	Flowering	Panicle	Height
Caqui	244.01a	119.25b	53,25a	132,75a
CO 18	146.76b	118.50b	58,50a	122,25a
Metica-1	127.22b	98.75c	40,75b	69,00e
Ciwini	101.54c	106.50c	34,50b	76,00d
Padi Senemok	100.33c	115.25b	33,25b	108,88b
CT13581	98.36c	99.00c	36,50b	99,00c
Mearin	92.60c	100.75c	35,25b	64,75e
Agulhinha	91.25c	106.50c	18,75c	118,38a
Moroberekan	88.76c	106.50c	15,25c	100,88b
Oryzica Lhanos 4	75.83c	114.00b	27,25b	63,63e
Bluebonnet	73.78c	102.50c	16,50c	103,75b
GZ 809-4-1-2	72.01c	103.50c	31,75b	53,25e
A12-286	69.54c	92.50d	15,75c	80,88d
IPSL 0574	61.83d	89.50d	13,25c	94,88c
N2583	61.44d	127.25a	20,75c	109,25b
Esav X Matão	56.45d	104.25c	14,50c	119,13a
AB172678	55.46d	84.00e	34,75b	65,25e
Rio Verde	54.53d	95.00d	16,50c	81,88d
CNA 108-B	49.14d	87.75d	17,50c	84,50d
BRS Esmeralda (C)	48.84d	82.00e	31,00b	72,25d
M 44	48.20d	96.00d	12,00c	91,50c
Amarelão	47.89d	77.00f	25,00c	79,25d
Irat 112	47.89d	76.00f	30,50b	65,25e
BRS A503	47.35d	88.75d	15,50c	72,13d
AB172748	46.59d	84.75e	29,00b	69,63e
N7441	44.66d	83.25e	21,50c	64,50e
LS 85-125	43.74d	91.25d	8,75c	93,00c
CT 11216	38.73d	87.25d	11,75c	62,13e
Primavera (C)	38.49d	78.25f	19,00c	75,13d
Cambará	37.23d	92.75d	17,75c	67,75e
CNAx 4914	36.83d	75.50f	22,00c	70,58e

The 15 TaqMan® assays genotyped 31 materials evaluated in the greenhouse experiment, and the molecular data of SNP markers were analyzed together with the phenotypic data by means of simple linear regression analysis. There was significance for the effects in seven of the 15 SNP markers, which were related to the traits number of panicles (five SNPs), percentage of whole grains (three SNPs) and grain yield (four SNPs) (Table 5).

The proportions of trait variation explained by the individual marker are an important estimate for the implementation of marker-assisted selection. For the trait percentage of filled grains, the variations (R^2) explained by the effects of the markers were high, ranging from 34.3 to 50.3%.

For number of panicles, the variations ranged from 17.5 to 22.6%, while for grain yield, they ranged from 7 to 56.1%. SNPs S1_33836751, S1_39254110, S2_17491825 and S3_2150128 explained the variations for both panicle number and grain yield traits. In turn, the markers S6_26027620, S6_30346911 and S1_32807862 were associated only with the trait percentage of filled grains. The largest variations were explained by markers S3_2150128 (56.1%; grain yield), S2_17491825 (55.3%; number of panicles) and S1_32807862 (50.3%; percentage of filled grains).

A direct association with the phenotype of a complex inheritance trait cannot be attributed to a single gene. Furthermore, the highest level of complexity of the genetic control of a quantitative trait also results from the interaction between genes. In this study, all genes modified by the SNPs associated with the traits evaluated were associated with other genes, ranging from 31 to 87 genes, some of which were transcription factors, which play a fundamental role in gene expression (Table 6).

Table 5. Summary of regression analysis between SNP markers and respective traits. DF: degrees of freedom; SS: sum of squares; MS: mean sum of squares.

