LSU 500: A mid-density SNP Panel for the USA Rice

Introduction:

Rice (Oryza Sativa) is the most important food crop to human nutrition and caloric intake, providing more than 20% of the calories consumed by humans worldwide. Rice is a key agricultural commodity, being the third most produced crop worldwide.

Oryza sativa is a monocot of the family Poaceae (grasses) and contains two major subspecies: long-grained indica rice variety and the sticky, short-grained japonica or sinica variety. Oryza sativa japonica was first domesticated in the Yangtze River basin in China 13,500 to 8,200 years ago, while O. sativa indica was domesticated around the Ganges River in India 8,500-4,500 years ago.

US Rice

Rice has been grown in the US since the mid-19th century. U.S. rice farming started in South Carolina and Georgia where rice plantations were built on rice cultivation culture brought Western African slaves. After the Civil War, the southeastern rice culture became less profitable and almost disappeared. Further west in southern Arkansas, Louisiana, and east Texas, still in the 19th century, many farmers grew rice in wet marshes and low-lying prairies where they could also farm crawfish when the fields were flooded.

Presently US rice is grown in Arkansas, Mississippi, Missouri, Louisiana, Texas, and California. The southern states grow predominately long-grain types, whereas California grows mostly medium-grain types. Three states—Arkansas, California, and Louisiana—collectively account for approximately 82% of U.S. production. The majority of domestic utilization of U.S. rice is purposed for direct food consumption (58%), 16% is used for processed foods and 16% for brewing beer. The remaining 10% is found in pet food.

The U.S. Department of Agriculture's National Agricultural Statistics Service reported that the year-end production estimate of 2021 was 191.8 million hundredweight (cwt, 1cwt = 45.3592 kg in the US) and the total harvested area was 2.49 million acres. These values are slightly lower from previous years, which is mostly attributed to lower land usage. However, the average yield in the US was 7,709 pounds/acre, being 90 pounds above the previous year and the highest yield on record.

	Class & State	Area Planted (1000 Acres)	Area Harvested (1000 Acres)	Yield/Acre (Pounds/Acre)	Production (1000cwt)
Long Grain	Arkansas	1,095	1,085	7,660	83,111
(Indica)	California	7	7	7,200	504
	Louisiana	380	375	6,890	25,838
	Mississippi	105	100	7,540	7,540
	Missouri	195	190	8,050	15,295
	Texas	188	179	6,900	12,351
US Total		**1,970**	**1,936**	**7,471**	**144,639**
Medium Grain	Arkansas	115	108	7,380	7,970
(Indica)	California	365	363	9,240	33,541
	Louisiana	40	39	6,690	2,609
	Mississippi	NA	NA	NA	NA
	Missouri	4	4	7,600	304
	Texas	2	2	3,500	70
US Total		**526**	**516**	**8,623**	**44,494**
Short Grain	Arkansas	1	1	5,500	55
(Japonica)	California	35	35	7,450	2,608
US Total		**36**	**36**	**7,397**	**2,663**
All	Arkansas	1,211	1,194	7,630	91,136
	California	407	405	9,050	36,653
	Louisiana	420	414	6,870	28,447
	Mississippi	105	100	7,540	7,540
	Missouri	199	194	8,040	15,599
	Texas	190	181	6,860	12,421
US Grand Total		**2,532**	**2,488**	**7,709**	**191,796**

Table 1. Rice Area Planted and Harvested, Yield, and Production by Class – States and United States: 2021 (Adapted from U.S. Department of Agriculture's National Agricultural Statistics Service (NASS), January 2022, Crop Production 2021 Summary); NA – No data available, ctw: Hundred weight

The top yield of the US rice sector can be credited to recent achievements in rice breeding. These modern breeding practices are dependent on establishing a suitable marker set a reliable, and effective genotyping platforms capable of providing informative marker data with fast turnaround time and low cost.

The LSU 500 SNP Panel

The Louisiana State University (LSU) 500 SNP Panel was developed by Chris Hernandez (Postdoctoral Associate) and Tommaso Cerioli (Ph.D. Student) in the laboratories of Prof. Adam Famoso at the Louisiana State University AgCetner and Prof. Kelly Robbins at Cornell University.

The panel is mostly derived from the C7AIR Rice 7K SNP array (Morales et al., 2020). SNPs that were polymorphic in 74 tropical Oryza *japonica* varieties and well distributed along the genome were selected. Initially, 2086 SNPs were identified; this SNP selection was taken through two cycles of reduction (Figure 1) that lead to the final 500 SNPs panel. The goal of the reduction was to reduce the size of the marker set to minimize costs for routine genotyping, without significantly decreasing the ability to capture the genetic variation of the target germplasm, and to enable the use of the panel for Genomic Selection implementation within Southern U.S. rice breeding germplasm.

