Development and validation of an optimized marker set for genomic selection in Southern U. S. rice breeding programs

Tommaso Cerioli[1], Christopher O. Hernandez[1,2], Brijesh Angira[1], Susan R. McCouch[2,3], Kelly R. Robbins[2], Adam N. Famoso[1]

Affiliations:

[1]Louisiana State University Agricultural Center, H. Rouse Caffey Rice Research Station, Rayne, LA 70578, USA

[2]Section of Plant Breeding and Genetics, School of Integrative Plant Sciences, Cornell University, Ithaca, NY 14850, USA

[3]Cornell Institute for Digital Agriculture, Cornell University, Ithaca, NY 14850, USA

Abbreviations:

AmpSeq = Amplicon Sequencing
BLUE = Best Linear Unbiased Estimator
BLUP = Best Linear Unbiased Predictor
CV = Cross-validation
DP = Diversity panel
GBS = Genotyping-by-sequencing
GEBV = Genomic Estimated Breeding Value
GS = Genomic selection
HRCRRS = H. Rouse Caffey Rice Research Station
LD = Linkage disequilibrium
MAF = Minor Allele Frequency
MAS = Marker Assisted Selection
PYT = Preliminary Yield Trial
RGS = Representative Germplasm Set

ABSTRACT

The potential of genomic selection (GS) to increase the efficiency of breeding programs has been clearly demonstrated; however, the implementation of GS in rice (*Oryza sativa* L.) breeding programs has been limited. In recent years, we have begun to work towards implementing GS into the LSU AgCenter rice breeding program. One of the first steps for successful GS

implementation is to establish a suitable marker set for the target germplasm and a reliable, cost-effective genotyping platform capable of providing informative marker data with an adequate turnaround time. In this study, we develop an optimized a marker set, the LSU500, for application of routine GS in Southern U.S. rice germplasm. The utility of the LSU500 was demonstrated using four years of breeding data across 8,473 experimental lines and four elite bi-parental populations. The predictive ability of GS ranged from 0.13 to 0.78 for key traits across different market classes and yield trials. Comparisons between phenotypic selection and GS within bi-parental populations using the LSU500 provided evidence of the potential of GS to improve the efficiency of a rice breeding program. The design of this marker set followed a continuous integration strategy, whereby GS is initially introduced into a breeding program while technical and strategic aspects of GS implementation are evaluated, optimized, and integrated into the breeding pipeline on-the-go. The LSU500 marker set has been established through the genotyping service provider Agriplex Genomics, and in the future, it will undergo improvements to reduce the cost and increase the accuracy of GS.

INTRODUCTION

DNA marker information is widely used for implementing marker-assisted selection (MAS) to improve the efficiency and precision of conventional plant breeding programs in different crops, including rice (*Oryza sativa*) (Collard & Mackill, 2008). The utility of MAS is greatest when the target of selection involves a single gene or locus of large phenotypic effect, and is often used to introgress useful traits into breeding populations (Cobb et al., 2019). However, many important traits in a plant breeding program, including yield and quality, are inherited in a highly quantitative manner and are determined by many genes of small effect. Thus, MAS is not an effective method for selection of these traits. Highly polygenic traits are often difficult to select because they are affected by the environment and have lower heritability making phenotypic selection in the early breeding stages inefficient.

The introduction of genomic selection (GS), which uses genome-wide marker information to predict breeding values, has enabled prediction of quantitative traits within breeding populations (Bernardo, 1994; Meuwissen et al., 2001). Genomic selection leverages information from a training population that is both genotyped and phenotyped to train a statistical model that predicts the performance of related individuals referred to as the prediction set. The prediction set is genotyped but not phenotyped, and the genomic estimated breeding value (GEBV) is calculated using genotypic information. The implementation of GS can increase the genetic gain of a breeding program by improving the accuracy of selection, increasing selection intensity, reducing breeding cycle time, and also decreasing the cost of the breeding program operations (Crossa et al., 2017; Heffner et al., 2010; Hickey et al., 2017; Voss-Fels et al., 2019).

Despite the demonstrated potential of GS to improve the efficiency of rice breeding programs (Grenier et al., 2015; Monteverde et al., 2018; Spindel et al., 2015; Xu et al., 2021), the implementation of GS in applied breeding programs has been limited. This can be attributed to changes in required skill sets and processes within the breeding program that are necessary for routine applications of GS. Among these are the requirement to adopt a data management system for storing and accessing phenotypic and genotypic data, routine use of data analysis and breeding decision-support tools, optimization of the breeding pipeline, and employment of people with new skill-sets within the traditional breeding program (Santantonio et al., 2020).

One of the first steps for successful GS implementation is to establish a suitable marker set for the target germplasm and a reliable, cost-effective genotyping platform capable of providing informative marker data with fast turnaround time and low cost. The optimal marker density for GS will depend on the extent of linkage disequilibrium (LD) across the genome in the target germplasm (Heffner et al., 2009), the number of independent chromosomal segments that need to be tracked in the target population (Daetwyler et al., 2010), and the heritability of the trait (Goddard & Hayes, 2007). In an applied and/or commercial plant breeding program, the effective population size is often small, leading to strong genome-wide LD (Flint-Garcia et al., 2003); therefore, a relatively small number of selected markers is capable of tagging each independent LD block and can be used to predict breeding values with good accuracy. This has been demonstrated in wheat (Juliana et al., 2019), barley (Abed et al., 2018), and rapeseed (Werner et al., 2018), where a few hundred to a few thousand markers enable high GS prediction accuracies that are comparable to high marker densities. In addition, breeding programs often breed for different market classes or heterotic groups, creating distinct population structures within the breeding population. Taking advantage of this structure in defining the training

populations helps to maintain high LD within the segments and consequently reduce the number of markers required for GS.

Rice is a model crop species, and numerous marker sets are available with varying marker densities; examples include the 1k-RiCA (Arbelaez et al., 2019), the C7AIR (7K) (Morales et al., 2020), the 44K-SNP chip (Zhao et al., 2011), and the 700K-SNP High Density Rice Array (McCouch et al., 2016). While these arrays represent a rich resource of validated SNPs for genome-wide association analysis and genetic diversity analysis, they are not ideal for breeding applications where the genetic diversity of the material is relatively narrow. The cost per sample and inflexibility of these arrays hinders their use for routine genotyping for GS where the majority of SNPs tend to be monomorphic in the breeding germplasm. However, these established marker sets represent invaluable resources for identifying tailored subsets of reliable, high-quality SNPs known to segregate in the target breeding germplasm.

Sequence-based genotyping approaches, including genotyping-by-sequencing (GBS), have also been used for breeding applications (Poland & Rife, 2012); however, their application requires bioinformatic skills to process the raw data and use the genotypic information, adding an additional layer of complexity for a breeding program.

In recent years, the utilization of highly multiplexed amplicon sequencing (AmpSeq) technologies (Yang et al., 2016) has provided the opportunity to efficiently genotype hundreds to thousands of selected SNPs at a relatively low cost per sample and offer flexibility in marker set design. Thus, these technologies strike the ideal balance between cost per sample, marker density, and flexibility making it possible to continually update and optimize the design of the market set. The strategy of designing highly informative, lower density chips for breeding

applications using the AmpSeq technology has been recently pursued for tropical rice (Arbelaez et al., 2019).

Another strategy to reduce the cost while increasing the efficiency of genotyping is to outsource to a genotyping service provider, which eliminates the need to buy and maintain sophisticated equipment and to hire trained personnel for the breeding program. The large-scale operations of genotyping service providers allow them to reduce the genotyping costs and offer lower prices per sample to consumers. Agriplex Genomics, Diversity Array Technology (DArT), Intertek AgriTech, NRGene Technologies, Eurofins Genomics, and LGC Ltd are examples of such service providers.

Existing rice genotyping marker sets have been previously developed to represent the range of genetic variation within the species and are reliable sources of validated SNPs. We targeted the C7AIR rice SNP array (7K) (Morales et al., 2020) as the primary source of informative and validated SNPs for inclusion in our AmpSeq genotyping platform. The 7K array was developed to detect genome-wide polymorphism within and between subpopulations of *O. sativa, O. glaberrima,* O. *rufipogon,* and O. *nivara*. Its ability to detect polymorphism in different species and subpopulations of rice was demonstrated based on genotyping of a set of 544 diverse *Oryza* accessions. Given that our breeding program is focused almost entirely on *tropical japonica* varieties, we were interested to observe that 2,086 SNPs well distributed across the rice genome were polymorphic in the 74 *tropical japonica* varieties tested (Morales et al., 2020). Thus, the 7K array was considered sufficient to capture the range of genetic diversity relevant to our breeding program and was used to develop an optimized set of markers for an AmpSeq marker set. An initial set of 1,200 informative and well-distributed SNPs from the 7K array were identified for the target germplasm, and the LSU1200 marker set was developed for

Agriplex Genomics, Cleveland, OH, USA (https://agriplexgenomics.com/) AmpSeq genotyping. Subsequently, a subset of 550 SNPs (LSU500) was selected and validated for GS implementation within Southern U.S. rice breeding germplasm.

MATERIALS AND METHODS

Genetic material

A set of 342 lines, the representative germplasm set (RGS), was genotyped with the C7AIR rice SNP array (7K) (Morales et al., 2020). The RGS was selected to represent the genetic diversity of a Southern U.S. rice breeding program and consists of 66 modern and historical varieties, advanced breeding lines, and 276 experimental lines from the largest segment of the target breeding program. Subsequently, the LSU1200 marker set was used to genotype a total of 4,230 lines. These lines consisted of 2,067 experimental lines from preliminary yield trials (PYTs) grown at the H. Rouse Caffey Rice Research Station (HRCRRS) from 2017 to 2019, 1,779 lines from six different bi-parental mapping populations, and 384 lines from a diversity panel (DP) that represents the genetic diversity of the U.S. germplasm, including accessions from different regions of the world (Addison et al., 2020, 2021; Angira et al., 2019). Finally, the LSU500 set was used to genotype 6,406 experimental lines tested from 2018 to 2020 in PYTs and multi-environmental trials. Four bi-parental populations – MPA, MPB, MPC, MPD – were developed through single seed descent in the greenhouse, then grown as $F_{3:4}$ panicle rows in 2019 and tested as $F_{3:5}$ plots in 2020 at HRCRRS. All MP2 lines were genotyped with the LSU1200 at the generation of derivation (F_3). The MPA population originated from the cross CL111/RoyJ, MPB from CL153/LaKast, MPC from CL172/Cypress, and MPD from Presidio/Catahoula. MPA consisted in 297 unique row entries and 130 plot entries, MPB consisted in 296 row entries and 100 plot entries, MPC consisted in 279 row entries and 100 plot entries, and MPD consisted in 286 row entries and 100 plot entries.

Genotyping

The RGS was genotyped with the C7AIR rice SNP array (7K) (Morales et al., 2020) at the Louisiana State University Health Sciences Center, New Orleans, LA 70112, USA. The

LSU1200 and LSU500 were genotyped through PlexSeq™, a mid-density multiplex Next Generation amplicon sequencing (AmpSeq) through the genotyping service provider Agriplex Genomics, Cleveland, OH, USA (https://agriplexgenomics.com/). Leaf samples were collected from fresh leaves and stored at freezing temperatures. Experimental breeding lines were sampled from plots when first tested, while the bi-parental populations were sampled when harvesting individual plants at the generation of derivation.