Source of variation		DF	SS	MS	F	p-value	R^2 (%)
S6_26027620 % filled grains	TT vs CC(1)	1	0.042144	0.042144	15.63	5.00E-04	34.3
	Residual	27	0.072782	0.003			
	Total	28	0.114926	0.0			
S6_30346911 % filled grains	AA vs GG(1)	1	0.048134	0.048134	19.46	1.8E-04	39.7
	Residual	27	0.066792	0.002			
	Total	28	0.114926	0.1			
S1_32807862 % filled grains	AA vs GG(1)	1	0.059813	0.059813	29.30	1.01E-05	50.3
	Residual	27	0.055113	0.002			
	Total	28	0.114926	0.1			
S1_33836751 Panicle Number	AA vs GG(1)	1	1072.3	1072.3	9.67	0.004	22.4
	Residual	29	3214.8	110.9			
	Total	30	4287.1	1183.2			
S1_39254110 Panicle Number	CC vs GG(1)	1	856.2	856.2	6.74	0.015	17.5
	Residual	26	3302.8	127.0			
	Total	27	4159	983.2			
S3_2150128 Panicle Number	CC vs TT(1)	1	1069.3	1069.3	9.47	0.005	22.6
	Residual	28	3161.9	112.9			
	Total	29	4231.2	1182.2			
S6_30346911 Panicle Number	AA vs GG(1)	1	469.6	469.6	3.50	0.072	7.9
	Residual	28	3761.6	134.3			
	Total	29	4231.2	603.9			
S2_17491825 Panicle Number	AA vs GG(1)	1	2300	2300	33.12	5.3E-06	55.3
	Residual	25	1736.2	69.4			
	Total	26	4036.2	2369.4			
S1_33836751 Grain Yield	AA vs GG(1)	1	5359	5359	3.26	0.081	7.0
	Residual	29	47628	1642.3			
	Total	30	52987	7001.3			
S1_39254110 Grain Yield	CC vs GG(1)	1	20072	20072	16.87	0.0004	37.0
	Residual	26	30936	1189.8			
	Total	27	51008	21261.8			
S3_2150128 Grain Yield	CC vs GG(1)	1	5121	5121	3.08	0.0903	6.7
	Residual	28	46581	1663.6			
	Total	29	51702	6784.6			
S2_17491825 Grain Yield	AA vs GG(1)	1	28761	28761	34.28	4.1E-06	56.1
	Residual	25	20977	839.1			
	Total	26	49738	29600.1			

(1) Contrast considered in the regression analysis between the marker locus and the evaluated trait.

Table 6. Information on genes modified by SNPs associated with the evaluated traits. PFG: Percentage of filled grains; PAN: Panicle number; GY: Grain yield.

SNP	Trait	Modified Gene	Gene Product	Associated Genes*
S6_26027620	PFG	LOC_Os06g43304	Cytochrome P450	37 (5)
S6_30346911	PFG/PAN	LOC_Os06g50100	Tyrosine kinase	50 (7)
S1_32807862	PFG	LOC_Os01g56810	Cytokinin Dehydrogenase/oxidase	65 (2)
S1_33836751	GY/PAN	LOC_Os01g58550	Methyladenine glycosylase	31
S1_39254110	GY /PAN	LOC_Os01g67530	AMP-binding enzyme	29 (7)
S3_2150128	GY /PAN	LOC_Os03g04590	Ribosomal protein	87
S2_17491825	GY /PAN	LOC_Os02g29464	SMC N-terminal domain	38

* In parentheses: number of transcription factors

Discussion

Although rice is native to Asia, it has been cultivated in Brazil since the 16th century (Pereira, 2002), leading to unique genomic variations driven by adaptation to tropical soils and climate. This diversity likely explains why certain SNPs from Brazilian germplasm are absent from international rice genome databases. Specifically, 13 SNPs were not found in these databases and appear to be exclusive to rice germplasm grown in Brazil. In the present study, we focused on SNPs present and validated in international databases. Most of the 61 SNPs associated with grain yield were located in intergenic regions. The importance of SNPs in intergenic regions is that they can alter the expression of adjacent genes, due to the interference of the SNP in the expression of regulatory elements of these genes (Chen and Tian, 2016). SNPs located in intergenic regions, introns, 3'UTR and 5' UTR have modifier gene effects, while SNPs located in exons have moderate or low effects. An SNP in an exon can result in an amino acid change (missense variant), without an amino acid modification (synonymous variant), or result in a stop codon (nonsense variant).

Rice is the most studied grass, and this has generated a large accumulation of knowledge about the putative function of genes (Jiang et al., 2018). The accessibility of this information is a great advantage and represents an important criterion in the selection of SNPs that may be present in a gene, or altering the expression of these genes. Most SNPs affected genes with known putative products, helping the identification of metabolic pathways and interacting genes, which can support various biotechnological applications, including genome editing.

The seven TaqMan® assays, based on greenhouse data, were associated with panicle number, percentage of filled grains, and grain yield — key traits for productivity. This supports that the SNPs identified by the ML approach using CNAE phenotypic data are indeed linked to grain yield. These SNPs, located in non-coding regions (3'UTR, 5'UTR, introns, and intergenic areas), showed a modifier effect on nearby genes. According to Cano-Gamez and Trynka (2020), several significant SNPs by GWAS methodology were identified in non-coding regions and were assigned as regulatory variants of adjacent genes. Furthermore, enhancers, promoters and long non-coding RNA are the main genome elements strongly influenced by SNPs, which implies that most of the SNPs associated with phenotypes are located in non-coding regions of the genome and are the most likely to regulate gene expression of QTLs (Fabo and Khavari, 2023).