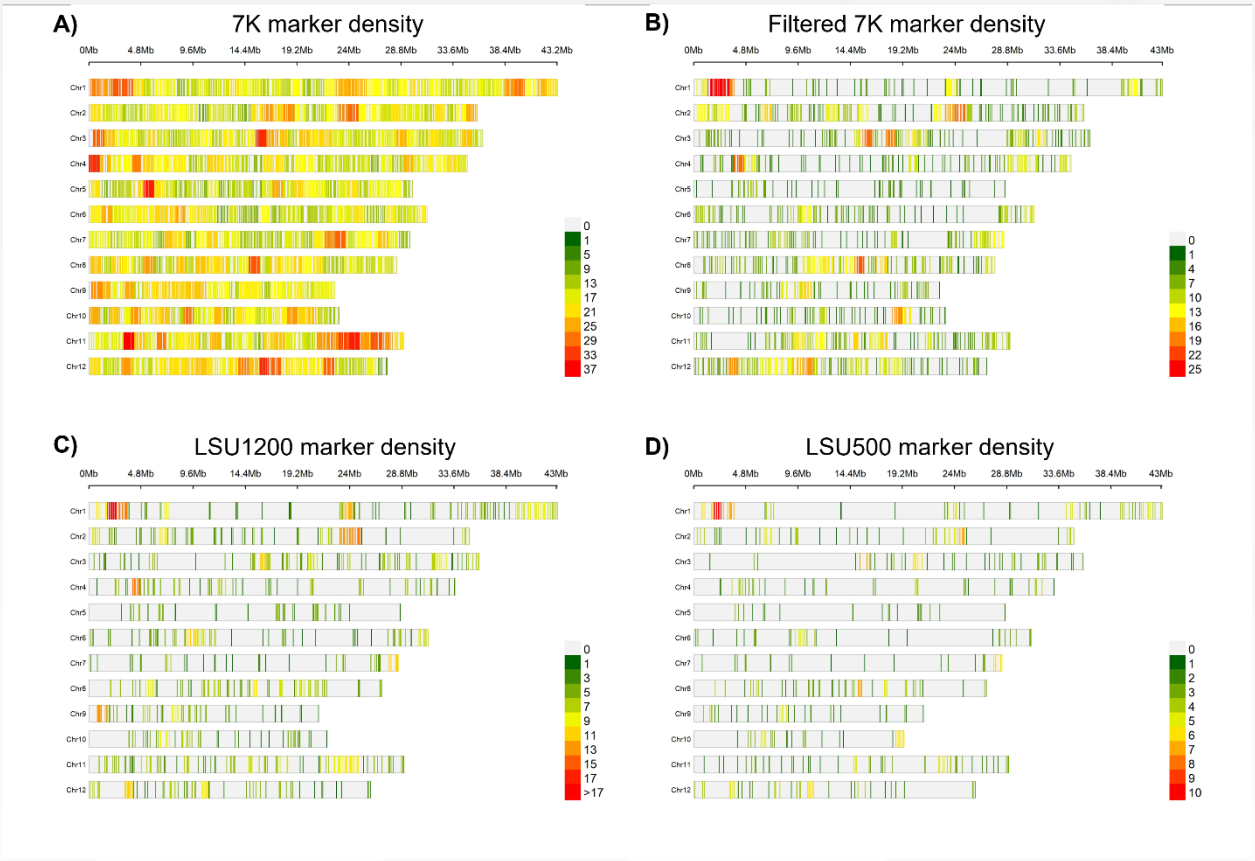

Figure 1. Marker density across different marker sets. A) C7AIR rice SNP array (7K) marker density; B) 7K filtered with MAF >0.1 and missing data <20% marker density; C) LSU1200 marker density; D) LSU500 marker density. From Cerioli et al., 2022.

The reduction process involved filtering markers that demonstrated genotyping success rate higher than 80% (less than 20% missing data), and a minor allele frequency (MAF) greater than 0.1 (average MAF among a geographical diversity set of rice lines = 0.22).

After filtering, haplotype blocks were identified representing groups of SNPs with alleles that are co-inherited due to linkage, and the minimum number of SNP markers needed to uniquely identify each haplotype block were selected (average distance between markers = 663.1 kb). The panel was augmented with trait associated and genome wide markers commonly used in US rice germplasm.