Phenotypic data

All PYTs were grown in randomized complete block designs at the HRCRRS, near Crowley, LA (30°14'30"N, 92°20'46"W) and were separated according to different grain classes (long grain and medium grain) and different herbicide resistances (Clearfield®, Provisia®, and conventional). Two rows for each bi-parental line were grown as $F_{3:4}$ at the HRCRRS in 2019. The 2020 bi-parental trial was tested with an augmented complete block design at the HRCRRS as $F_{3:5}$ plots. All PYTs and bi-parental plots were drill seeded in plots 1.42m wide and 4.12m long with a Heavy-Duty Grain Drill (ALMACO) at a seeding rate of 100 kg ha-1 at a depth of 2 cm. The water management and fertilization followed the standard practices for Louisiana rice (Saichuk et al., 2014). Grain yield was determined by harvesting the entire plot with a Delta Plot combine (Wintersteiger AG), and a Harvest Master Grain Gauge (Juniper Systems) was used to collect grain weight and moisture. Moisture values of each plot were adjusted to 120 g kg-1 water content. Milling samples (100 g) milled on a PAZ laboratory mill (ZaccariaUSA) to measure milling yield expressed as g kg-1 of whole milled kernels over rough rice seed. Days to heading was measured as the number of days between the date of emergence and the day when 50% of the plot or row had panicles emerging from the sheath of the flag leaf. Plant height was measured as cm from the soil surface to the panicle tip at maturity; two measures per line were averaged on

the rows, while only one measure was recorded for the plots. Chalk, reported as percentage of chalky area of the milled grain and grain length was measured with SeedCount (Next Instruments). These grain traits were measured only on 147 lines for MPA, 112 for MPB, 133 for MPC, and 155 for MPD. Grain yield and milling yield data were not collected on the bi-parental rows.

Filtering and marker set design

Data from the RGS were used to develop the LSU1200 marker set and then further filtered to obtain the LSU500 set. The RGS was split into two sets prior to filtering: key modern and historical lines and the most recent experimental lines from a PYT representing the program's largest market class segment. As the PYT lines made up the majority of the RGS, the data were split and independently filtered to avoid over-prioritizing markers that happened to capture the diversity in one year's PYT but may not be as informative for Southern U.S. germplasm as a whole. After splitting the data, TASSELv5 (Bradbury et al., 2007) was used to filter markers for a minor allele frequency (MAF) greater than 0.1 and less than 20% missing data. After filtering, the –blocks method from PLINK 1.9 (Purcell et al., 2007) was used to identify haplotype blocks. The haplotype blocks represent groups of SNPs with alleles that are co-inherited due to linkage. Haplotype alleles were identified for each block and a custom Python script was used to select the minimum number of SNP markers needed to uniquely identify each haplotype block using a greedy algorithm. The ~1,000 markers chosen from this analysis were combined with trait and purity markers to obtain the set of 1,218 markers referred to as the LSU1200. Pedigree strings from lines grown from 2017-2019 were used to identify a set of lines that were representative of the program's diversity. In addition to these lines, lines from several bi-parental mapping populations and the DP were included to give a total of 4,230 experimental lines. These

experimental lines were used to validate the LSU1200 set. The LSU500 set was developed by filtering the LSU1200 within the validation set. After removing markers with high rates of missing values, the MAF was calculated for all markers. Adjacent markers were considered in blocks of two, and the markers with the lowest MAF were dropped in each pair to obtain the LSU500.

Analysis

The analyses were conducted on the statistical computing software environment R (R Core Team, 2020). Marker density plots were designed with the R package "CMplot" (Yin et al., 2021). The linear mixed models were fit with "ASReml-R v4" (Butler et al., 2018). The genomic relationship matrix (**K**) was calculated from molecular markers with "ASRgenomics" (Gezan et al., 2021) using the (VanRaden, 2008) equation. Outliers in PYTs were removed with studentized residuals above 3 and below -3. The marker subsets were calculated by removing adjacent markers by physical distance. The pedigree-based additive relationship matrix **A** was obtained with the R package "AGHmatrix" (Amadeu et al., 2016). The narrow-sense heritability (h^2) was calculated as the ratio of additive genetic variance component over the sum of the additive genetic and residual error components as follows:

$$h^2 = \frac{{\sigma_a}^2}{{\sigma_a}^2 + {\sigma_e}^2} \quad (1)$$

${\sigma_a}^2$ is the additive genetic variation calculated fitting **K** in the model, and ${\sigma_e}^2$ is the residual error. The broad-sense heritability (H^2) was calculated as the ratio of total genetic variance component over the sum of the genetic and residual error components as follows:

$$H^2 = \frac{{\sigma_g}^2}{{\sigma_g}^2 + {\sigma_e}^2} \quad (2)$$

The Best Linear Unbiased Predictors (BLUPs) or Genomic Estimated Breeding Values (GEBVs) were calculated with a single trait GBLUP with the form:

$$y = Xb + Zu + e \quad (3)$$

Where y are the phenotypic observation, **X** the incidence matrix of fixed effects, **b** the intercept, **Z** the incidence matrix of the random effects, **u** the random effects, and **e** the error. It is assumed that $u \sim N(0, \sigma_a^2 K)$ and $e \sim N(0, R)$, where **K** is the genomic relationship matrix and **R** = ${\sigma^2}_e \, \mathrm{AR1}_{row} \otimes \mathrm{AR1}_{col}$, where $\mathrm{AR1}_{row} \otimes \mathrm{AR1}_{col}$ is the spatial adjustment. For every cross-validation (CV) iteration, 80% of the lines within a trial were selected randomly as the training set, while the remaining 20% were used as the validation set. The same training and validation sets were used for every trait and marker subset iteration. For every trial and trait, the values of 20 CV iterations are presented. The predictive ability is calculated as the Pearson correlation between GEBVs and Best Linear Unbiased Estimates (BLUEs); BLUEs were computed with lines as a fixed effect, calculated for each trial individually, and used as true estimated breeding values (EBVs). These values were calculated using the same model for the GEBVs (2), but allowing **u** to be fixed instead of random, with all observed records included. The GEBVs from the bi-parental rows were calculated using a leave-one-out CV scheme where each line was masked and predicted by the remaining lines. No spatial model was used with the rows. The predictive ability of the phenotype was calculated as the Pearson correlation of the 2019 phenotype with the 2020 BLUEs.

The linkage disequilibrium (LD) was calculated as the pairwise correlation (r^2) of all markers within the same chromosome and averaged across a 100 kb window for the PYTs lines and bi-parental populations separately with a custom script. Heterozygous were set as missing data.

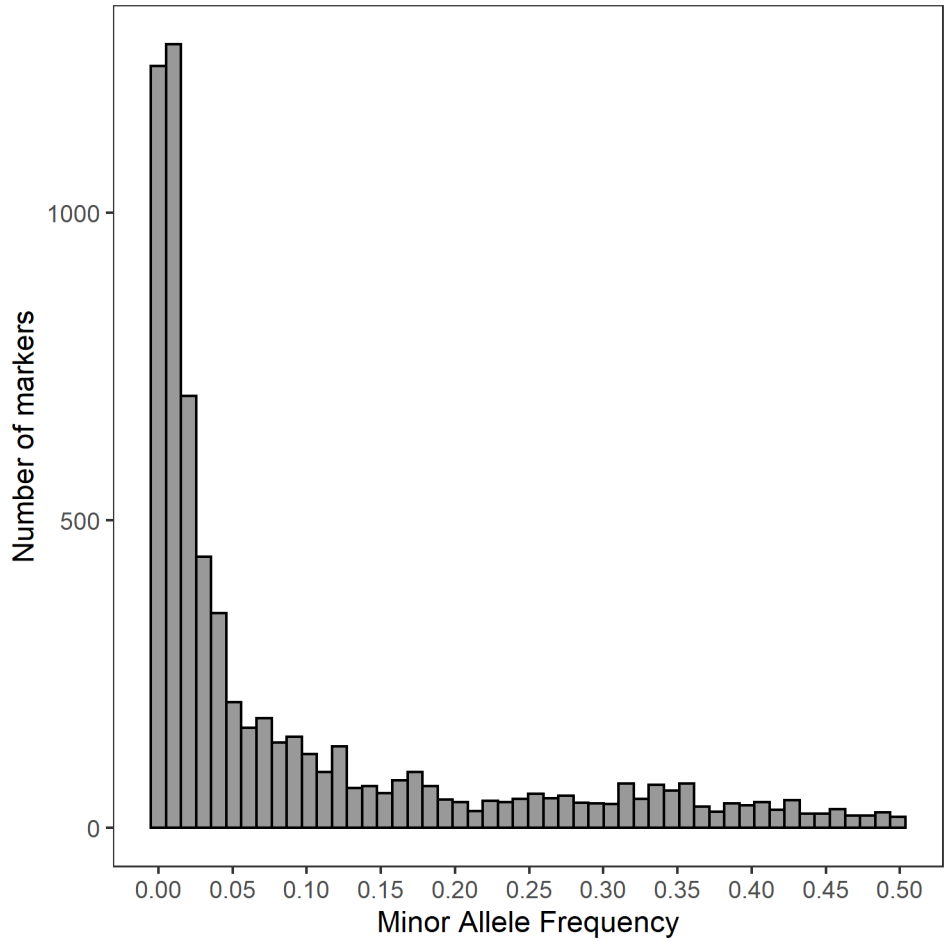
RESULTS

1. Marker set design

C7AIR rice SNP array (7K) genotyping

To develop a marker set optimized for implementing genomic selection (GS) in the Southern U.S. rice germplasm, a set of 342 lines, hereby referred to as the representative germplasm set (RGS) representing the genetic diversity of a Southern U.S. rice breeding program, was genotyped with the rice C7AIR rice SNP array (7K) (Morales et al., 2020). The RGS consisted of 66 modern and historical varieties and advanced breeding lines, as well as 276 experimental lines from the 2017 Clearfield Long Grain Preliminary Yield Trial, the largest segment of the breeding program. The 7K array consisted of 7,098 markers, of which 27 failed and 379 had more than 20% missing data across the RGS. A total of 6,831 markers were polymorphic, with an average Minor Allele Frequency (MAF) of 0.09. Sixty percent of the polymorphic markers had a MAF lower than 0.05, 11% between 0.05 and 0.1, 14% between 0.1 and 0.25, 10% between 0.25 and 0.4, and 4% between 0.4 and 0.5 (**Sup.1**). The average distance between the polymorphic markers was 54.3 kb.

Sup. 1. *Histogram of Minor Allele Frequency of the C7AIR rice SNP array (7K) polymorphic markers across the 342 lines of the representative germplasm set.*



265

266 The 7K genotypic data was used to calculate the additive relationship coefficients of the
267 66 modern and historical varieties within the RGS lines to correlate them with the coefficients
268 calculated with pedigree information alone. A correlation of 0.67 was observed, providing a
269 baseline for how well higher density marker data can capture relationships compared to pedigree
270 (**Sup. 2**). The marker data is more effective at capturing the variation present as it can capture

both within-family and across-family variation; whereas, the pedigree information is limited to across-family variation, and pedigree records are not available for some lines.

