2009

# Molecular, statistical and genetic analyses of complex agronomic traits in rice

Samuel Agbayani Ordonez Jr.
*Louisiana State University and Agricultural and Mechanical College*

# MOLECULAR, STATISTICAL, AND GENETIC ANALYSES OF COMPLEX AGRONOMIC TRAITS IN RICE

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The School of Plant, Environmental, and Soil Sciences

by
Samuel A. Ordonez Jr.
B.S. Biology, Central Luzon State University, Philippines, (*cum laude*), 1997
M.S. Plant Breeding, University of the Philippines Los Banos, 2003
August, 2009

To my loving wife ***Carol*** and sons –

***Samuel III and Ferdinand Amadeo***

This dissertation is humbly dedicated….

## ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge and thank my Professor, Dr. James H. Oard for my Ph.D. program. For the best mentoring I have had, for the knowledge, guidance and kind support he had provided me, instilling in me discipline, dedication and hard work as key to success; kind words are not enough to express and describe how thankful and grateful I am to have him as my major professor.

To the members of my graduate committee, Dr. Don R. Labonte, Dr. Xueyan Sha, Dr. James P. Geaghan, Dr. Jing Wang, and Dean's Representative, Dr. Michael J. Stout, I thank you all for the knowledge, guidance and help extended to me, especially the very kind interactions and mentoring I have had with you.

To the School of Plant, Environmental and Soil Sciences, headed by Dr. Freddie R. Martin, thank you for all the support and kind friendship I have had throughout my stay in the department.

To fellow graduate students in the department - Marvelous, Suman, Ashok, Reddy, James, Wenting, Suresh, and Nengyi for the healthy discussions, friendships, and fun we have had as graduate students and classmates. Special mention goes to my friend, Marvelous, who shared with me the ideals of how it is to be a good Ph.D. student, and all his sincere help and assistance throughout my Ph.D. program.

To the Filipino Community at LSU and Baton Rouge in general, for providing a family-like atmosphere for me and my family to dwell on. For the friendship and fun, we are forever grateful. Special mention goes to Jenny for the assistance in editing and proofreading my dissertation, as well as the friendship and fun for my whole family, thank you.

To my family, my wife Carol and children, Samuel III and Ferdinand Amadeo for the understanding, love, patience and sacrifices you have had as I embarked on furthering my studies. You have provided me the inspiration and love, and the strength and will to dream big for our future.

To my parents, brother and sisters, who keep on believing in my capabilities and for wishing me the best of luck in this endeavor, I would forever be grateful to you all.

And above all to **God Almighty** for allowing me to surpass all the struggles and hardships of a Ph.D. program. Thank you for providing me wisdom and knowledge.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## ABSTRACT

Novel molecular and statistical approaches are needed for identification of DNA markers associated with complex traits in rice. The first research objective was to evaluate mixed-model and multiple regression approaches for their ability to identify molecular markers associated with complex traits in rice. A combined mixed model and multiple regression approach was optimal for selecting the smallest number of DNA markers associated with relatively high $R^2$ values and for consistency with previous mapping studies.

Support Vector Regression (SVR) was evaluated in the second research objective for the ability to generate high levels of accuracy and power for markers associated with complex traits. High levels of prediction accuracy and power were observed for the selected markers. SVR produced greater model accuracy and ability to explain trait variation than multiple linear regression.

Single nucleotide polymorphic (SNP) markers for aroma, amylose content and gelatinization temperature were evaluated in the third research objective for marker-assisted improvement of breeding lines. This strategy increased frequency of desired alleles by an average of 26 percent in only two generations. Genetic analysis of pollen sterility was conducted in the fourth research objective for an $F_2$ population derived from an outcross between a weedy biotype and a commercial variety. Segregation analyses revealed that seed fertility was governed by two dominant genes, a result similar to the cytoplasmic male sterile (CMS)-WA system used to develop commercial hybrids. Pollen sterility was controlled by two recessive genes. The pollen sterility trait could be exploited as a new source of CMS for hybrid rice breeding. Additional research is needed to confirm if lines developed from this natural outcross represent a new source of CMS. Overall results show that both standard and new data mining approaches

can be used to successfully identify candidate genes and DNA markers associated with complex agronomic traits. In addition, the SNP markers were shown to rapidly enrich frequency of desired alleles associated with rice grain and cooking quality traits. All results demonstrated that a combination of molecular, statistical, and genetic approaches created an effective strategy to advance our understanding of factors that govern complex traits in rice.

