

Development of SNP markers for grain yield screening of Brazilian rice cultivars

Abstract – The objective of this work was to identify and validate single-nucleotide polymorphism (SNP) markers related to grain yield in rice (*Oryza sativa*) core collection. The genome-wide association studies (GWAS) methodology was applied for genotyping of 541 rice accessions by 167,470 SNPs. The grain yield of these accessions was estimated through the joint analysis of nine field experiments carried out in six Brazilian states. Fifteen SNPs were significantly associated with grain yield, and out of the ten SNPs converted to TaqMan assays, four discriminated the most productive accessions. These markers were used for the screening of rice accessions with favorable alleles. The selected accessions were, then, evaluated in field experiments in target environments, in order to select the most productive ones. This screening reduces the number of accessions evaluated experimentally, making it possible to prioritize those with higher productive potential, which allows of the increase of the number of replicates and, consequently, of the experimental accuracy.

Index terms: *Oryza sativa*, GWAS, marker-assisted selection, productivity.

Desenvolvimento de marcadores SNP para triagem de produtividade de grãos de cultivares brasileiras de arroz

Resumo – O objetivo deste trabalho foi identificar e validar marcadores de polimorfismo de nucleotídeo único (SNP) relacionados à produtividade de grãos em coleção nuclear de arroz (*Oryza sativa*). A metodologia de estudos de associação genômica ampla (GWAS) foi aplicada à genotipagem de 541 genótipos por 167.470 SNPs. A produtividade de grãos desses acessos foi estimada por meio da análise conjunta de nove experimentos de campo, realizados em seis estados brasileiros. Quinze SNPs foram significativamente associados à produtividade de grãos e, dos dez SNPs que foram convertidos em ensaios TaqMan, quatro discriminaram os acessos com maior produtividade. Esses marcadores foram utilizados para identificar acessos de arroz com os alelos favoráveis. Em seguida, os acessos selecionados foram avaliados em experimentos de campo, em ambientes-alvo, para identificar os mais produtivos. Essa triagem reduz o número de acessos avaliados experimentalmente, pois torna possível priorizar aqueles com maior potencial produtivo, o que permite aumentar o número de repetições e, consequentemente, a precisão experimental.

Termos para indexação: *Oryza sativa*, GWAS, seleção assistida por marcadores, produtividade.

Gabriel Feresin Pantalão⁽¹⁾ ,
Rosana Pereira Vianello⁽¹⁾ ,
Luíce Gomes Bueno⁽³⁾ ,
João Antônio Mendonça⁽¹⁾ ,
Alexandre Siqueira Guedes Coelho⁽²⁾ ,
Antônio Carlos Centeno Cordeiro⁽⁴⁾ ,
Paula Arielle Valdisser⁽¹⁾ ,
Ariadna Faria Vieira⁽¹⁾  and
Claudio Brondani⁽¹⁾ 

⁽¹⁾Embrapa Arroz e Feijão, Rodovia GO-462, Km 12, CEP 75375-000 Santo Antônio de Goiás, GO, Brazil. E-mail: gabrielferesin@hotmail.com, rosana.vianello@embrapa.br, joao.mendonca@embrapa.br, paula.valdisser@embrapa.br, aryadnafv@hotmail.com, claudio.brondani@embrapa.br

⁽²⁾Universidade Federal de Goiás, Escola de Agronomia, Rodovia GO-462, Km 0, CEP 74690-900 Goiânia, GO, Brazil. E-mail: asgcoelho@hotmail.com

⁽³⁾Embrapa Caprinos e Ovinos, Estrada Sobral-Groaíras, Km 4, CEP 62010-970 Sobral, CE, Brazil. E-mail: luice.bueno@embrapa.br

⁽⁴⁾Embrapa Roraima, Rodovia BR-174, Km 8, CEP 69301-970 Boa Vista, RR, Brazil. E-mail: antonio.cordeiro@embrapa.br

✉ Corresponding author

Received
September 18, 2019

Accepted
March 12, 2020

How to cite
PANTALÃO, G.F.; VIANELLO, R.P.; BUENO, L.G.; MENDONÇA, J.A.; COELHO, A.S.G.; CORDEIRO, A.C.C.; VALDISSER, P.A.; VIEIRA, A.F.; BRONDANI, C. Development of SNP markers for grain yield screening of Brazilian rice cultivars. *Pesquisa Agropecuária Brasileira*, v.55, e01643, 2020. DOI: <https://doi.org/10.1590/S1678-3921.pab2020.v55.01643>.

Introduction

There is a consensus on the need to increase food production, to meet the demand of the growing world population. A practical way to achieve this goal is to use molecular markers to increase the efficiency in the development of lines with a higher grain yield (Gupta et al., 2019). Single-nucleotide polymorphism (SNP) markers have been used in plant genomics due to their wide genome distribution and low cost for scoring (Voss-Fels et al., 2019). SNP markers can be used to genotype rice by medium (6K) or high-density chips (60K) (Tao et al., 2019), sequencing (Elshire et al., 2011), and resequencing (Barabaschi et al., 2016). The combination of high-throughput DNA sequencing and phenotyping of rice (*Oryza sativa* L.) cultivars can identify genes related to quantitative traits, through quantitative trait loci (QTL) analysis and genome-wide association studies (GWAS) (Mochida et al., 2018). Huang et al. (2010) and Lu et al. (2015) used the GWAS analysis to identify hundreds of SNP markers related to important agronomic traits in rice. Currently, the difficulty is not only to identify potentially useful molecular markers, but also to convert them into a useful and reproducible tool for breeding programs (Platten et al., 2019).