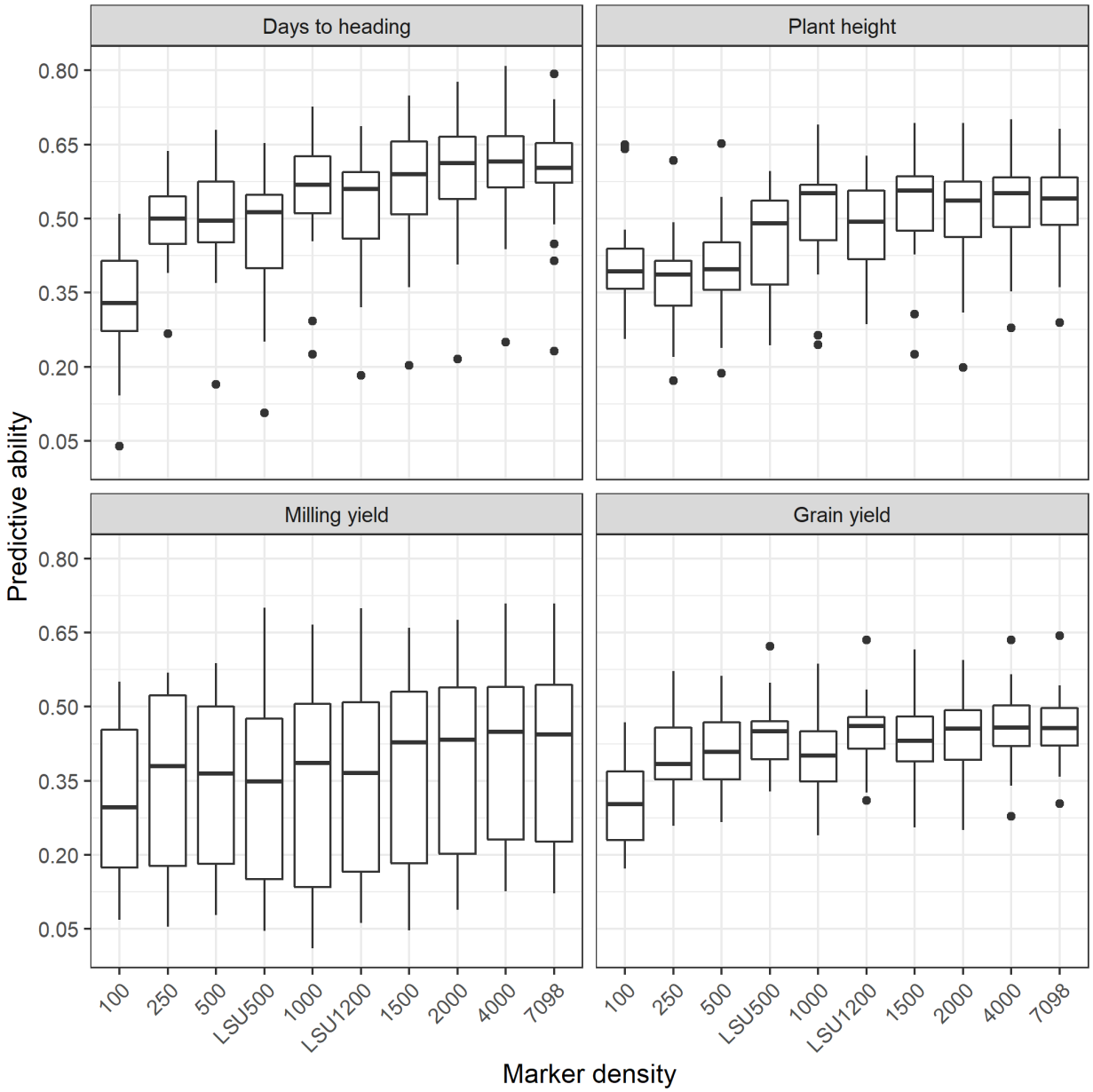
Sup*. *2. *Correlation between additive relationship matrices calculated with pedigree information, 7K SNP array, LSU1200 marker set, and LSU500 marker set of 342 lines of the representative germplasm set.*

	Pedigree	**7K**	**LSU1200**	**LSU500**
Pedigree	1			
7K	0.67	1		
LSU1200	0.63	0.96	1	
LSU500	0.63	0.93	0.99	1

A CV to measure the predictive ability of GS across different marker densities, obtained by removing adjacent markers by increasing physical distance, was performed with the 276 experimental lines of the 2017 Clearfield Long Grain Preliminary Yield Trial to estimate a potential size for the marker set for routine genotyping for GS (**Figure 1**). No major differences in predictive ability were observed between 7,098, 4,000, and 2,000 markers across all traits. Milling yield predictive ability started to decrease slightly with 1,000 markers, grain yield with 1,500 markers, days to heading with 1,500 markers, and plant height with 500 markers.

286

Figure 1. *Cross-validation conducted on the 2017 Clearfield Long Grain Preliminary Yield Trial across different marker densities. Boxplots show the predictive ability for each trait and marker density.*



287

LSU1200 marker set design

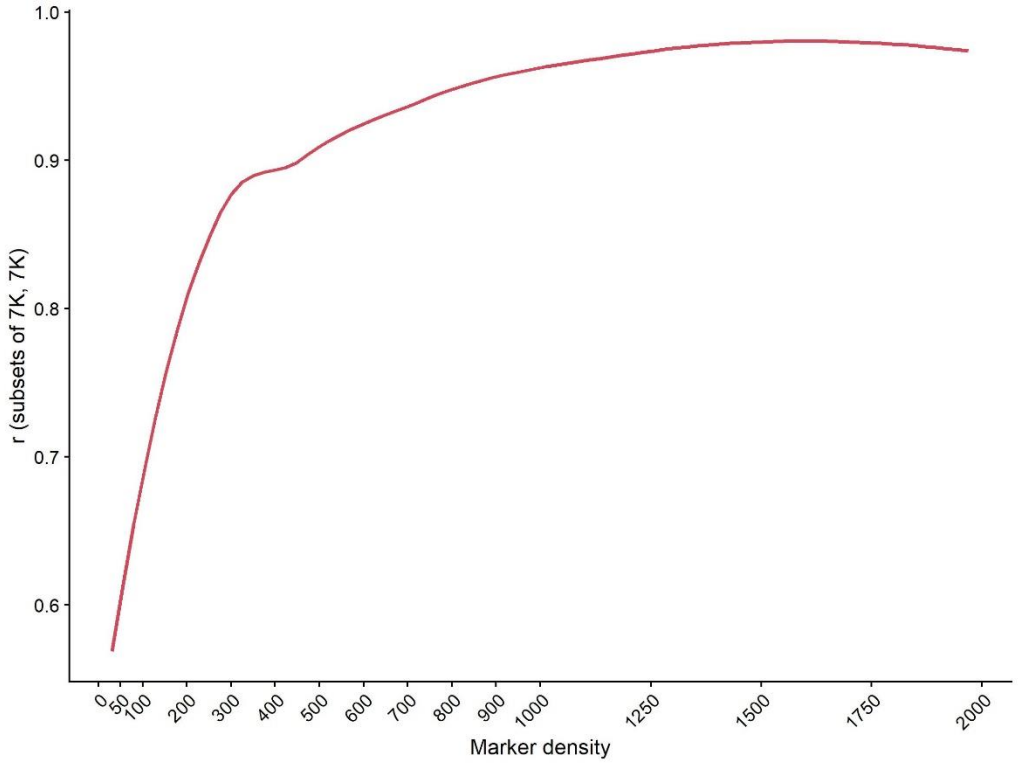
To identify a subset of the 7K markers that could capture as much genetic variation within the target germplasm as the full array, the additive relationship coefficients of the RGS lines calculated using the 7K polymorphic markers were correlated with the coefficients calculated using smaller marker densities (**Figure 2**). No increase was observed beyond 1,200 markers. A marginal decrease occurred at 300 markers and a significant reduction below 300 markers. To select the most informative SNPs from the 7K array, markers with more than 20% missing data and MAF < 0.1 were removed, resulting in 1,935 markers (**Figure 3**)**.** These markers were then used to characterize haplotype blocks and select the minimum number of SNPs needed to tag the different regions. The total number of optimal tag SNPs was 1,174. In addition, 28 trait-specific and 16 genome-wide KASP™ markers used routinely in the breeding program were added, resulting in a total of 1,218 markers, hereafter referred to as "LSU1200". The average MAF of these markers across the RGS lines was 0.18, while the average distance between markers was 291 kbp. It is interesting to note that 1cM in rice is estimated to represent approximately 250 kbp, so the LSU1200 marker set provides approximately 1cM resolution. These markers were tested across all target germplasm using the AmpSeq genotyping platform of the genotyping service provider Agriplex Genomics, Cleveland, OH, USA (https://agriplexgenomics.com/). A set of 4,230 experimental lines was genotyped with the LSU1200, consisting of 2,067 lines from preliminary yield trials (PYTs), 1,779 lines from six unique bi-parental mapping populations, and 384 lines from a U.S. rice breeding germplasm panel. Two hundred and five markers had more than 20% missing data and were excluded while 209 samples failed. Thus, a total of 1,013 SNPs remained for subsequent analysis. When comparing the relationship coefficients with the 7K array and pedigree information, the LSU1200 had a correlation of 0.96 with the 7K and 0.63

311 with the pedigree information, indicating that the LSU1200 was able to reliably capture the

312 overall genetic variation (**Sup. 2**).

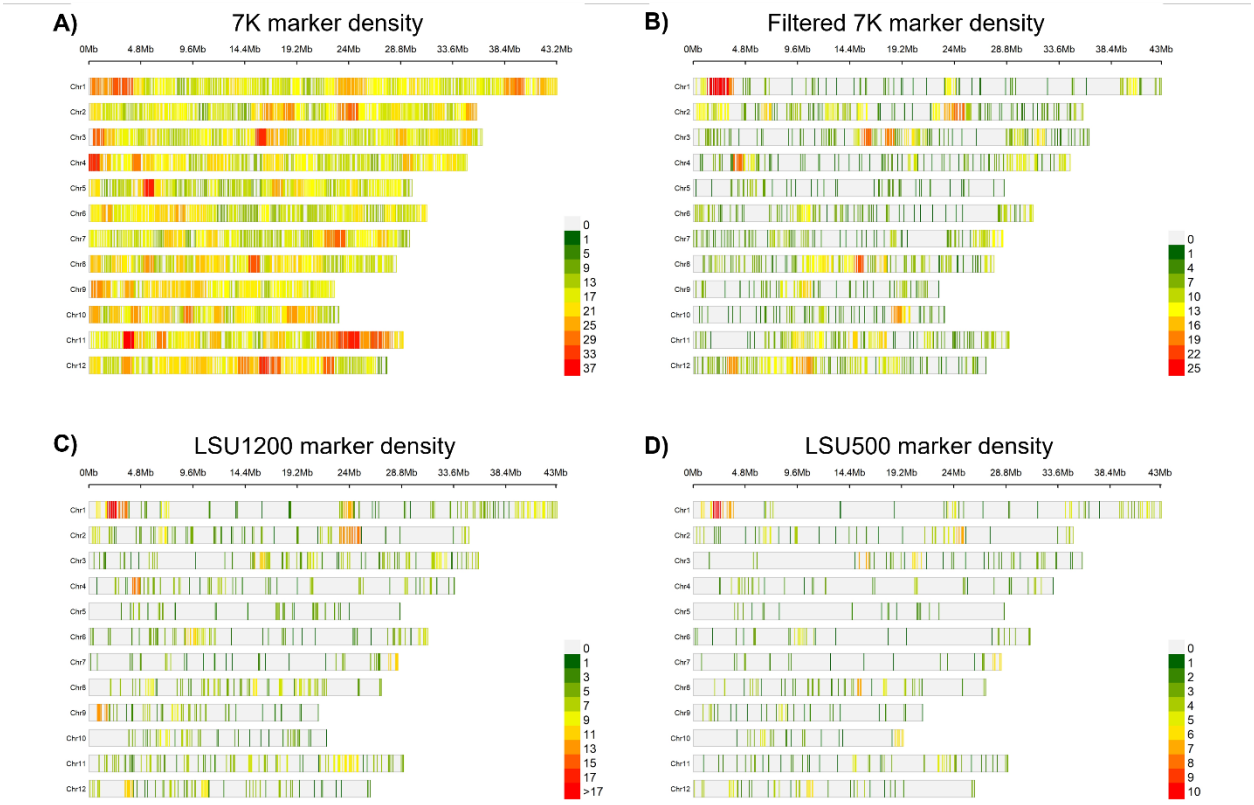
313

***Figure* 2.** *Correlation between additive relationship coefficients calculated with subsets of the 7K SNP array and the full 7K SNP array.*