# CHAPTER 1 GENERAL INTRODUCTION

## 1.1 Importance of Rice in Asia and the U.S.

Rice (*Oryza sativa* L.) is one of the most important food crops in the world, serving as the principal source of calories for more than half of the world's population (Singh and Khush, 2000). Asia produces and consumes approximately 90% of the rice on earth and by 2025 nearly 4 billion people, mostly poor, will consume rice as a basic food. Global production is projected at 417 million tons of milled rice in 2007, but global consumption continues to outpace production which is expected at 423.2 million tons of milled rice. (Grain: World Markets and Trade, May 2006).

Rice production and marketing in the United States is a multibillion dollar industry. At the farm level alone, rice generates more than $1.5 billion in revenues. In 2007, rice was planted on more than 1.1 million hectares in the United States with production estimated at 8.6 M MT. U.S. rice production is a viable commercial industry in Arkansas, California, Louisiana, Mississippi, Missouri, and Texas. The U.S. produces high quality varieties of short, medium and long grain rice, as well as specialty rice including jasmine and basmati types. U.S. rice farmers produce two percent of the world's annual rice supply and represent the world's fourth largest rice exporting country. Approximately half of the annual U.S. rice production is used domestically. Americans consume ~ 11 kg of rice per year which is substantially below world consumption levels of 85.9 kg per capita. Louisiana ranks third in terms of rice total production following Arkansas and California. The rice industry in Louisiana accounted for $235 M in 2006 from 350,000 acres with average yields of 5,820 lbs/acre for a total of 20.1 M cwt (Louisiana Farm Reporter; http://www.lsuagcenter.com/agsummary/ progressreport. aspx). Louisiana rice planting for 2007 was 360,000 acres, up 3 percent from a year earlier, but still the lowest acreage

planted since 1914. For 2008, the area planted was 464,000 acres with average yield of 5,830 lbs.

### 1.2 Molecular Marker Tools to Study Complex Traits in Rice

The application of molecular markers as a tool for rice improvement has resulted in rapid development of new and improved elite rice lines in the last decade (Collard and Mackill, 2008). Molecular markers allow selection for particular characters or traits on the basis of a simple laboratory test on a small amount of leaf or grain tissue, rather than direct measurement of the character itself. There are several types of molecular markers available for use. Among them are restricted fragment length polymorphisms (RFLP), random amplified polymorphic difference (RAPD), amplified fragment length polymorphisms (AFLP), simple sequence repeats (SSR) and single nucleotide polymorphism (SNP) that can detect a single nucleotide difference in the DNA sequence between two individuals. The utility of SNP markers has been reported in several crops with great success (Issiki et al., 1998; Bundock et al., 2004; Till et al., 2004), as well as its potential for plant genomic research (Feltus et al., 2004). In rice, *indica* and *japonica* genomic sequences have been published and are publicly available (Feltus et al., 2004; Shen et al., 2004; Takashi M, 2005) that allow development of SNP markers for efficient marker- assisted breeding.

### 1.3 Association Genetics

Association genetics is an alternative strategy to standard Quantitative Trait Loci (QTL) mapping approaches that is routinely used in human studies (Baker, 2008) and is gaining support in the plant research community (Hayes and Szucs, 2006). The principal advantage of this method, generally referred to as “linkage disequilibrium” mapping, is based on the ability to rapidly query informative regions of the genome among unrelated individuals that have generated numerous meiotic events over multiple generations. Linkage disequilibrium studies

have been conducted for various marker-trait associations in maize (Belo et al., 2008; Weber et al., 2007), rice (Mather et al., 2007), potato (Simko et al., 2006), barley (Casa et al., 2008) and wheat (Breseghello and Sorrells, 2006). Methods for association genetics other than LD mapping have also been evaluated. For example, Zhang et al. (2005) investigated a discriminant analysis procedure for rice inbreds while the same procedure identified candidate loci associated with agronomic traits in sweetpotato (*Ipomoea batatas*) (Mcharo et al., 2004).

**1.4 Statistical Methods for Association Mapping**

The TASSEL software package is a popular strategy for association genetics to incorporate population structure (Q) and kinship (K) estimates into a mixed model framework for marker-trait evaluation of unrelated individuals (Yu et al., 2006). Consideration of kinship relationships and population structure did improve power and reduce Type 1 error in association mapping (Yu et al., 2006). The TASSEL software has been used recently in association studies of complex agronomic traits in barley (Rostoks et al., 2006), potato (Simko et al., 2006), sorghum (Casa et al., 2008) and wheat (Breseghello and Sorrells, 2006). Parisseaux and Bernardo (2004) developed a mixed model for hybrid crops incorporating effects for general combining ability of markers associated with agronomic traits. Arbelbide et al. (2006) developed a mixed model for self-pollinating plants that accounted for multiple location effects and kinship based on pedigree records. Arbelbide and Bernardo (2006) applied single and multiple marker analyses in the mixed model format for candidate loci and genes associated with bread quality traits in wheat (*Triticum aestivum* L.).