The greatest impact of the utility of marker-assisted selection occurred on the incorporation of resistance to rice blast disease, for which out of the more than 100 resistance genes identified, 28 have been cloned and validated. The pyramiding of these genes has led to the development of cultivars of longer resistance to rice blast (Wu et al., 2019). However, for traits with more complex inheritance, such as grain yield, the correlation between a particular marker and QTL is weak, hindering the improvement of the genotype of interest (Cobb et al., 2019). Even though it was identified and validated, this marker-QTL correlation has been ineffective beyond the original work population in different environments and years of experimentation. Begum et al. (2015), through the GWAS analysis of a rice breeding population, found three SNP markers related to a high-yield haplotype that explained approximately 9% of the phenotypic variation, while Zhang et al. (2017) described a SNP that showed high heritability for productivity. According to Cobb et al. (2019), a QTL may exhibit a high percentage of the phenotypic variation explained, but may have little biological significance. Moreover, the values of the

markers associated with productivity are individually small and, as a whole, they are not responsible for most of the phenotypic variation. This makes marker-assisted selection difficult, as genotyping will not be a good predictor for grain yield. Platten et al. (2019) observed that markers associated with characters of interest are rarely validated, that is, the results generally does not meet the marker-assisted breeding routine.

Therefore, it would be valuable to have an alternative strategy that could be effectively incorporated into the routine of assisted selection, in order to increase the chance of success of this process, beyond the identification of poorly reproducible associations between favorable markers and phenotypes in a given study population.

The objective of this work was to identify and validate SNP markers related to grain yield in rice core collection.

Materials and Methods

The GWAS panel consisted of the 550 accessions of Embrapa's rice core collection (Abadie et al., 2005), and included 94 Brazilian lines and cultivars (57 upland and 37 lowland accessions), 148 international lines and cultivars (76 upland and 72 lowland accessions), and 308 Brazilian landraces (148 upland and 77 lowland accessions, and 83 accessions for both cropping systems). The 550 rice accessions and four checks ('BR/IRGA 409', 'Metica 1', 'BRS Caiapó', and 'BRS Colosso') were evaluated in nine experiments, in six Brazilian states (Table 1). Federer's augmented-block experimental design was carried out with 23 blocks. The plot size consisted of three rows of 4 m, with a density of 100 seeds per meter. The statistical analysis of grain yield data (kg ha^{-1}) was performed using the lme4 package of the R platform (R Core Team, 2018). In the joint analysis procedure, the random effects and the experimental error were considered for blocks and genotypes (except for checks). The estimates of variance components were obtained by the residual maximum likelihood (REML). The estimates for the genetic values of grain yield of each accession were performed by BLUPS (best linear unbiased prediction).

The genomic DNA of the 550 accessions was obtained from young leaves using the DNeasy 96 Plant Kit (Qiagen, Germantown, MD, USA). The

SNP markers were obtained by GBS (genotyping by sequencing), a method proposed by Elshire et al. (2011). To determine the marker quality, the parameters “reproducibility” (percentage of technical replicate pairs scoring identically for given marker) and “call-rate” (percentage of samples for which a marker was scored) were used. For the genetic analysis only the SNPs showing 0.01 as minor allele frequency (MAF) set, 0.9 inbreeding coefficient, and 0.1 minimum locus coverage were considered. Data input was performed by FastPHASE 1.3 software (Scheet & Stephens, 2006). The input accuracy was estimated by the concordance rate (proportion of correctly input genotypes), in which 10% of the genotypes were randomly masked, followed by input and comparison with the true results. Population structure was estimated using the Bayesian model of the Markov chain Monte Carlo (MCMC) implemented in the Structure program (Pritchard et al., 2000). Five iterations were performed for each number of populations (k) tested from 1 to 10. Burn-in value and number of replicates of MCMC were set at 50,000 and 100,000, respectively. The K value was determined by the data log likelihood [LNP (D)] and delta K, based on the change rate of [LNP (D)] between successive values of k. These analyses were performed using the Structure Harvester program (Earl & VonHoldt, 2011). The structuring data and the relationship matrix (K matrix or kinship) were obtained by the Tassel 4.0 software. From the data on population structure and kinship matrix, GWAS analysis was performed based on the mixed model method, correcting spurious associations that could

occur due to the genetic similarity between accessions. The SNP markers identified as significantly associated with rice yield, and the structuring data were considered as factors of fixed effect, while the kinship matrix was considered as a factor of random effect. For better analysis reliability, rare alleles were removed, by filtering the SNP data input with 0.05 minimum allele frequency (MAF).

The GWAS analysis used the GAPIT package (Lipka et al., 2012); the stepwise regression analysis was performed in the GCTA software (Yang et al., 2011), and the removal of the SNPs with overlapping effect was obtained by the R software, according to the methodology described in Pantalhão et al. (2016). The selected SNPs were positioned in haplotypic blocks using the software Haploview (Barrett et al., 2005), which allowed of the identification of the candidate genes that cosegregated with the SNPs identified by GWAS as associated to the grain yield trait. Subsequently, the transcribed sequences of these genes were obtained to search for their putative functions in the Rice Genome Annotation Project (Kawahara et al., 2013).

A subset of grain-yield associated SNPs were selected for validation, using the TaqMan (Thermo Fischer Scientific, San Diego, CA, USA) probe-based chemistry designed for genotyping. The target SNPs were aligned with the reference genome (Os-Nipponbare-Reference-IRGSP-1.0-release 7), and a flanking region within 250 bp up- and downstream from the target SNPs was selected (Woodward, 2014). Before the design of allele-specific molecular markers, a DNA fragment

Table 1. Location, coordinates, cropping system, and yield of breeding rice (*Oryza sativa*) accessions from Coleção Nuclear de Arroz da Embrapa (CNAE) in field experiments.