314

Figure 3. *Marker density across different marker sets. A) C7AIR rice SNP array (7K) marker density; B) 7K filtered with MAF >0.1 and missing data <20% marker density; C) LSU1200 marker density; D) LSU500 marker density.*



315

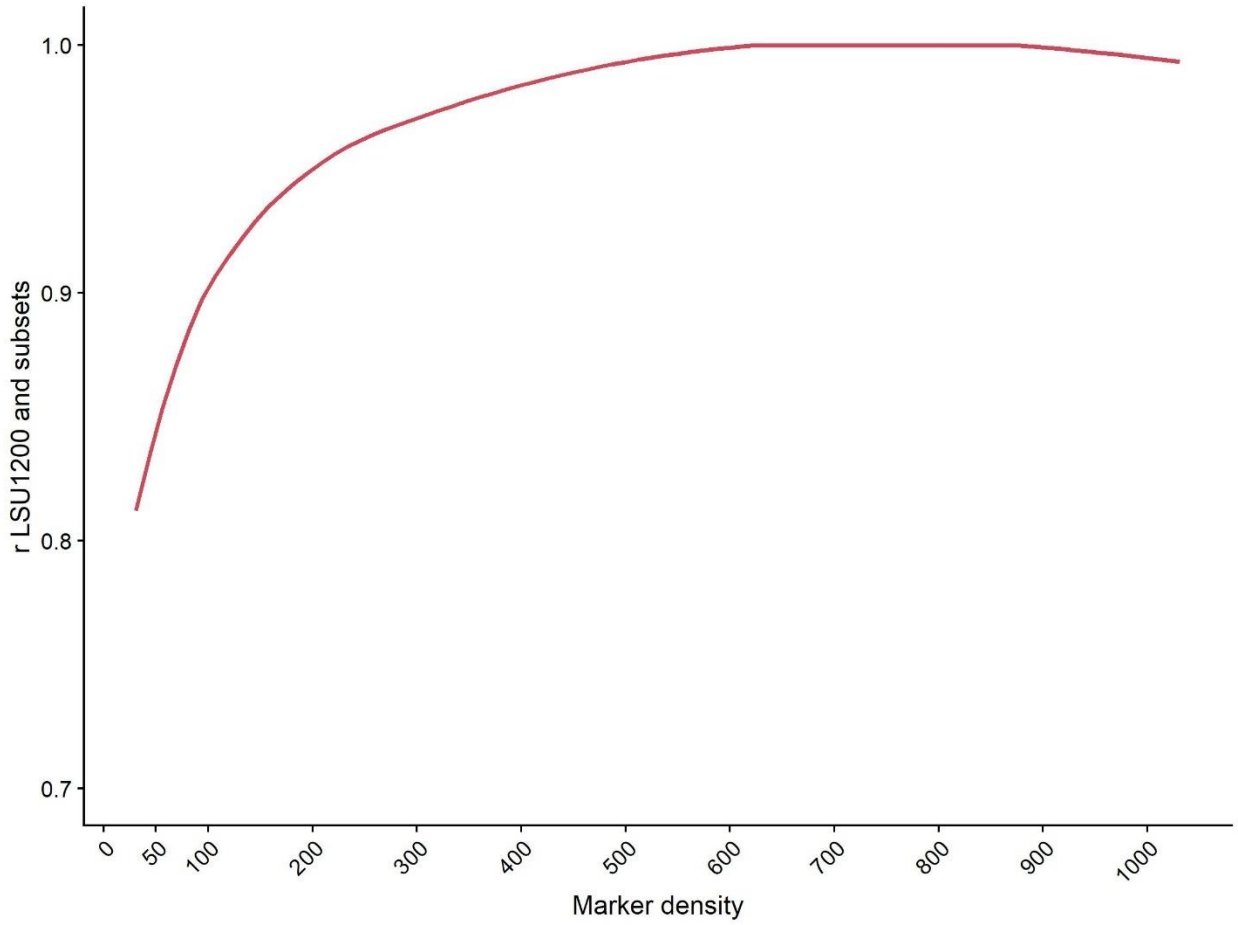
316 ***LSU500 marker set design***

317 Upon identifying the 1,013 SNPs that converted well to the AmpSeq platform, analysis was
318 conducted to determine whether a more stream-lined marker set might be useful for
319 implementation of GS in U.S. rice. The primary goal was to reduce the size of the marker set to
320 minimize costs for routine genotyping of breeding materials, without significantly decreasing the
321 ability to capture the genetic variation of the target germplasm. Correlation between the
322 relationship coefficients calculated with different marker densities (**Figure 2**) showed that
323 marker densities above 400 markers were still highly correlated to the 7K marker set ($r > 0.85$).

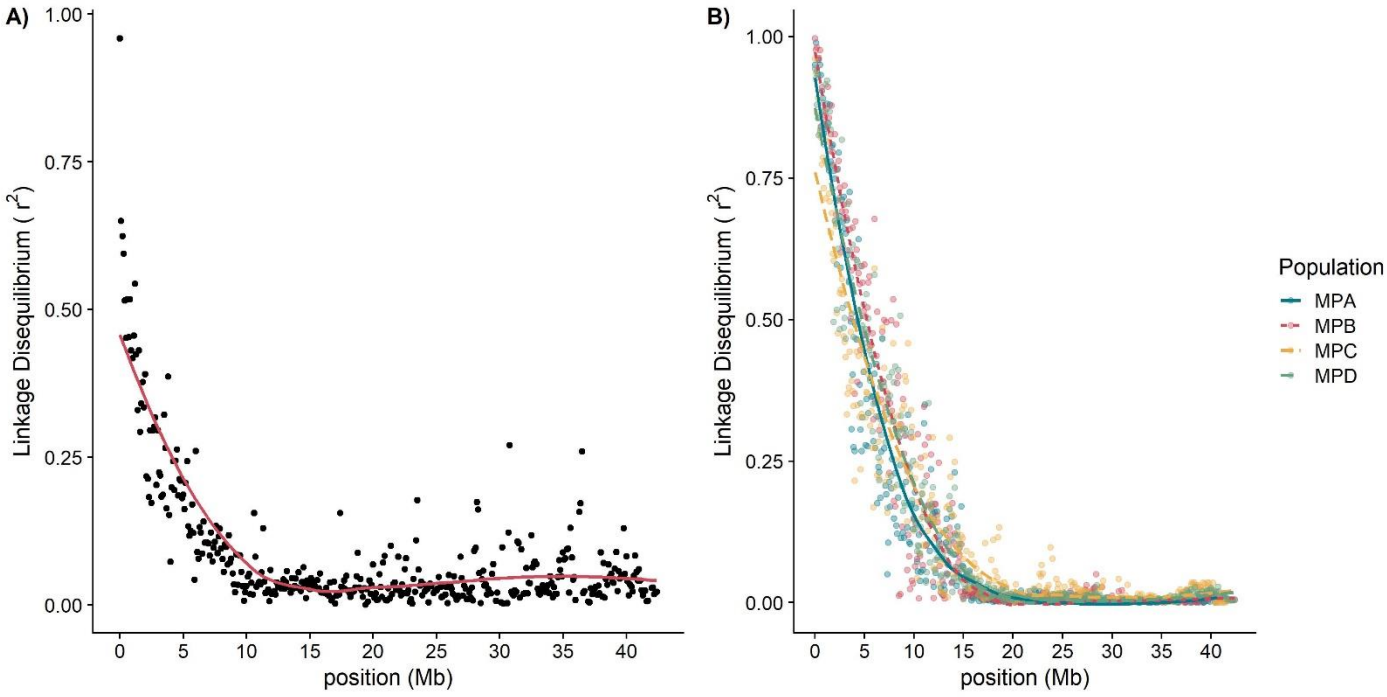
To explore the potential of minimizing the size of the LSU1200 marker set, a diversity panel (DP) of 384 lines that represent the diversity of Southern U.S. rice breeding germplasm and includes accessions from different regions of the world was used to compare the ability of smaller marker densities to estimate the genetic relationship between lines (**Sup. 3**). The correlation of different marker subsets of the LSU1200 to the full LSU1200 set was above 0.95 with marker densities above 200, 0.98 with 500 markers, and the correlation did not increase with marker densities above 600. Marker densities of 50 to 200 were still highly correlated, ranging from 0.85 to 0.95. Based on this observation, a marker set of 500 SNPs the "LSU500" was designed by removing markers with lower call rates and prioritizing SNPs with higher MAF among closely located markers (**Figure 3**). In consultation about the cost per sample with the genotyping service provider, the cost per sample was the same for 500 or 550 SNPs, so the LSU500 SNP set included 550 SNPs. The average MAF across the RGS lines was 0.22, while the average distance between markers was 663.1 kb. The LD calculation within four bi-parental populations and all the PYTs breeding lines showed that LD in the breeding program is high (~10Mb) (**Sup. 4**); and therefore, a well-chosen set of ~300 markers can tag the chromosomal segments segregating within the breeding population. The LD within the bi-parental populations was higher compared to the PYTs and was estimated based on LD decay within every chromosome. The correlation between the additive relationship coefficients calculated with the LSU500 and the 7K SNP set was 0.93 and 0.99 with LSU1200. The correlation with the relationship coefficients generated with pedigree information was 0.63, consistent with what was observed with the LSU1200 (**Sup. 2**). To compare the efficiency of the LSU500 and LSU1200 marker sets for GS, a CV was performed to measure the predictive ability of the two sets across 7 traits within two Clearfield Long Grain PYTs conducted in 2018 and 2019, consisting of 750

different entries in each trial, for a total of 1,500 each year. The results of the CV were, on average, nearly identical between the two marker sets for each trait and trial, with mean values ranging from 0.19 for chalk to 0.69 for plant (**Sup. 5**). A slight reduction was observed with the LSU500 set for grain length in the 2018 trial, but this was not observed in 2019.