The multiple regression approach, based on information criteria such as Bayesian Information Criterion (BIC; Schwarz, 1978) and Akaike Information Criterion (AIC; Akaike, 1974), has been investigated to address selection bias present in standard QTL mapping techniques (Bogdan et al., 2004; Bogdan and Doerge, 2005; Piepho and Gauch, 2001; Ball,

2001). The multiple regression strategy proposes to identify the fewest number of variables that minimize BIC or other information criteria as opposed to standard hypothesis testing (F test) to build the optimal predictive model. Multiple regression used with various selection criteria has been reported to be superior to Composite Interval Mapping in simulated studies (Broman and Speed, 2002). Software programs such as GLMSelect (SAS Institute) can readily implement multiple regression with multiple fixed effects and epistatic interactions based on standard F tests or different selection criteria.

Although the mixed model procedure for association genetics has been successful in identifying individual QTLs in several crop plants, a genome-wide test for multiple effects and two-way interactions (epistasis) is not feasible in this method. Therefore, a two-step method was developed to identify epistatic interactions and to characterize allelic variation at the *barren inflorescence2* (*bif2*) locus in maize (Pressoir et al., 2009). The mixed-model approach by Yu (2006) was used to identify associated effects and the SAS GLMSelect procedure was then used to identify QTL and conduct genome-wide scans for potential interactions with *bif2*. Similar methods were employed by Manicacci et al. (2009) to identify epistatic interactions between *Opaque2* and *CyPPDK1* loci for kernel quality traits in maize.

**1.5 Support Vector Regression**

The support vector regression (SVR) method was developed by Vapnik (1995) to increase model accuracy and power by approximating the unknown nonlinear relationship between the continuous response variables and corresponding predictors. SVR has gained broad popularity due to its robustness to noise, computational efficiency, and simplicity of the method. Implementation of SVR to study the relationship between maize hybrid and inbred lines has been previously investigated (Maenhout et al., 2007; De Baets et al., 2008).

### 1.6 SNP Marker Development for Marker-Assisted Breeding in Rice

The demand for high quality or special purpose aromatic and basmati rices in the U.S. and elsewhere has increased during the past two decades (Cordeiro et al., 2000; Jin et al., 2003). In the U.S., ~ 12% of the total rice consumed is aromatic, primarily imported for the Asian-American community (Sha, 2005). With increasing market demand at the U.S. and international level, breeding for special purpose aromatic rice is attractive as it creates the option of securing higher returns over conventional rice due to higher price (Jin et al., 2003). However, the traditional breeding method of crossing and selection is tedious and labor-intensive, particularly for recessive traits such as aroma that may be lost through selfing and subsequent segregation of desired allelic combinations. Moreover, grain evaluation through taste to determine aroma is often difficult, time-consuming and unreliable at times.

Molecular markers that can differentiate rice lines for different quality traits are therefore desirable to expedite development of new aromatic rice varieties. Kadaru et al. (2006) developed a modified procedure based on standard Ecotilling (Comai et al., 2004) for rice SNP discovery and genotyping referred to as Alternative Ecotilling (AE). Four previously reported and 14 new SNPs in the *alk* and *waxy* genes among 57 rice accessions based on comparisons with sequencing results were characterized by AE for GT and AC, respectively. In addition, new SNP markers for haplotype-specific markers in exon 7 of the *BAD2* gene for marker-assisted identification and introgression of the aroma gene in U.S. rice were developed in Dr. Oard's laboratory. These SNPs can distinguish aromatic and non-aromatic phenotypes that were consistent with corresponding marker haplotypes for all progeny tested.

### 1.7 Male Sterility and Its Importance to Rice Breeding and Genetics

Male sterility is a characteristic found widely in plants (Zuo et al., 2008) with more than 100 different male sterile mutants reported in rice (Bruskiewich et al., 2003). Male sterility