A direct association with phenotype of a complex inheritance trait cannot be attributed to a single gene. PFG has been associated with three genes, Cytochrome P450, Tyrosine kinase and Cytokinin Dehydrogenase/oxidase.

A homologue of this Cytochrome P450 gene (LOC_Os06g15680) has already been related to the 100-grain weight trait, one of the components of rice productivity, by Silva et al. (2022). This independent study indicates that this gene, which belongs to a superfamily of enzymes, actively participates in pathways related to increased productivity, and participates in the biosynthesis of several endogenous molecules, essential secondary metabolites and fatty acids (Naveed et al., 2018). Protein tyrosine kinase was associated with increased biomass (Liu et al., 2015) and response to abiotic stresses (Elangovan et al., 2020). The enzyme cytokinin dehydrogenase/oxidase was associated with increased biomass and response to abiotic stress (Zhang et al., 2021). An AMP (adenosine

monophosphate) binding protein gene, associated with the traits yield and panicle number, was related to the defense response to *Magnaporthe oryzae* (Zhang et al., 2009), the main rice disease. The other genes were associated with proteins involved in general cellular metabolism (methyladenine glycosylase, ribosomal protein and MEC - structural maintenance of N-terminal chromosomes), and showed a large number of genes regulated by these three genes.

These seven genes are linked to over 300 others, suggesting that the corresponding SNPs act as modifiers and influence the expression of these additional genes. This is a very important point and justifies further analyses to validate the use of these SNPs in assisted selection in rice. Furthermore, among all the more than 300 genes, 21 are transcription factors, which are a group of proteins that regulate multiple genes simultaneously and, therefore, are promising targets for improving important traits, such as grain yield (Watt et al., 2020). As a perspective, the seven SNPs will be integrated into the routine genetic characterization of the parents and lines of the rice breeding program, aiming to validate the response of these markers to the respective associated phenotypes.

Conclusion

The identification of robust SNPs associated with traits using machine learning (ML) methodology, followed by the evaluation of these SNPs through TaqMan® genotyping assays, highlights the potential for integrating these markers into the routine marker-assisted selection process in rice breeding programs. This approach combines advanced genomic tools with practical breeding strategies, enhancing the precision and efficiency of selection processes. It holds promise for developing rice varieties with improved yield and resilience to environmental stress, ultimately contributing to more sustainable and productive rice cultivation.

References

Breiman, L. 2001. Random Forests. **Machine Learning**, 45:5–32. https://doi.org/10.1023/A:1010933404324.

Bueno, L. G.; Vianello, R. P.; Rangel, P. N.; Utumi, M. M.; Centeno, A. C.; Pereira, J. A.; Franco, D. F.; Neto, F. M.; Mendonça, J. A.; Coelho, A. S.; Oliveira, J. P.; Brondani, C. 2012. Adaptabilidade e estabilidade de acessos de uma coleção nuclear de arroz. **Pesquisa Agropecuária Brasileira**, 47(2):216–226. https://doi.org/10.1590/S0100-204X2012000200010.

Cano-Gamez, E.; Trynka, G. 2020. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. **Frontiers in Genetics**, 11:424. doi: 10.3389/fgene.2020.00424

Chen, J.; Tian, W. 2016. Explaining the disease phenotype of intergenic SNP through predicted long range regulation. **Nucleic Acids Research**, 44(8641–8654). doi: 10.1093/nar/gkw519.

Elangovan, A.; Dalal, M.; Krishna, G.M.; Devika, S.; Kumar, R.R.; Sathee, L.; Chinnusamy, V. 2020. Characterization of atypical protein tyrosine kinase (PTK) genes and their role in abiotic stress response in rice. **Plants**, 9:664. doi: 10.3390/plants9050664.

Fabo, T.; Khavari, P. 2023. Functional characterization of human genomic variation linked to polygenic diseases. **Trends in Genetics**, 39(6):462-490. doi: 10.1016/j.tig.2023.02.014.

Flood, P.J.; Hancock, A.M. 2017. **Current Opinion in Plant Biology**, 36:88–94. https://doi.org/10.1016/j.pbi.2017.02.003.

Gentzbittel, L.; Ben, C.; Mazurier, M.; Shin, M.; Lorenz, T. 2019. WhoGEM: an admixture-based prediction machine accurately predicts quantitative functional traits in plants. **Genome Biology**, 20: 106. https://doi.org/10.1186/s13059-019-1697-0.

Jamora, N.; Ramaiha, V. 2022. Global demand for rice genetic resources. **CABI Agriculture and Bioscience**, 3:26. https://doi.org/10.1186/s43170-022-00095-6.

Jiang, J.; Xing, F.; Wang, C.; Zeng, X. 2018. Identification and analysis of rice yield-related candidate genes by walking on the functional network. **Frontiers in Plant Science**, 9:1685. doi: 10.3389/fpls.2018.01685.