	Location ⁽¹⁾	Coordinates			Altitude (m)	Cropping system	Yield (kg ha ⁻¹)	Yield range (kg ha ⁻¹)
		S	N	W				
1	Santo Antônio de Goiás, GO	16°28'	-	49°17'	779	Upland	1,383.3	3,030.0–270.0
2	Sinop, MT	11°51'	-	55°30'	345	Upland	2,290.9	2,525.3–2,081.4
3	Teresina, PI	5°05'	-	42°48'	72	Upland	3,203.8	5,645.9–2,057.7
4	Vilhena, RO	12°47'	-	60°05'	600	Upland	1,108.4	1,794.9–750.5
5	Goianira 1, GO	16°26'	-	49°23'	728	Lowland	4,117.8	5,777.2–2,756.1
6	Goianira 2, GO	16°26'	-	49°23'	728	Lowland	4,258.8	5,261.2–3,114.3
7	Boa Vista, RR	-	2°48'	60°39'	61	Lowland	5,080.7	7,664.5–3,307.8
8	Uruguaiana, RS	29°45'	-	57°05'	74	Lowland	4,032.7	4,220.8–3,902.8
9	Pelotas, RS	31°52'	-	52°21'	13	Lowland	6,272.4	9,178.9–4,438.5

⁽¹⁾Brazilian States: GO, Goiás; MT, Mato Grosso; PI, Piauí; RO, Rondônia; RR, Roraima; RS, Rio Grande do Sul.

of 501 bp length was evaluated for the presence of repetitive sequences, using the program RepeatMask (Smith et al., 2019), and for the presence of nontarget SNPs, using the SNPseek program (Mansueto et al., 2017) derived from the 3,000 Rice Genomes Project (Alexandrov et al., 2015). Those sequences without

repetitive elements and containing only the target SNPs were used for primer design. Two sets of plant material were used for the SNP validation analysis, as follows: 27 inbred lines from Value of Cultivation and Use experiments (17 upland and 10 lowland rice inbred lines), from Embrapa's rice breeding program

Table 2. Rice (*Oryza sativa*) inbred lines and checkers of Embrapa's value of cultivation and use (VCU) experiments genotyped by 10 TaqMan assays.

Genotypes	1 (A/G)	2 (G/T)	3 (C/T)	4 (C/G)	5 (G/C)	6 (G/A)	7 (G/T)	8 (C/T)	11 (A/G)	15 (AG)
Genotypes for upland cropping system										
BRS Esmeralda	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
BRS A501 CL	AA	GG	CC	CC	GG	GG	GG	CC	AA	AA
AN Cambará	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB112349	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB112345	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB112342	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB112305	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB162666	AA	GG	CC	GG	GG	GG	GG	CC	AA	AG
AB112315	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB112313	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB162664	AA	GG	CC	CC	GG	GG	GG	CC	AA	AG
AB142460	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB142538	AA	GG	CC	CC	GG	GG	TT	CC	AA	GG
AB142486	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB142485	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB142455	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB142481	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB142456	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB142472	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
AB142467	AA	GG	CC	CC	GG	GG	GG	CC	AA	GG
Genotypes for lowland cropping system										
IRGA 417	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
BR IRGA 409	AA	GG	CC	GG	GG	GG	TT	CC	AG	AA
BRS Pampeira	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
BRS Pampa	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
AB11502	AA	GG	CC	GG	GG	GG	TT	CC	AG	AA
AB13003	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
AB13006	AA	GG	CC	GG	GG	GG	GG	CC	AG	AA
AB13689	AA	GG	CC	GG	GG	GG	GG	CC	AA	AA
AB14727	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
AB14772	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
AB13715	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
AB14787	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
AB14764	AA	GG	CC	GG	GG	GG	TT	CC	AA	AA
AB14803	AA	GG	CC	GG	GG	GG	TT	CC	AG	AA

(Table 2), with average grain-yield data from 15 field experiments; and 20 high-yielding and 20 low-yielding rice cultivars from the joint analysis involving the nine experiments of the rice core collection (Table 3). PCR reactions, in duplicate, were carried out in a 5 μ L final volume of by the Custom TaqMan SNP Genotyping Assays 40X and TaqMan GTXpress master mix 2 X (Thermo Fisher Scientific, USA). The reactions were

Table 3. Genotyping of 20 high-yielding and 20 low-yielding rice (*Oryza sativa*) accessions from nine field experiments.