Sup. 3. *Correlation between additive relationship coefficients calculated with subsets of the LSU1200 marker set and the full LSU1200 of 384 lines from a diversity panel (DP) that represents the genetic diversity of the U.S. germplasm.*



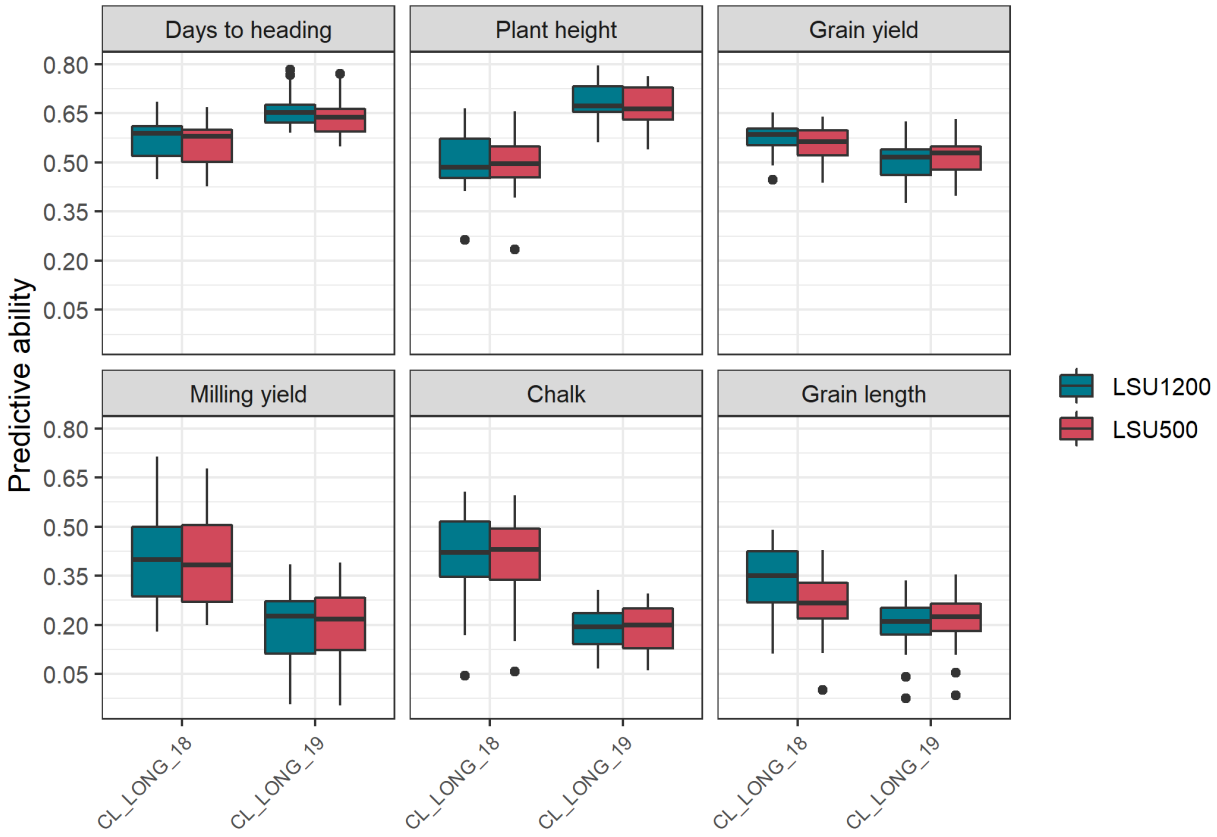
Sup. 4. *Linkage disequilibrium (LD) calculated as the average pairwise correlation (r^2) of markers in a 100 kb window. A) LD across all lines grown in preliminary yield trials from 2017 to 2020; B) LD within four bi-parental populations.*



353

354

Sup. 5. *Cross-validation conducted within two preliminary yield trials grown in 2018 and 2019 with the LSU500 marker set and LSU1200 marker set. Boxplots show the predictive ability for each trait and trial.*



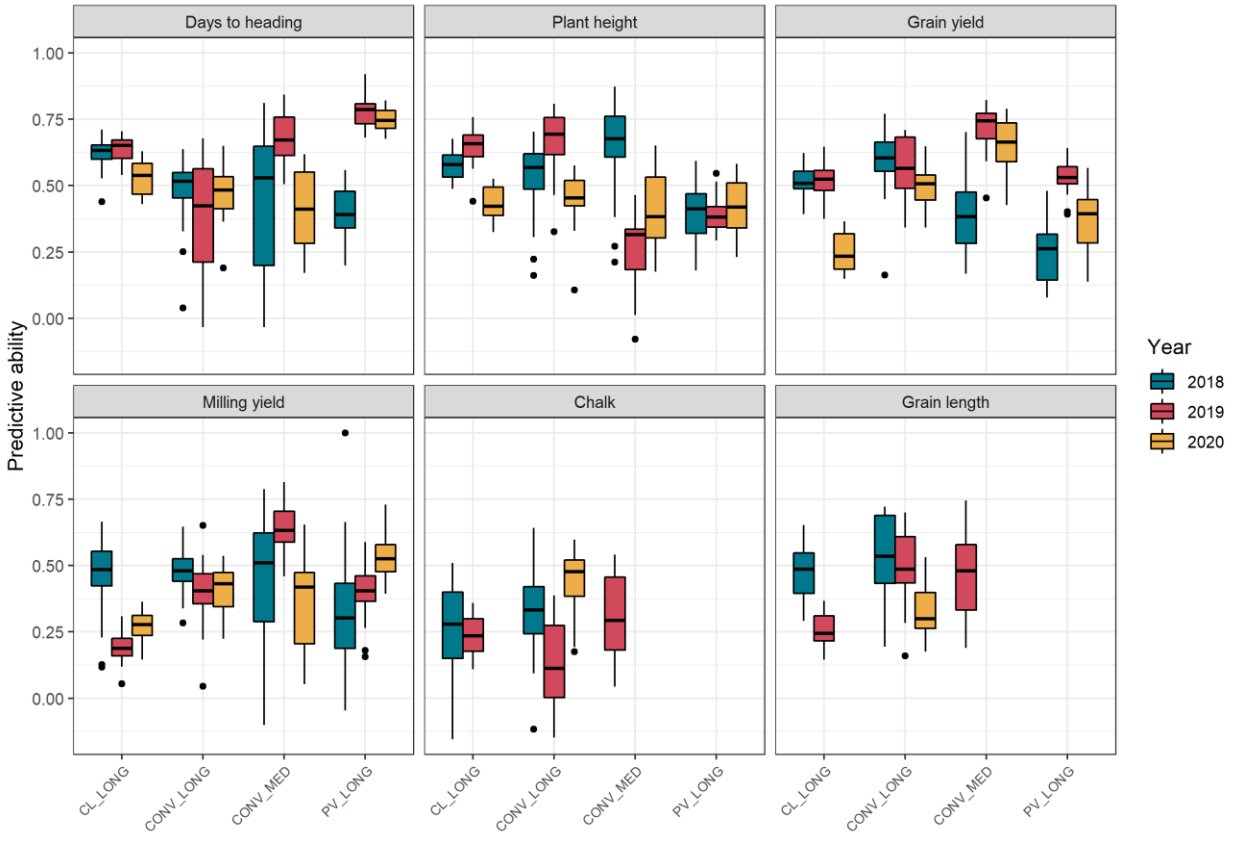
355
356

2. LSU500 for genomic selection

Predictive ability across preliminary yield trials

Once the LSU500 set was determined to be suitable for routine GS applications, the remaining 6,406 experimental lines of the breeding program pipeline from the three most recent years (2018-2020) were genotyped to conduct more extensive investigations into the potential of GS in U.S. rice and to serve as a training population for future predictions. A CV was conducted on 4,078 unique breeding lines (8,172 observations) tested in four different PYTs grown at the HRCRRS in 2018, 2019, and 2020 to evaluate the ability of the marker set to predict line performances within the different segments of the breeding program, Clearfield® Long Grain (CL_LONG), Conventional Long Grain (CONV_LONG), Conventional Medium Grain (CONV_MED), and Provisia® Long Grain (PV_LONG) (**Figure 4**). The predictive ability was influenced by the narrow-sense heritability (**Table 1**) of each trait-trial combination and by the dimension of the training set used in the CV. Lower narrow-sense heritability traits, such as chalk, produced lower predictive ability, and smaller trials like Conventional Medium, where the training population was small, produced lower predictive abilities despite average estimates of heritability. On average, the highest predictive ability was observed for days to heading (0.55); grain yield and plant height were similar (0.49), while grain length (0.43) and milling yield (0.41) had slightly lower values. The lowest values were observed for chalk (0.29). Overall, this marker set produced good predictive abilities across different trials; however, large differences can still be observed within different trials. For example, the average predictive ability of days to heading ranged from 0.78 in PV_LONG in 2019 to 0.40 in CONV_LONG in 2019 and PV_LONG in 2018, plant height ranged from 0.67 in CONV_LONG in 2019 to 0.27 in CONV_MED in 2019, and grain yield from 0.71 in CONV_MED in 2019 to 0.25 in PV_LONG in 2018 and CL_LONG in 2020.

Figure 4. *Cross-validation conducted within each breeding market class segment preliminary yield trial. Boxplots show the predictive ability for each trait and trial.*



381
382

Table 1. *Average broad-sense heritability (H^2) and narrow-sense heritability (h^2) calculated across four market classes preliminary yield trials grown from 2018 to 2020.*

	H^2	h^2
Grain yield	0.59	0.53
Days to heading	0.79	0.75
Plant height	0.66	0.59
Milling yield[a]	-	0.36
Grain length[a]	-	0.44
Chalk[a]	-	0.18

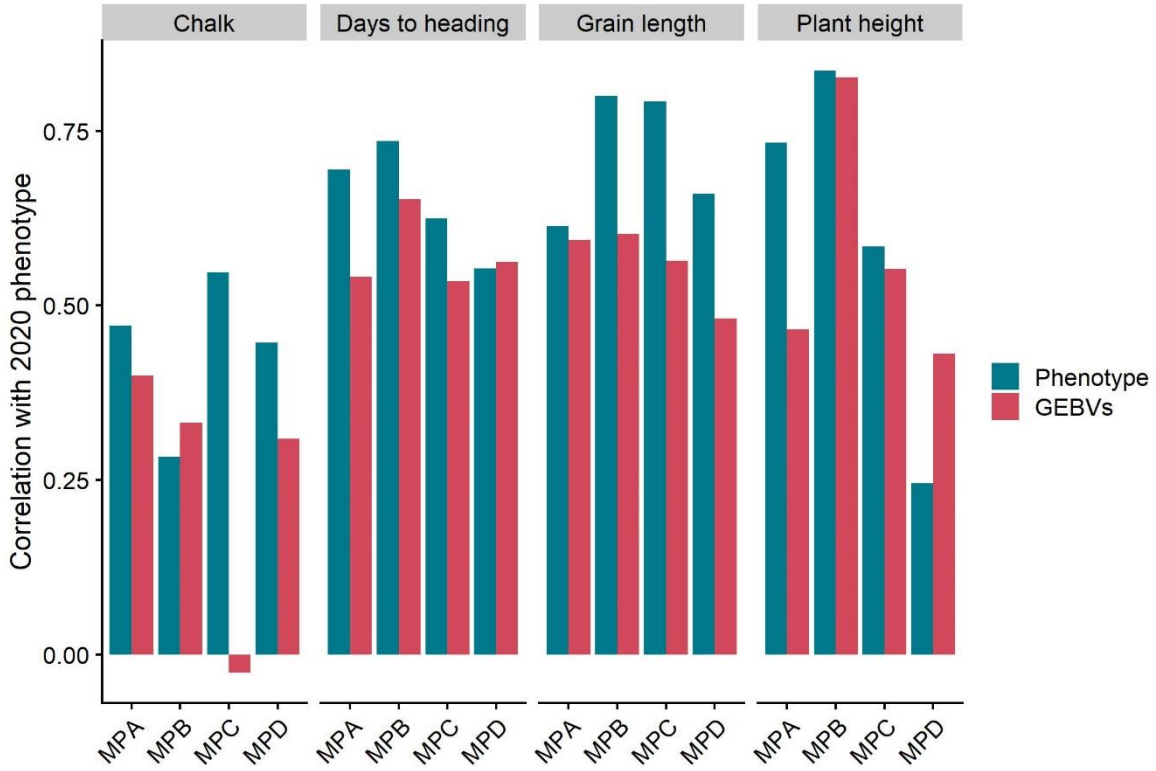
[a]H^2 is not reported for traits based on single measurements.