prevents self-fertilization, but represents tremendous value for basic research of plant reproduction and for commercial exploitation of heterosis (Zhang et al., 2008). Recently, several male sterile rice mutants from different sources (Zuo et al., 2008, Zhang et al., 2008) have been characterized and mapped. Inheritance studies of these different mutants showed that a single recessive nuclear gene (Zuo et al., 2008, Zhang et al., 2008) controlled expression of this trait. There are generally two categories of sterility recognized in rice: cytoplasmic male sterility (CMS) and nuclear male sterility (NMS) (Zhang et al., 2008). The CMS system is controlled by the interaction of cytoplasmic and nuclear genes (Virmani, 1994). The genetic factor(s) present in the cytoplasm-has been reported to occur in mitochondrial DNA (Levings and Pring, 1976; Forde and Leaver, 1980; Kadowaki et al., 1986). This phenomenon in rice was first reported by Weeraratne in 1954 (Li et al., 2007). Shinjyo and Omura reported in 1966 the first CMS observed in elite rice cultivars. A CMS line was designated CMS-BT, being the product of an inter-subspecific cross between *indica* Chinsurah Boro II and *japonica* Taichung native 65 (Shinjyo, 1975). In 1964, Yuan Long Ping discovered male sterility in the *indica* variety Dong-Ting-Wan-Xian, but the breakthrough came in 1970 when he discovered a spontaneous-male-sterile plant referred to as CMS-WA in a wild population in Hainan Island, China (Yuan, 1977). Four years later, the first hybrid rice combination, Nanyou-2, was released that produced higher yield potential as compared to inbred varieties. Since then, several CMS lines have been developed through inter-specific, inter-subspecific, and inter-varietal modes of hybridization (Li et al., 2007).

There have been only a few studies that investigated the molecular basis of CMS in rice (Liu et al., 1989). CMS is presumably controlled by variation in mitochondrial (mt) DNA (Virmani, 1994; Mignouna et al., 1987; Wang et al., 1987). Huang et al. (2006) characterized the diversity of rice CMS cytoplasm and the mechanism of CMS using RFLP markers. They

analyzed the sterile (A) and maintainer lines (B) of nine CMS sources that have been widely used in commercial production in China. The results showed that mitochondrial differences were detected between A and B lines and within different A lines.

CMS is broadly categorized into three types namely, CMS-WA, CMS-HL, and CMS-BT based on inheritance, morphology of abortive pollens and restoration-maintenance relationships (Li et al., 2007). However, commercial hybrid rice is almost exclusively based on CMS-WA, accounting for 90% of three-line hybrids in China (Yuan and Peng, 2005) and 100% outside China (Sattari et al., 2008). Moreover, the International Rice Research Institute in the Philippines relies heavily on CMS-WA in the development of rice hybrids. This scenario opens the vulnerability of the rice hybrids to narrowing genetic base due to one common CMS background.

**1.8 Research Objectives**

(1) Investigate the potential of mixed models as implemented in TASSEL and multiple regression as implemented in SAS GLMSelect to identify markers associated with complex traits and to identify new epistatic regions that play important role(s) in observed phenotypic variation.

(2) Evaluate the non-linear SVR technique for ability to generate high accuracy and power for candidate markers associated with three agronomic traits in rice.

(3) Evaluate potential of selected DNA markers to facilitate rapid introgression of aroma and cooking quality traits into elite Louisiana breeding lines with acceptable agronomic traits.

(4) Conduct a genetic analysis of pollen sterility/male sterility in a single $F_2$ population derived from a natural outcross of a red rice biotype with the commercial Louisiana variety Clearfield161.

## 1.9 References


Akaike H (1974) A new look at the statistical model identification. IEEE Trans. Automat. Contrl AC-19:716-723.

Arbelbide M, Yu J Bernardo R (2006) Power of mixed-model QTL mapping from phenotypic, pedigree and marker data in self-pollinated crops. Theor Appl Genet 112: 876-884.

Baker M (2008) Genetics by numbers. Nature 451: 516-518.

Ball RD (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics 159: 1351-1364.

Beló A, Zheng P, Luck S, Shen B, Meyer Dj, Li B, Tingey S, Rafalski A (2008) Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol genet genomics 279:1-10.

Bogdan M, Doerge RW (2005) Biased estimators of quantitative trait locus heritability and location in interval mapping. Heredity 95:476–484.

Bogdan M, Ghosh JK, Doerge RW (2004) Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. Genetics 167:989-999.

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-77.

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Statist Soc 64:641-656.

Bruskiewich RM, Cosico AB, Eusebio W, Portugal AM, Ramos LM, Reyes MT, et al., 2003. Linking genotype to phenotype: The International Rice Information System (IRIS). Bioinformatics 19 (Suppl. 1): 63-65.

Bundock PC, Henry RJ (2004) Single nucleotide polymorphism, haplotype diversity and recombination in the Isa gene of barley. Theor. Appl. Genet. 109(3):543-51.

Casa AM, Pressoir G, Brown P, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Sci 48:30-40.

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society. 363(1491) 557-572.

Comai L, Young K, Till BJ, Reynolds SH, Greene EA, Codomo CA, Enns LC, Johnson JE, Burtner C, Odden AR, Henikoff S (2004) Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. Plant J. 37(5):778-86.

Cordeiro GM, Christopher MJ, Henry RJ, Reinke RF (2002) Identification of microsatellite markers for fragrance in rice by analysis of the rice genome sequence. Mol Breed. 9:245-250.