Jiang, H.; Zhang, A.; Liu, X.; Chen, J. 2022. Grain size associated genes and the molecular regulatory mechanism in rice. **International Journal of Molecular Sciences**, 23:3169. https://doi.org/10.3390/ijms23063169.

Liu, S.; Hua, L.; Dong, S.; Chen, H.; Zhu, X.; Jiang, J.; Zhang, F.; Li, Y.; Fang, X.; Chen, F. 2015. OsMAPK6, a mitogen-activated protein kinase, influences rice grain size and biomass production. **The Plant Journal**, 84:672–681. doi: 10.1111/tpj.13025.

Mochida, K.; Koda, S.; Inoue, K.; Hirayama, T.; Tanaka, S.; Nishii, R.; Melgani, F. 2018. Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. **GigaScience**, 8:1. doi: 10.1093/gigascience/giy153.

Naveed, A.; Li, H.; Liu, X. 2018. Cytochrome P450s: blueprints for potential applications in plants. **Journal of Plant Biochemistry & Physiology**, 6:100204. doi: https://doi.org/10.4172/2329-9029.1000204.

Pantalião, G.F.; Narciso, M.; Guimarães, C.; Castro A.; Colombari, J.M.; Breseghello, F.; Rodrigues, L.; Vianello, R.P.; Borba, T.O.; Brondani, C. 2016. Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. **Genetica**, 144:651-664. 10.1007/s10709-016-9932-z.

Pathirana, R.; Carimi, F. 2022. Management and utilization of plant genetic resources for a sustainable agriculture. **Plants**, 11:2038. https://doi.org/10.3390/ plants11152038.

Pereira, J.A. 2002. **Cultura do Arroz no Brasil: Subsídios para sua história**. Embrapa Meio Norte, Teresina. 226 p.

Ray, D.K.; Mueller, N.D.; West, P.C.; Foley, J.A. 2013. Yield trends are insufficient to double global crop production by 2050. **Plos One**, 8:e66428. https://doi.org/10.1371/journal.pone.0066428.

R Core Team. 2023. **R: A Language and Environment for Statistical Computing**. Available from URL. http://www.r-project.org/.

Silva, D.A.R.; Mendonça, J.A.; Cordeiro, A.C.C.; Magalhães Júnior, A.M. De; Vianello, R.P.; Brondani, C. 2022. Identification of stable quantitative trait loci for grain yield in rice. **Pesquisa Agropecuária Brasileira**, 57:e02812, 2022. DOI: https://doi.org/10.1590/S1678-3921.pab2022.v57.02812.

Silva, J.C.F.; Teixeira, R.M.; Silva, F.; Brommonschenkel, S.H.; Fontes, E.P.B. 2019. Machine learning approaches and their current application in plant molecular biology: A systematic review. **Plant Science**, 284:37–47. *https://doi.org/10.1016/j.plantsci.2019.03.020.*

Tong, K.; Chen, X.; Yan, S.; Dai, L.; Liao, Y.; Li, Z.; Wang, T. 2024. PlantMine: a machine-learning framework to detect core SNPs in rice genomics. **Genes**, 15:603. *https://doi.org/10.3390/genes15050603*.

Voss-Fels, K.; Snowdon, R.J. 2016. Understanding and utilizing crop genome diversity via high-resolution genotyping. **Plant Biotechnology Journal**, 14:1086–1094. doi: *10.1111/pbi.12456*.

Wang, W.; Mauleon, R.; Hu Z. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. **Nature**, 557:43-52. *https://doi.org/10.1038/s41586-018-0063-9*.

Watt, C.; Zhou, G.; Li, C. 2020. Harnessing transcription factors as potential tools to enhance grain size under stressful abiotic conditions in cereal crops. **Frontiers in Plant Sciences**, 11:1273. doi: *10.3389/fpls.2020.01273*.

Zhang, W.; Peng, K.; Cui, F.; Wang, D.; Zhao, J.; Zhang, Y.; Yu, N.; Wang, Y.; Zeng, D.; Wang, Y.; Cheng, Z.; Zhang, K. 2021. Cytokinin oxidase/dehydrogenase *OsCKX11* coordinates source and sink relationship in rice by simultaneous regulation of leaf senescence and grain number. **Plant Biotechnology Journal**, 19:335–350. doi: *10.1111/pbi.13467*.

Zhang, X.; Yu, X.; Zhang, H.; Song, F. 2009. Molecular characterization of a defense-related AMP-binding protein gene, *OsBIABP1*, from rice. **Journal of Zhejiang University SCIENCE B**, 10(10):731-739. doi: *10.1631/jzus.B0920042*.