Genotypes	Code	Yield ⁽¹⁾ (kg ha ⁻¹)	1y (A/G)	2y (G/T)	3y (C/T)	5y (G/C)
High-yielding rice accessions						
Farroupilha	CA940008	4,524.20	AA	GG	CC	GG
Wir 5621	CNA0005853	4,474.18	AA	GG	CC	GG
Tapuripa 161	CNA0001423	4,179.41	AA	GG	CC	GG
Szu Maio	CNA0005478	4,097.16	AA	GG	CC	GG
Agulhinha	CA790332	4,052.58	AA	GG	CC	GG
Dawn	CNA0001006	4,000.79	AA	GG	CC	GG
Agulhinha	CA780366	3,944.83	AA	GG	CC	GG
Ipeaco-SL 1469	CNA0001344	3,940.11	AA	GG	TT	GG
Vitro	CNA0006961	3,937.36	AA	GG	CC	GG
CR 1113	CNA0004552	3,927.95	AA	GG	CC	GG
Caiana Grande	CA960008	3,863.36	AA	GG	CC	GG
Uba Laginha	CA780077	3,850.73	AA	GG	CC	GG
WU 10 B	CNA0005014	3,834.39	AA	GG	CC	GG
BR IRGA 413	CNA0008398	3,833.27	AA	GG	CC	GG
Bico Torto	CA940004	3,805.80	AA	GG	CC	GG
EEA 404	CNA0001108	3,802.60	AA	GG	CC	GG
Metica 1	CNA0004566	3,799.36	AA	GG	CC	GG
Pacholinha	CA780320	3,790.56	AA	GG	CC	GG
IR8	CNA0008416	3,773.26	AA	GG	CC	GG
Minami Hata Mochi	CNA0010432	3,771.48	AA	GG	CC	GG
Low-yielding rice accessions						
Cataguases	CA800117	3,076.99	AA	GG	TT	GG
Santa Catarina	CA780125	3,052.67	GG	TT	CC	CC
Vermelho	CA910003	3,052.24	AA	GG	CC	GG
Kataktara	CNA0010514	3,048.48	AA	GG	CC	GG
Paranazinho Palha	CA930003	3,048.47	AA	GG	CC	GG
Amarelo	CA850050	3,045.43	AA	GG	CC	GG
N.2583	CNA0004697	2,983.28	AA	GG	CC	GG
Branquinho	CA780391	2,980.86	AA	GG	CC	GG
Canarinho	CA960013	2,883.01	GG	TT	TT	CC
Lageado	CA840062	2,874.20	GG	TT	TT	CC
Bico preto	CA810011	2,859.52	GG	TT	TT	CC
Lageado	CA780392	2,830.15	GG	TT	TT	CC
Lageado	CA790255	2,827.47	GG	TT	TT	CC
Lageado	CA800169	2,812.40	GG	TT	TT	GG
Lageado	CA800183	2,812.00	GG	TT	TT	CC
Lageado	CA790011	2,792.66	GG	TT	TT	CC
Lageado	CA840029	2,773.44	GG	TT	TT	CC
Lageado	CA790110	2,765.72	GG	TT	TT	CC
Caianinha	CA960007	2,680.94	GG	TT	CC	CC
Maranhão	CA810036	2,660.93	GG	TT	TT	CC

⁽¹⁾Average yield (eBlups).

run in the QuantStudio 7 Flex Real-Time PCR System (Thermo Fisher Scientific, USA), with 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, finishing with 30 s at 60°C. The SNP genotyping was performed using the TaqMan Genotyper Software (Thermo Fisher Scientific, USA), and it was exported as a matrix of genotyping calls for each sample.

Results and Discussion

There was a great variation for grain yield in all nine field experiments, as expected because of the high diversity of the set of accessions of the core collection, and by the environmental variation (Table 1). The average grain yield of the experiments varied from 1,108.4 kg ha⁻¹ (Vilhena, RO) to 6,272.4 kg ha⁻¹ (Pelotas, RS). The genotyping of the 550 accessions provided 526,220 SNPs, distributed on the 12 chromosomes. After the data input, accessions showing more than 20% of missing data were excluded, which resulted in 445,589 SNPs from 541 accessions. When 0.05 minimum allele frequency was applied, the final number was 167,470 SNPs (Table 4). The average distribution was approximately 449 SNPs/Mbp (one SNP every 2.23 kbp), ranging from 366 SNPs/Mbp

on chromosome 5 to 507 SNPs/Mbp on chromosome 11. Chromosome 1 had the highest number of SNPs (21,662), while chromosome 9 had the lowest number of SNPs (9,788), with average 13,956 SNPs per chromosome. Considering the high-rice genome linkage disequilibrium, about 150 kbp, this average marker density is considered adequate to perform genome-wide association analyses (Rebolledo et al., 2015). The population structure analysis identified two groups of accessions (k=2), corresponding to the *Oryza sativa* subspecies indica and japonica. The GWAS was performed with 167,470 SNPs, and identified 31 SNP markers significantly associated with the grain yield trait. After the stepwise regression analysis, 15 SNPs remained in the model, with R² value p<0.001 without significant difference (0.496 and 0.490, for 31 and 15 SNPs, respectively). Out of the 15 SNPs, 9 were located in genes, 3 were located in genes present in linkage blocks, and 3 were located in intergenic regions (Table 5). The validation of the SNPs associated to characters identified by GWAS is an essential step to enable the effective use of these markers in breeding programs (Kikuchi et al., 2017). From the 15 SNPs maintained in the statistical model, 10 were used in the development of TaqMan probes. The remaining 5 SNPs showed a high percentage of repetitive DNA

Table 4. Number of single nucleotide polymorphism (SNP) markers obtained from genotyping by sequencing of rice (*Oryza sativa*) accessions (MAF <0.05).

Chromosome	Chromosome size (bp) ⁽¹⁾	MAF (<0.01)	Input	MAF (<0.05)	Kbp/SNP	SNPs/Mbp
1	43,270,923	68,224	58,447	21,662	2.00	500.61
2	35,937,250	52,953	45,349	17,052	2.11	474.49
3	36,413,819	54,931	47,088	15,990	2.28	439.12
4	35,502,694	49,341	41,609	15,442	2.30	434.95
5	29,958,434	39,227	33,493	10,963	2.73	365.94
6	31,248,787	44,076	37,485	14,857	2.10	475.44
7	29,697,621	40,509	34,486	12,291	2.42	413.87
8	28,443,022	39,156	32,750	12,783	2.23	449.42
9	23,012,720	31,642	26,711	9,788	2.35	425.33
10	23,207,287	32,961	27,711	11,631	2.00	501.18
11	29,021,106	41,306	34,138	14,716	1.97	507.08
12	27,531,856	31,894	26,322	10,295	2.67	373.93
Total	373,245,519	526,220	445,589	167,470	2.23	448.69 ⁽²⁾