De Baets B, Haesaert G, Van Bockstaele E (2008) Marker-based screening of maize inbred lines using Support Vector Machine Regression. Euphytica 161:123-131.

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Res. 14(9):1812-9.

Forde BG. Leaver CJ (1980) Nuclear and cytoplasmic genes controlling synthesis of variant mitochondrial polypeptides in male sterile maize. Proc Natl Acad Sci USA 77:418-422.

Hayes P, Szucs P (2006) Disequilibrium and association in barley: Thinking outside the glass. Proc Nat Acad Science 103:18385-18386.

Huang W, Wang L, Yi P, Tan XL, Zhang XM, Zhang ZJ, LI YS, Zhu YG (2006) RFLP analysis of mitochondrial genome of CMS rice. Acta Genetica Sinica 33 (4) 330-338.

Issiki M, Morino K, Okagaki RJ, Wressler SR, Izawa T, Shimamoto K (1998) A naturally occurring functional allele of the rice *waxy* locus has a GT to TT mutation at the 5'splice site of the first intron. Plant J. 15:133-138.

Jin Q, Waters DLE, Cordeiro GM, Henry RJ, Reinke RF (2003) A single nucleotide polymorphism (SNP) marker linked to the fragrance gene in rice (*Oryza sativa* L.). Plant Sci. 165: 359-364.

Kadaru SB, Yadav AS, Fjellstrom RG, Oard JH (2006) Alternative ecotilling protocol for rapid, cost effective single-nucleotide polymorphism discovery and genotyping in rice (*Oryza sativa* L.). Plant Molecular Biology Reporter 24:3-22.

Kadowaki K, Ishige T, Suzuki S, Harada K, Shinjyo C (1986) Differences in the characteristics of mitochondrial DNA below a normal and male sterile cytoplasms of japonica rice. Jpn J Breed 36: 333-339.

Levings CS III, Pring DR (1976) Restriction endonucleases of mitochondrial DNA from normal and Texas cytoplasmic male-sterile maize. Science 193:158-160.

Li S, Yang D, and Zhu Y (2007) Characterization and use of male sterility in hybrid rice breeding. Journal of Integrative Plant Biology 49 (6): 791-804.

Liu Z, Zha S, Zhan Q, Chen Y (1989) Mitochodrial genome translation products and cytoplasmic male sterility in rice. Acta Genetica Sinica 16(1) 19 (Abstract).

Maenhout S, De Baets B, Haesaert G, van Bockstaele E (2007) Support vector machine regression for the prediction of maize hybrid performance. Theor. Appl. Genet.115: 1003-13.

Manicacci D, Kulanddaivelu C, Fourman M, et al., (2009) Plant Physiology 150:506-520.

Mather KA, Caicedo AL, Polato NR, Olsen KM, Mccouch S, Purugganan MD (2007) The extent of linkage disequilibrium in rice (*Oryza sativa* L.). Genetics 177:2223-2232.

Mcharo M, Labonte D, Oard JH, Kays SJ, McLaurin WJ (2004) Linking quantitative traits with AFLP markers in sweet potato using Discriminant Analysis. Acta Horticultura Acta Hort 637:285-293.

Parisseaux B, Bernardo R (2004) In silico mapping of quantitative trait loci in maize. Theor Appl Genet (2004) 109: 508-514.

Piepho HP, Gauch HG (2001) Marker pair selection for mapping quantitative trait loci. Genetics 157:433-444.

Pressoir G, Brown PJ, Zhu W, Upadyahula N, Rochefrod T, Buckler ES, Kresovich S (2009) Natural variation in maize architecture is mediated by allelic differences at the PINOID co-ortholog *barren inflorescence2*. The Plant Journal 58:618-628.

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Genetics 103:18656–18661.

Sattari M, Kathiresan A, Gregorio GB and Virmani SS (2008) Comparative genetic analysis and molecular mapping of fertility restoration genes for WA, Dissi, and Gambiaca cytoplasmic male sterility systems in rice. Euphytica 160:305-315.

Sha XY (2005) Researchers make progress on new aromatic rice varieties. Rice Research Station News. Crowley: Louisiana Agricultural Experiment Station 2: 4.

Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, Wang G, Wang C, Qian L, Li X, Yu QB, Liu HJ, Chen DH, Gao JH, Huang H, Shi TL, Yang ZN (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol. 135(3):1198-205.

Shinjyo C (1975) Genetical studies of cytoplasmic male sterility and fertility restoration in rice (*Oryza sativa* L.) Sci Bull Coll Agric Univ Ryukus 22:1-57.

Simko I, Haynes KG, Jones RW (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 73(4): 2237–2245.