Genomic selection and phenotypic selection

The predictive ability from CV across PYTs (**Figure 4**, and **Sup. 5**) was calculated by correlating the GEBVs with the phenotype of a single year-one location PYT, thereby assuming the single-year phenotype as the true breeding value of a line. However, a more relevant estimate of the GS potential for applied breeding applications should be calculated comparing both the ability of GEBVs and phenotype to predict future performances of a line. To make this comparison, phenotype data collected in 2019 from four bi-parental populations was used to calculate GEBVs and compare them and the 2019 phenotype data to the observed 2020 phenotype (**Figure 5**). The LSU500 market set contained 265 polymorphic markers with MAF < 0.05, on average, and ranged from 243 to 297 polymorphic markers within each of the four populations. Across all populations, a comparison of the average GS predictive ability to phenotype for plant height was 0.57 and 0.60, respectively. For days to heading, the comparison

was 0.57 and 0.65, for grain length 0.56 and 0.72, and for chalk 0.26 and 0.44. However, both GS predictive ability and phenotype differed significantly across traits and across populations. The correlation of the phenotype data across years ranged from 0.28 to 0.55 for chalk, from 0.26 to 0.84 in plant height; while in days to heading and grain length, the values were more homogeneous, from 0.55 to 0.74 and from 0.61 to 0.80, respectively. On the other hand, the correlation between GEBVs and phenotype ranged from 0.83 to 0.43 in plant height, -0.02 to 0.4 in chalk, 0.54 to 0.65 in days to heading, and 0.48 to 0.60 in grain length. Across the 16 different trait-population comparisons, the reduction of accuracy of the GEBVs compared to the phenotype was 20% or less in 9 cases. In three cases, the GEBVs were more accurate than the phenotype.

Figure 5. *Correlation of 2019 phenotype and 2019 GEBVs with the observed 2020 phenotype within four bi-parental populations (MPA, MPB, MPC, MPD).*



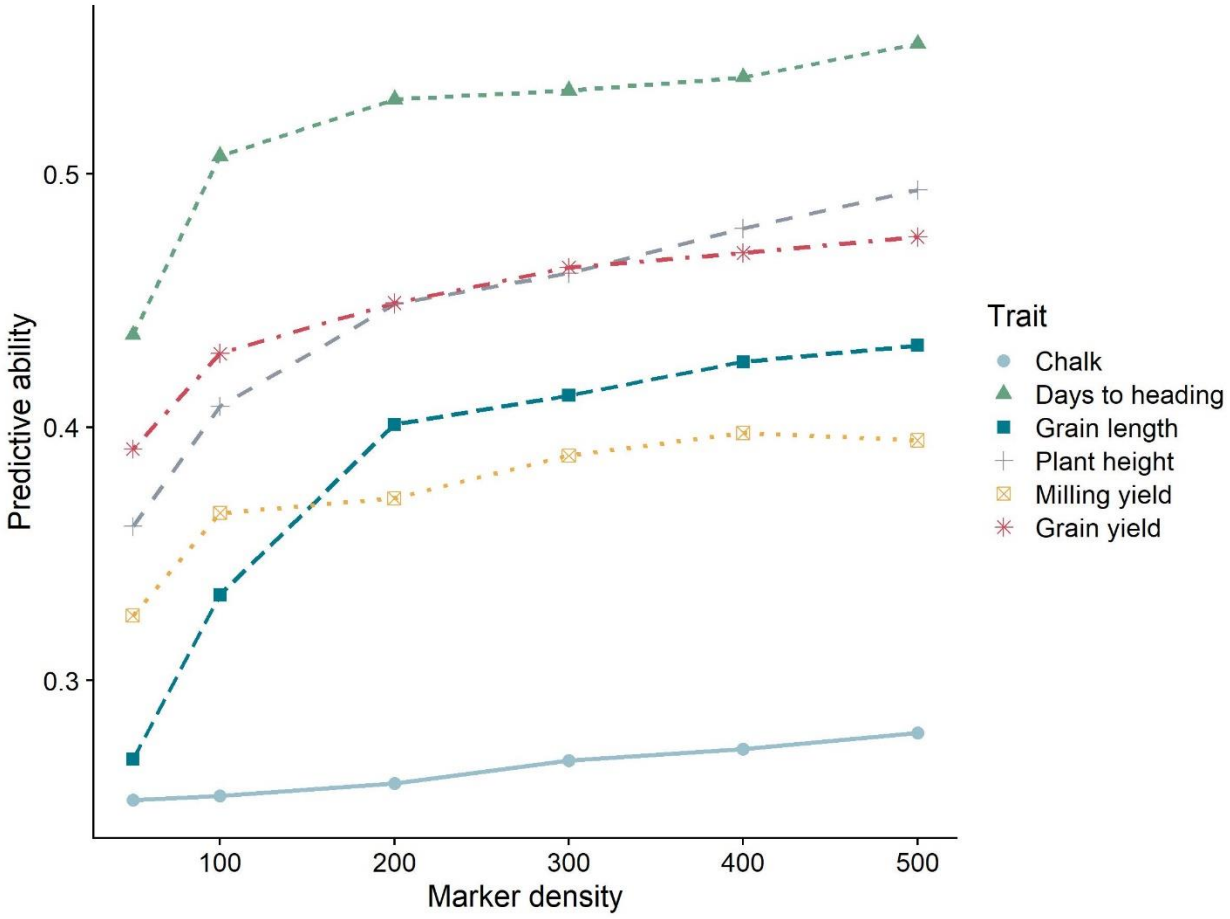
Within the MPC population, very low prediction accuracies were observed for chalk, while the other traits within the MPC had similar accuracies as observed in the other populations. This observation may be explained by sparse marker coverage within the MPC population in the genomic region(s) underlying the variation of the trait in the population.

3. Moving forward: reducing marker density

Marker density across experimental trials

The ability of the LSU500 marker set to capture the genetic variation of the target breeding germplasm is nearly identical to the LSU1200 marker set (**Sup. 2**). However, it was observed that marker subsets below 500 were effective at capturing the relationships among the germplasm as well (**Figure 2**). If it is possible to further reduce the marker set in the future without impacting the predictive ability, this would be desired from a cost standpoint. To investigate the potential to further reduce the size of the marker set, the average predictive ability of CV of the PYTs tested at HRCRRS from 2018 to 2020 was compared across different marker densities by removing adjacent markers at increasing physical distances (**Figure 6**). Overall, no significant differences were recorded with marker densities as low as 200, except for plant height. Results show that the trend across marker densities was different for different traits: plant height predictive ability slightly decreased with densities below 500 and significantly decreased below 200 markers; milling yield and chalk started to decrease with 300 markers, while grain yield, days to heading, and grain length did not significantly lose predictive ability until 200 markers. Overall, higher predictive ability traits with the full set had higher predictive ability with smaller densities except for plant height and grain length that were more affected by lower marker densities than other traits.

Figure 6. *Average cross-validation predictive ability across different marker densities. Cross-validation was conducted within each breeding market class segment using the 2018, 2019, and 2020 preliminary yield trials. The average across all trials is reported for each marker density.*

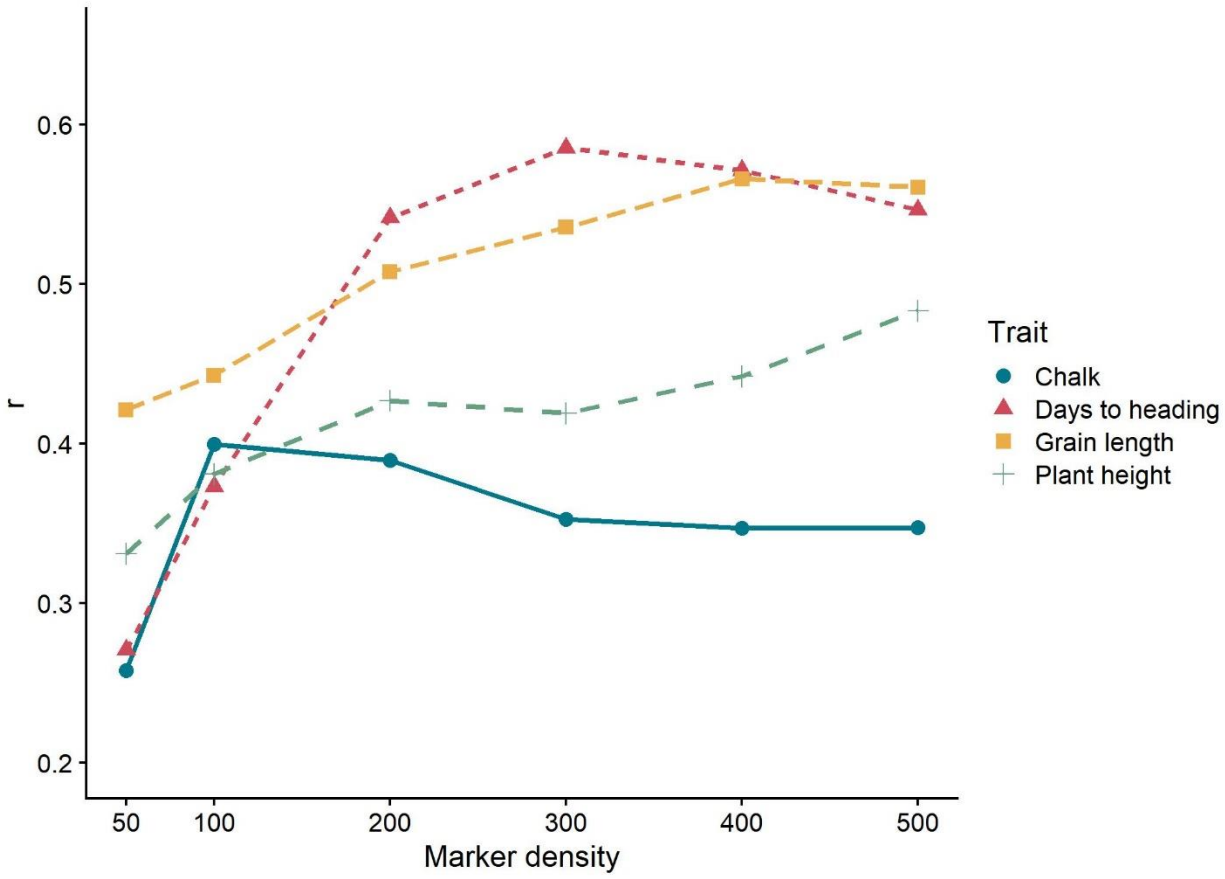


Marker density within bi-parental populations

To quantify the impact of marker density reduction when predicting lines within bi-parental families, the correlation of GEBVs with the phenotype of the following year was compared across different bi-parental populations and different marker densities (**Sup 6**). On average, the predictive ability for days to heading did not show major variation across marker densities of 500, 400, 300, and 200 with values ranging from 0.57 to 0.58. It decreased to 0.37 with 100

markers and to 0.25 with 50 markers. Grain length predictive ability (0.56) decreased gradually with densities below 400 markers reaching 0.42 predictive ability with 50 markers. Plant height predictive ability (0.57) decreased gradually with smaller marker densities reaching 0.27 with 50 markers. Chalk predictive ability increased with lower marker densities from 0.25 up to 0.32 with 100 markers and then decreased to 0.22 with 50 markers.

Sup. 6. *Comparison of GEBVs calculated with 2019 phenotypes and correlated with the observed 2020 phenotype within four different bi-parental populations. Marks show average correlation (r) across populations for each trait and marker density.*



DISCUSSION

1) LSU500 for GS implementation

The objective of this work was to develop and validate a mid-density marker set for routine genotyping of Southern U.S. rice breeding lines and to determine its usefulness for GS. The LSU500 marker set is designed for multiplex AmpSeq and is available for outsourcing at the genotyping service provider Agriplex Genomics, Cleveland, OH, USA (https://agriplexgenomics.com/). This marker set addresses the need for a low-cost, efficient, and germplasm-tailored platform ideal for GS in Southern U.S. rice. A practical example of the initial steps of defining a genotyping method, developing a preliminary marker set, and testing the marker sets suitability for GS was presented. This process can be difficult since it requires multiple different skill sets that are not always present in a traditional breeding program (Santantonio et al., 2020). We provide a usable marker set for programs that may not have the resources to develop it on their own and demonstrate an overall strategy that can be adopted by other programs interested in independently developing a custom marker set.