⁽¹⁾Ox-Nipponbare-Reference-IRGSP-1.0-release 7. ⁽²⁾Average value. MAF: minimum allele frequency; kbp (kilobase pairs); Mbp (megabase pairs).

in the adjacent sequences, thus it was not possible to design Taqman assays (S1_33418191, S9_1062037, S10_2231343, S12_3544726, and S12_17681142). Considering the 10 SNPs loci that derived the TaqMan assays, 28 genes were identified (25 genes located in linkage blocks), 20 of which were putatively annotated and related to metabolic processes, such as responses to biotic and abiotic stresses, responses to endogenous and exogenous stimuli, post-embryonic multicellular development, growth, and morphogenesis (Table 6). The marker #7 is located in a gene (*OsWAK*, LOC_Os10g01410), which was previously related to grain yield and panicle number traits (Zhang et al., 2017). The SNPs located in genes are candidates to be explored by genetic engineering (Chen et al., 2018).

The rice inbred lines genotyped by TaqMan assays showed the following results: 7 markers were monomorphic; and 3 markers discriminated genotypes of lowland and upland cropping systems (S7_939762, #4; S10_251060, #7; S2_22142097, #15) (Tables 2 and 3). In Brazil, most of upland rice cultivars are of the japonica subspecies, and lowland cultivars

are of the indica subspecies (Khush, 1997); therefore, these three markers can be helpful to identify the materials of unknown rice cropping systems. The exclusive alleles for indica or japonica, may have originated during the independent domestication of these two subspecies (Civan & Brown, 2018; Wang et al., 2018). The markers #6 (S9_20925193), #8 (S4_21506318), and #11 (S9_7799399) were unable to discriminate the most productive accessions. The accessions with average grain yield above 2,981 kg ha⁻¹ were all discriminated by the markers #1 (S1_23079331), #2 (S2_26805540), #3 (S6_5353837), and #5 (S9_12051077) (Table 3). The exception was for the landrace 'Santa Catarina', whose average productivity was 3,052 kg ha⁻¹, but showed markers #1, #2, and #5, which are the SNP pattern of plants of lower productivity. A possible explanation is that these four loci were selected independently, during the domestication and genetic breeding of cultivated rice, as observed by Xie et al. (2015). According to the selective sweep model, a series of variants in the genome that lead to adaptation is rapidly fixed

Table 5. Single nucleotide polymorphism (SNPs) associated with grain yield rice (*Oryza sativa*) accessions after GWAS and stepwise regression analyses.

SNP	Chromosome	SNP alleles		SNP location ⁽¹⁾	Marker acronym	TaqMan ID
		A1	A2			
S1_23079331	1	A	G	LOC_Os01g40820	#1	AH1SFDM
S1_33418191	1	G	A	LOC_Os01g57780	-	-
S2_26805540	2	G	T	8 genes (LOC_Os02g44310)	#2	AH21DJU
S2_22142097	2	G	A	(intergenic)	#15	AN2W9E4
S4_21506318	4	C	T	LOC_Os04g35370	#8	ANYMKAA
S6_5353837	6	C	T	2 genes (intergenic)	#3	AH4ABP2
S7_939762	7	C	G	LOC_Os07g02610	#4	AH5I9WA
S9_12051077	9	G	C	6 genes (LOC_Os09g20110)	#5	AH6R72I
S9_1062037	9	T	A	LOC_Os09g02550	-	-
S9_7799399	9	A	G	LOC_Os09g13470	#11	ANZTEU7
S9_20925193	9	G	A	7 genes (intergenic)	#6	AH7058Q
S10_2231343	10	C	G	LOC_Os10g04674	-	-
S10_251060	10	G	T	LOC_Os10g01410	#7	AH894EY
S12_17681142	12	G	C	2 genes (LOC_Os12g29650)	-	-
S12_3544726	12	G	T	LOC_Os12g07210	-	-

⁽¹⁾Between parenthesis: SNP is located in a gene belonging to a linkage block.

in a population, and this creates a selection signature that consists in reducing genetic diversity, extending linkage disequilibrium in the genome region around the locus under selection (Gentzittel et al., 2019).

Due to the low heritability of the productivity trait, and the inconstancy of markers associated with characters of interest in different backgrounds (Sandhu et al., 2018), or when considering different environments (Kumar et al., 2018), an alternative strategy needs to be implemented to make these markers useful for breeding programs.

Our suggestion is to use the four markers to select rice accessions from a gene bank, and then evaluate these accessions in field experiments to identify the most productive ones. This screening would reduce the number of accessions that would be evaluated in the field, making it possible to prioritize efforts in those with greater grain-yield potential. The genotypes with the greatest adaptability to a given location would be selected, and, then, they are used as genitors for the development of inbred lines targeted at specific locations.

Table 6. Genes related to selected single nucleotide polymorphism (SNP) markers and their putative functions.