Singh RK, Khush GS, Singh US, Singh AK, Singh S (2000) Breeding aromatic rice for high yield, improved aroma and grain quality. P71-106. In Aromatic rices. Singh RK, Singh US, Khush GS ed. Oxford and IBH Publishing Co. Pvt. LTD. New Delhi. 289 p.

Takashi M (2005) The map-based sequence of the rice genome. Nature 436: 793-800.

Till BJ, Reynolds SH, Weil C, Springer N, Burtner C, Young K, Bowers E, Codomo CA, Enns LC, Odden AR, Greene EA, Comai L, Henikoff S (2004) Discovery of induced point mutations in maize genes by TILLING. BMC Plant Biol. 4(1):12.

Vapnik V (1995) The nature of statistical learning theory, Springer-Verlag, New York, pp. 1-338.

Virmani SS (1994) Monographs on Theoretical and Applied Genetics 22. Springer-Verlag.

Weber A, Clark RM, Vaughn L, Sánchez-Gonzalez de J, Yu J, Yandell BS, Bradbury P, Doebley J (2007) Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp. Parviglumis). Genetics 177(4):2349-59.

Yu J, Pressoir G, Briggs WH, Vroh BI, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208.

Yuan LP (1977) The execution and theory of developing hybrid rice. Chin Agric Sci 1:27-31.

Yuan LP and Peng JM (2005) Hybrid Rice and World Food Security. China Science and Technology Press, Beijing, China.

Zhang N, Xu Y, Akash M, McCouch S, Oard JH (2005) Identification of candidate markers associated with agronomic traits in rice using Discriminant Analysis. Theor Appl Genet 110: 721-729.

Zhang Y, Mao JX, Yang K, Li YF, Zhang J, Huang YX, Shen FC, and Zhang CD (2008) Characterization and mapping of a male-sterility mutant, *tapetum desquamation* (t), in rice. Genome 51:368-374.

Zhang Y, Li YF, Zhang J, Shen FC, Huang YX, and Wu Z (2008) Characterization and mapping of anther advanced dehiscence (t) in rice. J. Genet. Genomics 35:177-182.

Zuo L, Li SC, Chu M, Wang S, Deng Q, Ding L, Zhang J, Wen Y, Zheng A, and Li P (2008) Phenotypic characterization, genetic analysis, and molecular mapping of a new mutant gene for male sterility in rice. Genome 51:303-308.

# CHAPTER 2 EVALUATION OF MIXED MODEL AND MULTIPLE REGRESSION APPROACHES FOR ASSOCIATION GENETICS IN RICE

## 2.1. Introduction

Completion of the rice genome sequencing project (Takashi et al., 2005) will serve as a powerful springboard for functional characterization of rice genes by a variety of methods that include identity and validation of DNA markers associated with complex traits. Standard QTL mapping approaches such as Composite Interval Mapping (Zeng, 1994) require screening of potential parents for polymorphic markers and extensive periods of three to six years for development of segregating or recombinant inbred lines. However, power and precision may be compromised by limited recombination in segregating/recombinant inbred lines and by relatively small sample size of most mapping populations (Flint-Garcia et al., 2003; Beavis, 1998; Kearsey and Farquhar, 1998). Large intermating populations can be developed to enhance recombination rates, but time, labor, and financial investment limit this strategy, particularly when collecting phenotypic data for complex traits in replicated plots. Cross validation methods have been proposed to obtain unbiased estimates of QTL position and effect for marker-assisted selection (Beavis, 1994; Schon et al., 2004; Utz et al., 2000).

### 2.1.1 Association Mapping Procedures in Plants

Association genetics is an alternative strategy to standard QTL methods that is routinely used in human studies (Baker, 2008), and that is gaining support in the plant research community (Hayes and Szucs, 2006). The principal advantage of this approach, generally referred to as "linkage disequilibrium" mapping, is based on the ability to rapidly query informative regions of the genome among unrelated individuals that have generated numerous meiotic events over multiple generations. Linkage disequilibrium studies have been conducted for various marker-trait associations in maize (Belo et al., 2008; Weber et al., 2007), rice (Mather et al., 2007; Wen

et al., 2009), potato (Simko et al., 2006), sorghum (Casa et al., 2008) and wheat (Breseghello and Sorrells, 2006). Methods for association genetics other than LD mapping have also been evaluated. For example, Zhang et al. (2005) investigated a discriminant analysis procedure for rice inbreds while the same procedure identified candidate loci associated with agronomic traits in sweetpotato (*Ipomoea batatas*) (Mcharo et al., 2004).