The design of the set was done by including validated SNPs from the C7AIR rice SNP array (7K) (Morales et al., 2020) that capture the genetic variation of global rice germplasm and some key trait markers. The high LD in the breeding population compared to the LD in global rice facilitated the reduction of the marker density found in the 7K array without losing predictive ability. The SNPs were selected based on their ability to tag individual haplotypes of the breeding population; with this approach, the predictive ability of this set was similar to the full 7K array in a CV experiment. In addition, the ability of the LSU500 to capture the relationships between breeding lines was comparable to the 7K array.

Lowering the number of markers significantly reduces the cost per sample, making the LSU500 cost-effective for an applied breeding program. The cost of outsourcing the genotyping with the 7K array is between $35-$40, genotyping-by-sequencing (GBS) is about $25, the LSU1200 is $10, and our current cost per sample with the LSU500 is $4.80 through Agriplex Genomics. Agriplex Genomics offers a four- to six-week genotyping turnaround time, which is sufficient for our breeding activities and calendar. A significant consideration to implementing GS for our program centers around the logistics of routine activities and ensuring that key breeding processes are not delayed, so it is critical to have service providers who can be relied upon to deliver data in a defined timeframe. Geographic location of the genotyping service provider within the United States was also an important consideration to reduce shipping time, phytosanitary constraints, and paperwork.

2) LSU500 and genomic selection

The LSU500 marker set enables GS across different traits and groups of germplasm at a level of accuracy that is sufficient for implementation in a breeding program. When testing GS with CV across different trials and environments, the predictive ability varied significantly according to the heritability of the trait-trial combination and the size of the training population used in the CV. It is important to note that the heritability of a trait in a specific trial or environment is affected by many aspects, such as the accuracy of the phenotyping method, the trait genetic architecture, the phenotypic variation of the environment, and the presence of large differences between the genotypes tested (Covarrubias-Pazaran, 2019).

Population structure is another important factor to account for in GS experiments. Predictive ability estimates can be biased relative to accuracy obtained in practice if family-size and structure differ between the CV dataset and real life prediction scenarios (Werner et al., 2020).

For example, in the early stages of a breeding program, many full-sib and half-sib individuals are tested together. This is in contrast to later stages where there are many different families represented, each of small size. Thus, in earlier stages, prediction is largely within families, and at later stages, prediction is across families. Both scenarios were considered by testing prediction within bi-parentals (early stage) and PYTs (later stage).

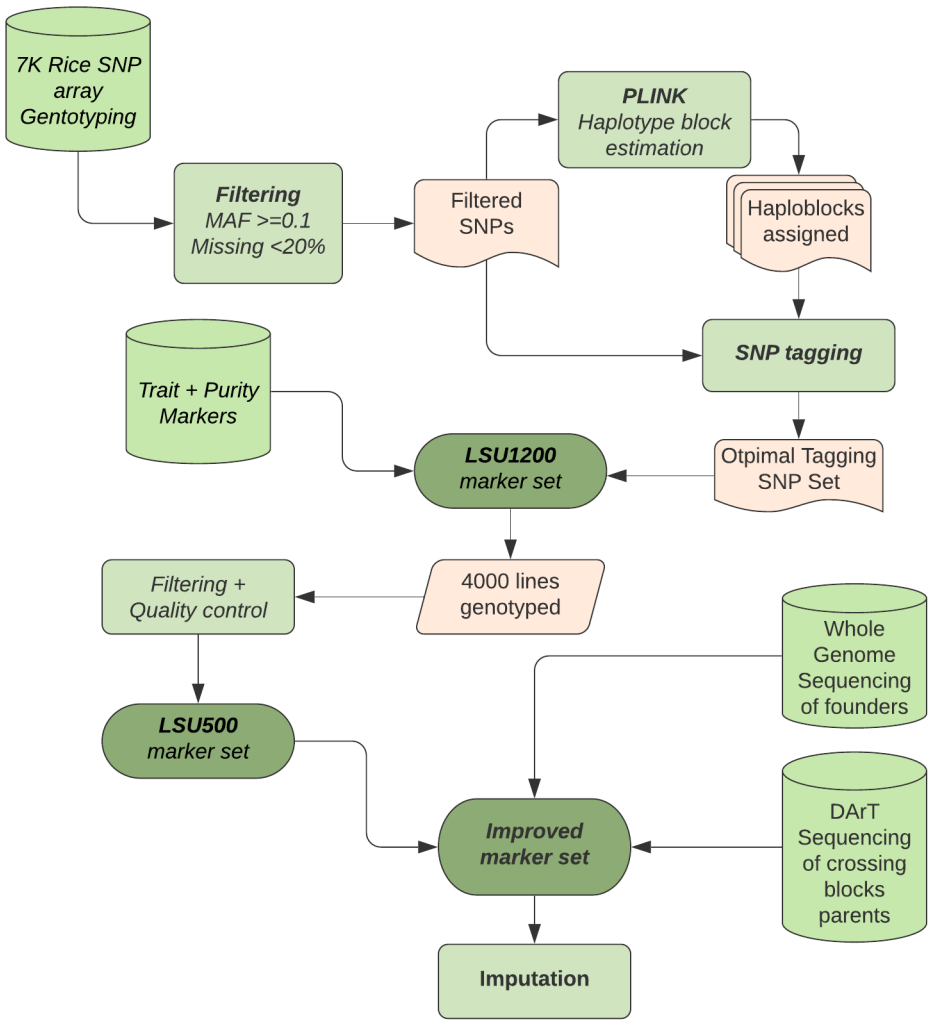
Four bi-parental populations, developed from 8 elite varieties, were phenotyped for key traits across two years, providing the opportunity to compare the correlation of the GEBVs versus correlations between phenotypes measured in different years. This is a useful measure of the predictive ability of GS and provides additional context when compared to using CV alone. In the CV experiments, the prediction accuracy was calculated by comparing the GEBVs to the phenotypes observed in the CV experiment, which assumes that the observed phenotypes are the true breeding value of the line. However, for the traits explored in these populations across two years, the correlation of the phenotypes ranged from 0.25 to 0.84 across year 1 and year 2. When compared to the predictive ability of GS using year 1 as the training set, the correlation of the GEBVs to the year 2 phenotypes was on average 0.49, ranging from -0.02 to 0.83. These results confirmed that GS based on the LSU500 genotypes produced good predictive ability across full-sibs and different bi-parental populations.

3) Beyond the LSU500

The objective of the LSU500 is to provide a robust marker set for implementation of GS in Southern U.S. rice breeding programs, with reliable outsourcing to third-party genotyping providers. The LSU500 is not intended to be the final or fixed marker set moving forward. This initial marker set will enable U.S. rice breeding programs to rapidly and economically explore the potential of GS in their materials. The design and implementation of the LSU500 followed a

continuous integration strategy (**Sup. 7**), whereby GS is initially introduced into a breeding program while technical and strategic aspects of GS implementation are evaluated, optimized, and integrated into the breeding pipeline on-the-go. This strategy allows resources to be invested progressively and reveals possible implementation problems as early as possible. We will continue to improve upon the marker set as more data and information are obtained. The ability to easily update the marker set is another benefit of the Agriplex AmpSeq platform over existing fixed assays. Future iterations of the GS marker set will aim to address some physical gaps throughout the genome and regions of the genome in which the existing markers may not tag independent chromosomal segments. The inability of the LSU500 marker set to predict chalk in the bi-parental population MPC is an example where increasing marker coverage may improve the predictability of the trait. One reason for the presence of regions of low marker coverage can be traced to the original source of SNPs, the 7K array: when selecting markers from a specific array, gaps within the array are inevitably present in the selected subset. To address this issue, we generated whole-genome sequence for the 384 lines of the DP selected to represent the genetic diversity of Southern U.S. rice germplasm, including important founders. These data will be used to select new informative markers to augment the LSU500 set. Markers from other sources, such as the 1k-Rica, which was designed for *indica* rice germplasm, will also be used as sources of new markers (Arbelaez et al., 2019).

Sup. 7. *Diagram of the development of LSU1200 and LSU500 sets and future improvements to the LSU500.*



The LSU500 represents a short-term solution for starting GS in the breeding program. The next stage of our marker strategy will focus on imputation. Imputing progeny from low marker density to high density data obtained for parents can greatly increase prediction accuracy and reduce genotyping costs (Jacobson et al., 2015). Results from varying marker densities across PYTs and bi-parental populations suggest that the number of markers could be further reduced if an imputation strategy is adopted. Due to high LD in bi-parental populations, imputation of

marker information from a lower density (100-500 markers) to higher density (~10,000 markers) datasets obtained on parents can be achieved with high accuracy (Gonen et al., 2018). In a plant breeding program, a new breeding cohort is initiated every year by crossing elite parents. The group of parents is generally small, around 100, and some will be repeated across years. As a result, it is feasible to sequence new parents every year so that each breeding line can be imputed using its direct parents (Gorjanc et al., 2017). In addition, the varieties considered as the founders of the breeding program germplasm have whole-genome sequence data that can be used to impute to even higher densities using tools like the practical haplotype graph (Jensen et al., 2020).

Following this strategy, we used the DArTSeq genotyping technology through the genotyping service provider Diversity Array Technology (DArT), Bruce, ACT, Australia (https://www.diversityarrays.com/) to genotype 188 lines that have been utilized in our crossing blocks since 2017 and represent all the parents of our current breeding populations. This genotyping method provides higher DNA marker information (~10,000 SNPs) for a reduced cost per sample, compared to whole genome sequencing, and will be used for imputation purposes and to monitor program diversity. Information from the whole-genome sequence data of the 384 DP founders will be used as the basis for imputation to increase the density of genotypic information on the lines from the crossing blocks.

Moving forward, when integrating GS into the routine breeding pipeline, it is important that the markers remain informative as the breeding program evolves. When recycling advanced lines and bringing in new material for crossing, allele frequencies are altered, LD patterns are broken, and new alleles are introduced. For these reasons, the marker set for routine genotyping will need to be updated on a regular basis.

CONCLUSION

In this paper, we present our strategy for developing a routine genotyping assay, discuss the results of our efforts to date, and present our vision for how this assay will evolve in the future. Genomic selection promises to bring great advantages to a breeding program, including improving the accuracy of selection and increasing the number of tested experimental lines, both while reducing the cost and duration of operations. At the same time, implementation of this breeding approach requires significant changes to the existing breeding program's structure and practice. Multiple logistical barriers need to be addressed for the adoption of GS to be successful. A custom-tailored marker set for Amplicon Sequencing that is informative for the breeding program's germplasm and can be outsourced to a third-party genotyping provider is an essential ingredient for a rapid and cost-effective initial implementation of GS in the breeding program. The LSU500 is a reliable and efficient marker set with demonstrated utility for routine genotyping of rice in a Southern U.S. rice breeding program. It is currently available at a cost-effective price from the genotyping service provider, and in the future, it will undergo improvements to reduce the cost and increase the accuracy of GS. Moreover, due to the complex nature of GS implementation and the peculiarity of every breeding program, the genotyping method and GS implementation strategy will require further evaluation and customization to meet each program's specific needs.

ACKNOWLEDGMENTS

The authors thank Valerie Dartez for helpful contributions in manuscript preparation and proofreading. This work was supported by the Louisiana Rice Research Board (grant GR-00008264, PI: A. Famoso and grant GR-00006303, PI: B. Angira), Horizon Ag. (grant GR-00004555, PI: A. Famoso), National Science Foundation (grant award number (FAIN) 1826836, PI: A. Pereira), USDA National Institute of Food and Agriculture, USDA-NIFA grant #2014-67003-21858, PI: S. McCouch, and USDA-NIFA grant #2021-67014-34035 PI: J. Richards.