Marker	SNP position	ID	Block Size (bp)	Putative function
#1	S1_23079331	LOC_Os01g40820	-	Peptidase family M41 containing protein
#2	S2_26805540	LOC_Os02g44250	43,311	Unknown
		LOC_Os02g44260		Zinc-binding protein
		LOC_Os02g44270		Unknown
		LOC_Os02g44280		Zinc finger protein
		LOC_Os02g44290		Phosphatase
		LOC_Os02g44300		MSP domain containing protein
		LOC_Os02g44310		LTPL112 - Protease inhibitor/seed storage/LTP family protein precursor
	LOC_Os02g44320	LTPL113 - Protease inhibitor/seed storage/LTP family protein precursor		
#3	S6_5353837	LOC_Os06g10420	9,142	Nitrilase
		LOC_Os06g10430		Unknown
#5	S9_12051077	LOC_Os09g20110	26,272	Unknown
		LOC_Os09g20120		Transposon protein, putative, CACTA, En/Spm sub-class
		LOC_Os09g20130		Transposon protein, putative, CACTA, En/Spm sub-class
		LOC_Os09g20140		Unknown
		LOC_Os09g20150		C1-like domain containing protein
	LOC_Os09g20160	Transposon protein, putative, CACTA, En/Spm sub-class		
#4	S7_939762	LOC_Os07g02610	-	Unknown
#7	S10_251060	LOC_Os10g01410	-	OsWAK94 - OsWAK receptor-like protein OsWAK-RLP
#15	S2_22142097	-	-	-
#8	S4_21506318	LOC_Os04g35370		Broad complex BTB domain
#11	S9_7799399	LOC_Os09g13470		Unknown
#6	S9_20925193	LOC_Os09g36230	44,556	PPR repeat containing protein
		LOC_Os09g36240		deoxyribodipyrimidine photolyase family protein
		LOC_Os09g36244		Unknown
		LOC_Os09g36250		MYB family transcription factor
		LOC_Os09g36260		Unknown
		LOC_Os09g36270		Pantothenate kinase
	LOC_Os09g36280	Glycosyl hydrolases family 17		

Conclusions

1. The four SNP markers associated with grain yield, identified in this work, can select Brazilian cultivars of rice with a greater productive potential.

2. The accessions with average grain yield above 2,981 kg ha⁻¹ are all discriminated by the markers #1 (S1_23079331), #2 (S2_26805540), #3 (S6_5353837), and #5 (S9_12051077).

Acknowledgments

To Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), for financial support, and grants to second and ninth authors; to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), for grants to first author; and to Empresa Brasileira de Pesquisa Agropecuária (Embrapa), for financial support.

References

- ABADIE, T.; CORDEIRO, C.M.T.; FONSECA, J.R.; ALVES, R. de B. das N.; BURLE, M.L.; BRONDANI, C.; RANGEL, P.H.N.; CASTRO, E. de M. de; SILVA, H.T. da; FREIRE, M.S.; ZIMMERMANN, F.J.P.; MAGALHÃES, J.R. Construção de uma coleção nuclear de arroz para o Brasil. **Pesquisa Agropecuária Brasileira**, v.40, p.129-136, 2005. DOI: <https://doi.org/10.1590/S0100-204X2005000200005>.
- ALEXANDROV, N.; TAI, S.; WANG, W.; MANSUETO, L.; PALIS, K.; FUENTES, R.R.; ULAT, V.J.; CHEBOTAROV, D.; ZHANG, G.; LI, Z.; MAULEON, R.; HAMILTON, R.S.; MCNALLY, K.L. SNP-seek database of SNPs derived from 3000 rice genomes. **Nucleic Acids Research**, v.43, p.D1023-D1027, 2015. DOI: <https://doi.org/10.1093/nar/gku1039>.
- BARABASCHI, D.; TONDELLI, A.; DESIDERIO, F.; VOLANTE, A.; VACCINO, P.; VALÈ, G.; CATTIVELLI, L. Next generation breeding. **Plant Science**, v.242, p.3-13, 2016. DOI: <https://doi.org/10.1016/j.plantsci.2015.07.010>.
- BARRETT, J.C.; FRY, B.; MALLER, J.; DALY, M.J. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics**, v.21, p.263-265, 2005. DOI: <https://doi.org/10.1093/bioinformatics/bth457>.
- BEGUM, H.; SPINDEL, J.E.; LALUSIN, A.; BORROMEO, T.; GREGORIO, G.; HERNANDEZ, J.; VIRK, P.; COLLARD, B.; MCCOUCH, S.R. Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). **PLoS ONE**, v.10, e0119873, 2015. DOI: <https://doi.org/10.1371/journal.pone.0119873>.
- CHEN, K.; LUAN, M.; XIONG, H.; XIONG, H.; CHEN, P.; CHEN, J.; GAO, G.; HUANG, K.; ZHU, A.; YU, C. Genome-wide association study discovered favorable single nucleotide polymorphisms and candidate genes associated with ramet number in ramie (*Boehmeria nivea* L.). **BMC Plant Biology**, v.18, art.345, 2018. DOI: <https://doi.org/10.1186/s12870-018-1573-1>.
- CIVAN, P.; BROWN, T.A. Role of genetic introgression during the evolution of cultivated rice (*Oryza sativa* L.). **BMC Evolutionary Biology**, v.18, art.57, 2018. DOI: <https://doi.org/10.1186/s12862-018-1180-7>.
- COBB, J.N.; BISWAS, P.S.; PLATTEN, J.D. Back to the future: revisiting MAS as a tool for modern plant breeding. **Theoretical and Applied Genetics**, v.132, p.647-667, 2019. DOI: <https://doi.org/10.1007/s00122-018-3266-4>.
- EARL, D.A.; VONHOLDT, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conservation Genetics Resources**, v.4, p.359-361, 2011. DOI: <https://doi.org/10.1007/s12686-011-9548-7>.
- ELSHIRE, R.J.; GLAUBITZ, J.C.; SUN, Q.; POLAND, J.A.; KAWAMOTO, K.; BUCKLER, E.S.; MITCHELL, S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS ONE**, v.6, e19379, 2011. DOI: <https://doi.org/10.1371/journal.pone.0019379>.
- GENTZBITTEL, L.; BEN, C.; MAZURIER, M.; SHIN, M.-G.; LORENZ, T.; RICKAUER, M.; MARJORAM, P.; NUZHIDIN, S.V.; TATARINOVA, T.V. WhoGEM: an admixture-based prediction machine accurately predicts quantitative functional traits in plants. **Genome Biology**, v.20, art.106, 2019. DOI: <https://doi.org/10.1186/s13059-019-1697-0>.
- GUPTA, P.K.; KULWAL, P.L.; JAISWAL, V. Association mapping in plants in the post-GWAS genomics era. **Advances in Genetics**, v.104, p.75-154, 2019. DOI: <https://doi.org/10.1016/bs.adgen.2018.12.001>.
- HUANG, X.; WEI, X.; SANG, T.; ZHAO, Q.; FENG, Q.; ZHAO, Y.; LI, C.; ZHU, C.; LU, T.; ZHANG, Z.; LI, M.; FAN, D.; GUO, Y.; WANG, A.; WANG, L.; DENG, L.; LI, W.; LU, Y.; WENG, Q.; LIU, K.; HUANG, T.; ZHOU, T.; JING, Y.; LI, W.; LIN, Z.; BUCKLER, E.S.; QIAN, Q.; ZHANG, Q.-F.; LI, J.; HAN, B. Genome-wide association studies of 14 agronomic traits in rice landraces. **Nature Genetics**, v.42, p.961-967, 2010. DOI: <https://doi.org/10.1038/ng.695>.
- KAWAHARA, Y.; BASTIDE, M. de la; HAMILTON, J.P.; KANAMORI, H.; MCCOMBIE, W.R.; OUYANG, S.; SCHWARTZ, D.C.; TANAKA, T.; WU, J.; ZHOU, S.; CHILDS, K.L.; DAVIDSON, R.M.; LIN, H.; QUESADA-OCAMPO, L.; VAILLANCOURT, B.; SAKAI, H.; LEE, S.S.; KIM, J.; NUMA, H.; ITOH, T.; BUELL, C.R.; MATSUMOTO, T. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. **Rice**, v.6, art.4, 2013. DOI: <https://doi.org/10.1186/1939-8433-6-4>.
- KHUSH, G.S. Origin, dispersal, cultivation and variation of rice. **Plant Molecular Biology**, v.35, p.25-34, 1997. DOI: <https://doi.org/10.1023/A:1005810616885>.
- KIKUCHI, S.; BHEEMANAHALLI, R.; JAGADISH, K.S.V.; KUMAGAI, E.; MASUYA, Y.; KURODA, E.; RAGHAVAN, C.; DINGKUHN, M.; ABE, A.; SHIMONO, H. Genome-wide association mapping for phenotypic plasticity in rice.