### 2.1.2 The Mixed Model Procedure in TASSEL

TASSEL is a popular software package used to analyze marker-trait associations in populations that incorporates population structure (K) and kinship (Q) estimates into a mixed model framework to increase power and reduce Type 1 and Type 2 errors (Yu et al., 2006). TASSEL was used recently in association studies of complex traits in barley (Rostoks et al., 2006), potato (Simko et al., 2006), sorghum (Casa et al., 2008) and wheat (Breseghello and Sorrells, 2006).

Mixed models using variance component approaches that account for kinship estimates have been exploited in animal research for over two decades (Henderson, 1984; George et al., 2000). Nagamine and Haley (2001) extended the mixed model of Henderson to detect QTL by interval mapping in animal systems. Parisseaux and Bernardo (2004) developed a mixed model for hybrid crops incorporating effects for general combining ability of markers associated with agronomic traits. Arbelbide et al. (2006) developed a mixed model for self-pollinating plants that accounted for multiple location effects and kinship based on pedigree records. Arbelbide and Bernardo (2006) applied single and multiple marker analyses in the mixed model format for candidate loci and genes associated with bread quality traits in wheat (*Triticum aestivum* L.).

### 2.1.3 Population Structure

Spurious associations between genotype and phenotype caused by population stratification must be detected among unrelated individuals in association studies to reduce Type

I errors. Clustering techniques are one approach to identify stratified populations. For example, the model-based "Structure" software program identifies putative population structure and assigns individuals to subgroups or clusters based on genotype frequencies (Pritchard et al., 2000). Other clustering approaches based on genetic distance include the weighted and unweighted pair-group methods (Sokal and Sneath, 1973). The Ward's method (Ward, 1963) is distinct from all other clustering strategies in that it minimizes the Sum of Squares (SS) of any two hypothetical clusters that can be formed at each step.

### 2.1.4 The Multiple Regression Procedure

Multiple regression with variable selection based on Bayesian Information Criterion (BIC; Schwarz, 1978) and Akaike Information Criterion (AIC; Akaike, 1974), has been investigated to address selection bias present in standard QTL mapping techniques (Bogdan et al., 2004; Bogdan and Doerge, 2005; Piepho and Gauch, 2001; Ball, 2001). Multiple regression with BIC or other information criteria proposes to identify the fewest number of variables to build the optimal predictive model. Multiple regression with variable selection options has been reported to be superior to Composite Interval Mapping in simulated studies (Broman and Speed, 2002). Software programs such as GLMSelect (SAS Institute) can readily implement multiple linear regression with fixed effects and epistatic interactions based on standard F tests or different selection criteria.

### 2.1.5 Significance of Epistatic Interactions

Epistatic interactions between alleles at different loci in rice have been reported to exert considerable influence on different characters such as hybrid vigor (Li et al., 2001; Goodnight, 1999; Yu et al., 1997), cooking quality (Fan et al., 2005), plant height and heading date (Yu et al., 2002), panicle number (Liao et al., 2001) and other complex traits in rice (Cao et al., 2001; Mei et al., 2003). Standard QTL models have therefore been developed to account for epistasis in

rice and other species (Cui et al., 2006; Wan et al., 2006; Cui and Wu, 2005; Bogdan et al., 2004). Recently, Dudley and Johnson (2009) fit a partial least square (PLS) with epistasis and found significant increase in predictive power for identification of DNA markers associated with oil, protein, starch and grain yield in corn.

Although the mixed model procedure for association genetics has been successful in identifying individual QTLs in several crop plants, a genome-wide test for epistasis is not feasible in this method (Pressoir et al., 2009). Therefore, a two-step method was proposed to model epistatic interactions and to characterize allelic variation at the *barren inflorescence2*(*bif2*) in maize (Pressoir et al., 2009). In the step, a mixed model approach by Yu (2006) was used to identify associated effects. For the second step, SAS GLMSelect was used to identify QTL and conduct genome-wide scans for interaction with *bif2*. Similar methods were employed by Manicacci et al. (2009) to identify epistatic interactions between *Opaque2* and *CyPPDK1* on kernel quality traits in maize.

The objective of our research was to investigate the potential of the mixed model as implemented in TASSEL and multiple regression as implemented in SAS GLMSelect to identify markers associated with complex traits and to identify new epistatic regions that play important role(s) in observed phenotypic variation. Results from our study showed that a combined mixed model-multiple regression procedure successfully identified markers within known QTL regions for three agronomic traits. New epistatic loci were also identified that helped explain a majority of the observed variation for the characters evaluated in this study.