PlexSeq™: The mid-density, multiplexed, SNP genotyping platform

Several attributes of the PlexSeq™ process contribute to its unique value as a genotyping platform:

- The proprietary multiplexing algorithm, *PlexForm™*: The software designs all possible primers around all requested SNPs. Artificial intelligence algorithms identify the optimal sets of primers that can be mixed in one PCR amplification reaction.
- Once the amplifications are completed, the amplicon mixture is equivalent to barcoded libraries produced from other NGS methods. The process is unique because the samples produce amplicon libraries that are equivalent in concentration and do not require any additional equalization steps. A mixture of all the libraries is subjected to one bead cleanup and are loaded onto the sequencer. The process saves time, plasticware, and expenses.
- The method requires only minute quantities of crude DNA that can be isolated from a variety of tissues, enabling a quick and inexpensive DNA isolation process to start the genotyping workflow.
- The *PlexSeq™* workflow consists of:
  - Crude DNA isolation
  - Primary PCR: highly multiplexed, low volume (3ul) PCR amplifications
  - Secondary, barcoding PCR amplifications
  - Pooling: barcoded amplicons are combined into one tube, purified and quantitated
  - Sequencing on an NGS sequencer
  - This relatively simple workflow is amenable to automation; all steps can be carried out on liquid handlers and high-capacity thermocyclers. This enables high-throughput genotyping

Once the sequencing is complete, a proprietary allele frequency-based genotype calling analysis software, *PlexCall™*, provides an automated sequencer to data workflow. This java-based software is tuned for each assay and is fully automated based on only the sequencing output files and a sample sheet indicating sample location on the plate.

Two other features make *PlexSeq™* a unique fit for molecular breeding and seed QA. These applications typically require the genotyping of large number of individuals. AgriPlex Genomics' extensive collection of barcode combinations allows simultaneous sequencing of thousands of individuals, limited only by the sequencer's capacity. Similarly, molecular breeding may require the addition or substitution of only some of the SNP markers as the individual breeding program advances or among programs as the parent's genetic makeup or focus changes. Those changes in SNP composition are also required for QA applications as the diversity of germplasm changes. The fact that the panel is a collection of PCR primers not tethered to a surface (e.g.: chips) provides the flexibility to dynamically customize and alter the composition of the markers in the panel so it best fits the germplasm or the application. The LSU 500 SNP panel is available as a service from AgriPlex Genomics and is also available as a kit to be used by in-house genotyping laboratories.

Applications

Genomic selection: The LSU 500 SNP panel enables genomic selection in *Japonica* breeding programs, and possibly in selected *indica* lines. In a comparison of four elite bi-parental populations, Cerioli et al. observed that the number of polymorphic markers varied between 243 to 297. This marker density is sufficient for imputation to a higher marker density level. In the same study, the average Genomic Selection predictive ability was tested and was found to vary among populations and years and was satisfactory for most traits observed.

The combination of rapid, cost-effective genotyping of a prediction population during the last generation of line fixation, saves expenses on the cost of field space for seed increase, and allows rapid recycling of progeny as parents.

Marker-assisted backcrossing: The LSU 500 panel can be used for background recovery estimates in marker-assisted backcrossing programs. The combination of informative background markers and a selection of trait markers allows accurate estimation of background recovery, ensures recovery of valuable genes from the recipient line, and can provide additional confirmation that a target gene is carried by the selected progeny. Background selection can reduce by 2 or more the number of backcross generations required to achieve >95% recipient parent recovery.

QTL mapping: Although not the primary target application, it is possible to use the LSU 00 panel for biparental mapping purposes. Density of polymorphic markers may, in some cases, be lower than desirable (largely dependent on the parents involved) which may lead to gaps in the linkage map. However, the panel does provide an option for genotyping much of the genome. Any remaining gaps could then be filled in with other marker systems or by further panel customization.

Trait profiling: The LSU 500 panel contains high-value trait markers, covering a range of traits related to disease resistance, grain quality, abiotic stress tolerance, heading date, hybrid rice production and others. These markers are designed to give accurate profiling of the targeted

genes and QTLs across all Oryza sativa genomic diversity, including both indica and japonica, and are highly informative within US rice germplasm.

Seed purity and hybridity: The LSU 500 panel includes the number of markers and genome-wide distribution that enables seed quality applications such as genetic purity testing (uniformity), F1 hybridity tests, and varietal identification fingerprinting

Conclusions

The LSU 500 panel implemented at AgriPlex Genomics provides an excellent, cost-effective alternative for applications requiring moderate to high sample throughput at a modest SNP density.

While it primarily enables genomic selection, the suite of trait markers expands its functionality to a range of additional applications. The panel fits in neatly with rapid line fixation protocols due to its low cost-per-sample and fast turnaround time; it enables locus selection and genomic selection to occur before field increase of seed. The flexibility of the platform also enables continual revision and upgrading of the marker system, ensuring that the panel keeps pace with current trait needs.

References

Rice Outlook: January 2022 Nathan W. Childs, coordinator; Bonnie LeBeau, contributor, USDA-ERS, RSC-22A, Jan.14, 2022, https://www.ers.usda.gov/webdocs/outlooks/103052/rcs-22a.pdf?v=1415.1

Tommaso Cerioli, Christopher O. Hernandez, Brijesh Angira, Susan R. McCouch, Kelly R. Robbins, Adam N. Famoso (2022). Development and validation of an optimized marker set for genomic selection in Southern U. S. rice breeding programs. Submitted to Plant Genome, https://doi.org/10.1002/essoar.10508975.1

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