- Plant, Cell and Environment**, v.40, p.1565-1575, 2017. DOI: <https://doi.org/10.1111/pce.12955>.
- KUMAR, A.; SANDHU, N.; DIXIT, S.; YADAV, S.; SWAMY, B.P.M.; SHAMSUDIN, N.A.A. Marker-assisted selection strategy to pyramid two or more QTLs for quantitative trait-grain yield under drought. **Rice**, v.11, art.35, 2018. DOI: <https://doi.org/10.1186/s12284-018-0227-0>.
- LIPKA, A.E.; TIAN, F.; WANG, Q.; PEIFFER, J.; LI, M.; BRADBURY, P.J.; GORE, M.A.; BUCKLER, E.S.; ZHANG, Z. GAPIT: genome association and prediction integrated tool. **Bioinformatics**, v.28, p.2397-2399, 2012. DOI: <https://doi.org/10.1093/bioinformatics/bts444>.
- LU, Q.; ZHANG, M.; NIU, X.; WANG, S.; XU, Q.; FENG, Y.; WANG, C.; DENG, H.; YUAN, X.; YU, H.; WANG, Y.; WEI, X. Genetic variation and association mapping for 12 agronomic traits in *indica* rice. **BMC Genomics**, v.16, art.1067, 2015. DOI: <https://doi.org/10.1186/s12864-015-2245-2>.
- MANSUETO, L.; FUENTES, R.R.; BORJA, F.N.; DETRAS, J.; ABRIOL-SANTOS, J.M.; CHEBOTAROV, D.; SANCIANGCO, M.; PALIS, K.; COPETTI, D.; POLIAKOV, A.; DUBCHAK, I.; SOLOVYEV, V.; WING, R.A.; HAMILTON, R.S.; MAULEON, R.; MCNALLY, K.L.; ALEXANDROV, N. Rice SNP-seek database update: new SNPs, indels, and queries. **Nucleic Acids Research**, v.45, p.D1075-D1081, 2017. DOI: <https://doi.org/10.1093/nar/gkw1135>.
- MOCHIDA, K.; KODA, S.; INOUE, K.; HIRAYAMA, T.; TANAKA, S.; NISHII, R.; MELGANI, F. Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. **GigaScience**, v.8, p.1-12, 2018. DOI: <https://doi.org/10.1093/gigascience/giy153>.
- PANTALIÃO, G.F.; NARCISO, M.; GUIMARÃES, C.; CASTRO, A.; COLOMBARI, J.M.; BRESEGHELLO, F.; RODRIGUES, L.; VIANELLO, R.P.; BORBA, T.O.; BRONDANI, C. Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. **Genetica**, v.144, p.651-664, 2016. DOI: <https://doi.org/10.1007/s10709-016-9932-z>.
- PLATTEN, J.D.; COBB, J.N.; ZANTUA, R.E. Criteria for evaluating molecular markers: comprehensive quality metrics to improve marker-assisted selection. **PLoS ONE**, v.14, e0210529, 2019. DOI: <https://doi.org/10.1371/journal.pone.0210529>.
- PRITCHARD, J.K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v.155, p.945-959, 2000.
- R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018. Available at: <https://www.R-project.org/>. Accessed on: Dec. 15 2019.
- REBOLLEDO, M.C.; DINGKUHN, M.; COURTOIS, B.; GIBON, Y.; CLÉMENT-VIDAL, A.; CRUZ, D.F.; DUITAMA, J.; LORIEUX, M.; LUQUET, D. Phenotypic and genetic dissection of component traits for early vigour in rice using plant growth modelling, sugar content analyses and association mapping. **Journal of Experimental Botany**, v.66, p.5555-5566, 2015. DOI: <https://doi.org/10.1093/jxb/erv258>.
- SANDHU, N.; DIXIT, S.; SWAMY, B.P.M.; VIKRAM, P.; VENKATESHWARLU, C.; CATOLOS, M.; KUMAR, A. Positive interactions of major-effect QTLs with genetic background that enhances rice yield under drought. **Scientific Reports**, v.8, art.1626, 2018. DOI: <https://doi.org/10.1038/s41598-018-20116-7>.
- SCHEET, P.; STEPHENS, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. **American Journal of Human Genetics**, v.78, p.629-644, 2006. DOI: <https://doi.org/10.1086/502802>.