REFERENCES

Abed, A., Pérez-Rodríguez, P., Crossa, J., & Belzile, F. (2018). When less can be better: How can we make genomic selection more cost-effective and accurate in barley? *Theoretical and Applied Genetics*, *131*(9), 1873–1890. https://doi.org/10.1007/s00122-018-3120-8

Addison, C. K., Angira, B., Cerioli, T., Groth, D. E., Richards, J. K., Linscombe, S. D., & Famoso, A. N. (2021). Identification and mapping of a novel resistance gene to the rice pathogen, Cercospora janseana. *Theoretical and Applied Genetics*. https://doi.org/10.1007/s00122-021-03821-2

Addison, C. K., Angira, B., Kongchum, M., Harrell, D. L., Baisakh, N., Linscombe, S. D., & Famoso, A. N. (2020). Characterization of Haplotype Diversity in the BADH2 Aroma Gene and Development of a KASP SNP Assay for Predicting Aroma in U.S. Rice. *Rice*, *13*(1). https://doi.org/10.1186/s12284-020-00410-7

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A. F., Resende, M. F. R., & Muñoz, P. R. (2016). AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. *The Plant Genome*, *9*(3), 1–8. https://doi.org/10.3835/plantgenome2016.01.0009

Angira, B., Addison, C. K., Cerioli, T., Rebong, D. B., Wang, D. R., Pumplin, N., Ham, J. H., Oard, J. H., Linscombe, S. D., & Famoso, A. N. (2019). Haplotype Characterization of the sd1 Semidwarf Gene in United States Rice . *The Plant Genome*, *12*(3), 190010. https://doi.org/10.3835/plantgenome2019.02.0010

Arbelaez, J. D., Dwiyanti, M. S., Tandayu, E., Llantada, K., Jarana, A., Ignacio, J. C., Platten, J. D., Cobb, J., Rutkoski, J. E., Thomson, M. J., & Kretzschmar, T. (2019). 1k-RiCA (1K-Rice Custom Amplicon) a novel genotyping amplicon-based SNP assay for genetics and breeding applications in rice. *Rice*, *12*(1). https://doi.org/10.1186/s12284-019-0311-0

Bernardo, R. (1994). Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. *Crop Science*, *34*, 20–25. https://doi.org/https://doi.org/10.2135/cropsci1994.0011183X003400010003x

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2018). ASReml-R Reference Manual Version 4. *ASReml-R Reference Manual*, 176. http://www.homepages.ed.ac.uk/iwhite/asreml/uop.

Cobb, J. N., Biswas, P. S., & Platten, J. D. (2019). Back to the future: revisiting MAS as a tool for modern plant breeding. *Theoretical and Applied Genetics*, *132*(3), 647–667. https://doi.org/10.1007/s00122-018-3266-4

Collard, B. C. Y., & Mackill, D. J. (2008). Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1491), 557–572. https://doi.org/10.1098/rstb.2007.2170

Covarrubias-Pazaran, G. E. (2019). Heritability : meaning and computation. *Excellenceinbreeding.Org*.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, *22*(11), 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Daetwyler, H. D., Pong-wong, R., Villanueva, B., & Woolliams, J. A. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics*, *1031*(July), 1021–1031. https://doi.org/10.1534/genetics.110.116855

Flint-Garcia, S. A., Thornsberry, J. M., & Edward IV, S. B. (2003). Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology*, *54*, 357–374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Gezan, S. A., de Oliveira, A. A., & Murray, D. (2021). *ASRgenomics: An R package with Complementary Genomic Functions* (pp. 0–39). VSN InternationalInternational, Hemel Hempstead, United Kingdom.

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, *124*, 323–330. https://doi.org/https://doi.org/10.1111/j.1439-0388.2007.00702.x

Gonen, S., Wimmer, V., Gaynor, R. C., Byrne, E., Gorjanc, G., & Hickey, J. M. (2018). A heuristic method for fast and accurate phasing and imputation of single-nucleotide polymorphism data in bi-parental plant populations. *Theoretical and Applied Genetics*, *131*(11), 2345–2357. https://doi.org/10.1007/s00122-018-3156-9

Gorjanc, G., Battagin, M., Dumasy, J. F., Antolin, R., Gaynor, R. C., & Hickey, J. M. (2017). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Science*, *57*(1), 216–228. https://doi.org/10.2135/cropsci2016.06.0526

Grenier, C., Cao, T. V., Ospina, Y., Quintero, C., Châtel, M. H., Tohme, J., Courtois, B., & Ahmadi, N. (2015). Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. *PLoS ONE*, *10*(8), e0136594. https://doi.org/10.1371/journal.pone.0136594

Heffner, E. L., Lorenz, A. J., Jannink, J. L., & Sorrells, M. E. (2010). Plant breeding with Genomic selection: Gain per unit time and cost. *Crop Science*, *50*(5), 1681–1690. https://doi.org/10.2135/cropsci2009.11.0662

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*(1), 1–12. https://doi.org/10.2135/cropsci2008.08.0512

Hickey, J. M., Chiurugwi, T., Mackay, I., & Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics*, *49*(9), 1297–1303. https://doi.org/10.1038/ng.3920

Jacobson, A., Lian, L., Zhong, S., & Bernardo, R. (2015). Marker Imputation Before Genomewide Selection in Biparental Maize Populations. *The Plant Genome*, *8*(2), 1–9. https://doi.org/10.3835/plantgenome2014.10.0078

Jensen, S. E., Charles, J. R., Muleta, K., Bradbury, P. J., Casstevens, T., Deshpande, S. P., Gore, M. A., Gupta, R., Ilut, D. C., Johnson, L., Lozano, R., Miller, Z., Ramu, P., Rathore, A., Romay, M. C., Upadhyaya, H. D., Varshney, R. K., Morris, G. P., Pressoir, G., …

673 Ramstein, G. P. (2020). A sorghum practical haplotype graph facilitates genome-wide
674 imputation and cost-effective genomic prediction. *The Plant Genome*, *13*(1), 1–15.
675 https://doi.org/10.1002/tpg2.20009

676 Juliana, P., Poland, J., Huerta-Espino, J., Shrestha, S., Crossa, J., Crespo-Herrera, L., Toledo, F.
677 H., Govindan, V., Mondal, S., Kumar, U., Bhavani, S., Singh, P. K., Randhawa, M. S., He,
678 X., Guzman, C., Dreisigacker, S., Rouse, M. N., Jin, Y., Pérez-Rodríguez, P., … Singh, R.
679 P. (2019). Improving grain yield, stress resilience and quality of bread wheat using large-
680 scale genomics. *Nature Genetics*, *51*(10), 1530–1539. https://doi.org/10.1038/s41588-019-
681 0496-6

682 McCouch, S. R., Wright, M. H., Tung, C. W., Maron, L. G., McNally, K. L., Fitzgerald, M.,
683 Singh, N., DeClerck, G., Agosto-Perez, F., Korniliev, P., Greenberg, A. J., Naredo, M. E.
684 B., Mercado, S. M. Q., Harrington, S. E., Shi, Y., Branchini, D. A., Kuser-Falcão, P. R.,
685 Leung, H., Ebana, K., … Mezey, J. (2016). Open access resources for genome-wide
686 association mapping in rice. *Nature Communications*, *7*.
687 https://doi.org/10.1038/ncomms10532

688 Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value
689 using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.
690 https://doi.org/10.1093/genetics/157.4.1819

691 Monteverde, E., Rosas, J. E., Blanco, P., Pérez de Vida, F., Bonnecarrère, V., Quero, G.,
692 Gutierrez, L., & McCouch, S. (2018). Multienvironment models increase prediction
693 accuracy of complex traits in advanced breeding lines of rice. *Crop Science*, *58*(4), 1519–
694 1530. https://doi.org/10.2135/cropsci2017.09.0564

695 Morales, K. Y., Singh, N., Perez, F. A., Ignacio, J. C., Thapa, R., Arbelaez, J. D., Tabien, R. E.,
696 Famoso, A., Wang, D. R., Septiningsih, E. M., Shi, Y., Kretzschmar, T., McCouch, S. R., &
697 Thomson, M. J. (2020). An improved 7K SNP array, the C7AIR, provides a wealth of
698 validated SNP markers for rice breeding and genetics studies. *PLoS ONE*, *15*(5), 1–14.
699 https://doi.org/10.1371/journal.pone.0232479

700 Poland, J. A., & Rife, T. W. (2012). Genotyping-by-Sequencing for Plant Breeding and Genetics.
701 *The Plant Genome*, *5*(3), 92–102. https://doi.org/10.3835/plantgenome2012.05.0005

702 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J.,
703 Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for
704 whole-genome association and population-based linkage analyses. *American Journal of*
705 *Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

706 R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation
707 for Statistical Computing, Vienna, Austria. https://www.r-project.org/

708 Saichuk, J. K., Harrell, D. L., Hollier, C. A., White, L. M., Stout, M. J., Brown, S., Webster, E.
709 P., Reagan, T. E., Schultz, B., Salassi, M., Oard, J. H., Groth, D. E., & Linscombe, S. D.
710 (2014). Louisiana Rice Production Handbook. In *LSU AgCenter Publication ID:2321*. LSU
711 AgCenter. https://doi.org/10.1021/ac60320a016

712 Santantonio, N., Atanda, S. A., Beyene, Y., Varshney, R. K., Olsen, M., Jones, E., Roorkiwal,
713 M., Gowda, M., Bharadwaj, C., Gaur, P. M., Zhang, X., Dreher, K., Ayala-Hernández, C.,
714 Crossa, J., Pérez-Rodríguez, P., Rathore, A., Gao, S. Y., McCouch, S., & Robbins, K. R.

(2020). Strategies for Effective Use of Genomic Information in Crop Breeding Programs Serving Africa and South Asia. *Frontiers in Plant Science*, *11*(March), 1–12. https://doi.org/10.3389/fpls.2020.00353

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J. L., & McCouch, S. R. (2015). Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. *PLoS Genetics*, *11*(2), 1–25. https://doi.org/10.1371/journal.pgen.1004982

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, *132*(3), 669–686. https://doi.org/10.1007/s00122-018-3270-8

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., Leckband, G., Snowdon, R. J., & Stahl, A. (2020). How Population Structure Impacts Genomic Selection Accuracy in Cross-Validation: Implications for Practical Breeding. *Frontiers in Plant Science*, *11*(December), 1–14. https://doi.org/10.3389/fpls.2020.592977

Werner, C. R., Voss-Fels, K. P., Miller, C. N., Qian, W., Hua, W., Guan, C., Snowdon, R. J., & Qian, L. (2018). Effective Genomic Selection in a Narrow-Genepool Crop with Low-Density Markers: Asian Rapeseed as an Example. *The Plant Genome*, *11*(2), 170084. https://doi.org/10.3835/plantgenome2017.09.0084

Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., Li, C., Li, P., Yang, Z., Xu, C., & Xu, S. (2021). Genomic selection: A breakthrough technology in rice breeding. *Crop Journal*, *9*(3), 669–677. https://doi.org/10.1016/j.cj.2021.03.008

Yang, S., Fresnedo-Ramírez, J., Wang, M., Cote, L., Schweitzer, P., Barba, P., Takacs, E. M., Clark, M., Luby, J., Manns, D. C., Sacks, G., Mansfield, A. K., Londo, J., Fennell, A., Gadoury, D., Reisch, B., Cadle-Davidson, L., & Sun, Q. (2016). A next-generation marker genotyping platform (AmpSeq) in heterozygous crops: A case study for marker-assisted selection in grapevine. *Horticulture Research*, *3*(January). https://doi.org/10.1038/hortres.2016.2

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2020.10.007

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A., Mezey, J., McClung, A. M., Bustamante, C. D., & McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature Communications*, *2*(1), 1–10. https://doi.org/10.1038/ncomms1467