- SMITH, A.F.A.; HUBLEY, R.; GREEN, P. **RepeatMasker Web Server**. Available at: <http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker>. Accessed on: Dec. 15 2019.
- TAO, Y.; ZHAO, X.; MACE, E.; HENRY, R.; JORDAN, D. Exploring and exploiting pan-genomics for crop improvement. **Molecular Plant**, v.12, p.156-169, 2019. DOI: <https://doi.org/10.1016/j.molp.2018.12.016>.
- VOSS-FELS, K.P.; STAH, A.; HICKEY, L.T. Q&A: modern crop breeding for future food security. **BMC Biology**, v.17, art.18, 2019. DOI: <https://doi.org/10.1186/s12915-019-0638-4>.
- WANG, W.; MAULEON, R.; HU, Z.; CHEBOTAROV, D.; TAI, S.; WU, Z.; LI, M.; ZHENG, T.; FUENTES, R.R.; ZHANG, F.; MANSUETO, L.; COPETTI, D.; SANCIANGCO, M.; PALIS, K.C.; XU, J.; SUN, C.; FU, B.; ZHANG, H.; GAO, Y.; ZHAO, X.; SHEN, F.; CUI, X.; YU, H.; LI, Z.; HEN, M.; DETRAS, J.; ZHOU, Y.; ZHANG, X.; ZHAO, Y.; KUDRNA, D.; WANG, C.; LI, R.; JIA, B.; LU, J.; HE, X.; DONG, Z.; XU, J.; LI, Y.; WANG, M.; SHI, J.; LI, J.; ZHANG, D.; LEE, S.; HU, W.; POLIAKOV, A.; DUBCHAK, I.; ULAT, V.J.; BORJA, F.N.; MENDOZA, J.R.; ALI, J.; LI, J.; GAO, Q.; NIU, Y.; YUE, Z.; NAREDO, M.E.B.; TALAG, J.; WANG, X.; LI, J.; FANG, X.; YIN, Y.; GLASZMANN, J.-C.; ZHANG, J.; LI, J.; HAMILTON, R.S.; WING, R.A.; RUAN, J.; GENGYUN ZHANG, G.; CHAOCHUN WEI, C.; ALEXANDROV, N.; MCNALLY, K.L.; LI, Z.; LEUNG, H. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. **Nature**, v.557, p.43-52, 2018. DOI: <https://doi.org/10.1038/s41586-018-0063-9>.
- WOODWARD, J. Bi-allelic SNP genotyping using the TaqMan® assay. In: FLEURY, D.; WHITFORD, R. (Ed.). **Crop Breeding: methods and protocols**. New York: Springer, 2014. p.67-74. (Methods in Molecular Biology, 1145). DOI: https://doi.org/10.1007/978-1-4939-0446-4_6.
- WU, Y.; XIAO, N.; CHEN, Y.; YU, L.; PAN, C.; LI, Y.; ZHANG, X.; HUANG, N.; JI, H.; DAI, Z.; CHEN, X.; LI, A. Comprehensive evaluation of resistance effects of pyramiding lines with different broad-spectrum resistance genes against *Magnaporthe oryzae* in rice (*Oryza sativa* L.). **Rice**, v.12, art.11, 2019. DOI: <https://doi.org/10.1186/s12284-019-0264-3>.
- XIE, W.; WANG, G.; YUAN, M.; YAO, W.; LYU, K.; ZHAO, H.; YANG, M.; LI, P.; ZHANG, X.; YUAN, J.; WANG, Q.; LIU, F.; DONG, H.; ZHANG, L.; LI, X.; MENG, X.; ZHANG, W.; XIONG, L.; HE, Y.; WANG, S.; YU, S.; XU, C.; LUO, J.; LI, X.; XIAO, J.; LIAN, X.; ZHANG, Q. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. **Proceedings of the National Academy of**

Sciences of the United States of America, v.112, p.E5411-E5419, 2015. DOI: <https://doi.org/10.1073/pnas.1515919112>.

YANG, J.; LEE, S.H.; GODDARD, M.E.; VISSCHER, P.M. GCTA: a tool for genome-wide complex trait analysis. **The American Journal of Human Genetics**, v.88, p.76-82, 2011. DOI: <https://doi.org/10.1016/j.ajhg.2010.11.011>.

ZHANG, Y.; ZHOU, L.; SHEN, X.; CHEN, D.; WU, W.; ZHAN, X.; LIU, Q.; ZHU, A.; LOU, X.; XU, H.; CHENG, S.; CAO, L. Genetic dissection of yield traits in super hybrid rice Xieyou9308 using both unconditional and conditional genome-wide association mapping. **Scientific Reports**, v.7, art.824, 2017. DOI: <https://doi.org/10.1038/s41598-017-00938-7>.