## 2.2 Materials and Methods

### 2.2.1 Plant Material, Phenotypic Data Collection and Analysis

A panel of 192 elite rice breeding lines and varieties representing a narrow germplasm base was evaluated in replicated field plot trials in 2000 in Crowley, Louisiana (LA); Beaumont,

Texas (TX); Stuttgart, Arkansas (AR), Stoneville, Mississippi (MS); and Cape Girardeau, Missouri (MO). The germplasm was composed of 52 lines from Arkansas, one from California, 55 from Louisiana, 25 from Mississippi and 58 lines from Texas. Based on grain length, 162 were long grain types, 24 were medium grain and 6 were short grain. All 192 inbred lines were planted from March to April, 2000 in each of the five states listed above in two to four replicated six-row plots, 2.0 m x 1.4 m, in a randomized complete block design. Standard agronomic practices at each location were carried out to minimize weed and insect damage for maximum grain yield. The center four rows of each plot were used to collect data for heading date (days from seedling emergence to panicle emergence from swollen stem or boot), and percent head rice (whole grains/whole grains + broken grains) x 100). Data for amylose content (percentage of starch in rice grain composed of the polysaccharide amylose) were collected from the Texas and Arkansas locations in 2000. Phenotypic data expressed as trait means across replications at each location were obtained from the University of Arkansas Rice Research and Extension Center, Stuttgart, AR. The agronomic data were averaged across replications within each location to compute mean and variances using PROC MIXED, SAS Institute, v. 9.0. ANOVA was performed to determine location and line differences. Least square (LS) means were used to compute significant location and line differences in SAS. Correlation analyses of each trait for all locations were done using PROC CORR in SAS.

**2.2.2 Molecular Marker Analyses**

Microsatellite (SSR) marker data for the 192 lines were obtained from Dr. Thomas Tai, USDA-ARS, UC-Davis, Davis, CA. A total of 97 SSR markers, evenly spaced over the 12 chromosomes at ~ 20 cm intervals, generated a total of 579 alleles with an average of six alleles/locus. Rare alleles at < 0.07 percent were removed from homozygous loci, but heterozygous loci were retained to provide 194 marker alleles at 97 bi-allelic loci for the final

analysis. PROC ALLELE, SAS Genetics, SAS Institute v. 9.1.4, was used to estimate polymorphism information content (PIC) and allelic diversity. Detection of potential population structure was carried out by the "Structure" software program, v. 2 (http://pritch.bsd. Uchicago.edu/ structure.html). Ward's hierarchical clustering of the 192 lines with all 579 marker alleles was performed in PROC CLUSTER, SAS Institute, v. 9.1.4.

### 2.2.3 TASSEL Mixed Model

Mixed model analysis as implemented in TASSEL was performed for three traits, (amylose content (AC), heading date (HD), and head rice (HR) using the 194 bi-alllelic SSR markers. Kinship (Q) was estimated in TASSEL and was incorporated into the mixed model analysis to account for errors associated with familial relatedness. Data for each trait were averaged across replications both within and across locations for a total of 14 different mixed model analyses. Marker-trait associations at P-value < 0.15 were selected. Correlations of observed and predicted values for of AC, HD, HR traits in each locations were computed in Microsoft Excel based on the predicted phenotype output from the Tassel-mixed model. Corresponding graphs were generated in Excel 2007. Heritability estimates were obtained from the TASSEL program.

### 2.2.4 Multiple Regression by GLMSelect Procedure

Multiple regression in this study was carried out in GLMSelect in the following three steps: Step1. Both forward and stepwise selection methods were used with all possible combinations of the CHOOSE, SELECT and STOP options with Bayesian Information Criterion (BIC), Coefficient of Variation (CV), Adjusted $R^2$ (Adjusted $R^2$), or SL (F test) selection criterion = 0.15 (default value), generating a total of 172 different models or combinations. Stepwise and forward multiple linear regression were performed on phenotypic values (y) of the inbred lines as dependent variables and SSR marker alleles $X_1$, $X_2$, ..., $X_{194}$ as independent

variables. Those independent variables producing a test statistic estimate better than the selection criteria values were added to the model. Completion of Step 1 helped reduce dimension or complexity of the data sets because only significant marker trait associations based on selection criteria were included in the models for additional analyses.

Step 2. To reduce Type I errors or false marker-trait associations, selected models from Step 1 were re-run to include a "leave-one-out" validation step (without epistasis) that was accomplished by the "PRESS" criterion in the "stop" option. The model that produced the highest Adjusted $R^2$ value for a given trait with < 30 effects was considered "optimal" in step 2.

Step 3. Selected models from Step 1 were further evaluated to include all possible two-way interaction effects (epistasis) in the model. The models include a "leave-one-out" validation step that was accomplished by the "PRESS" criterion in the "stop" option. Those selected models that produced the highest adjusted $R^2$ value with < 30 effects were considered "optimal" for a given trait.

### 2.2.5 Mixed Model (TASSEL) - Multiple Regression (GLMSelect) Procedure

The mixed model in TASSEL was used in this study to account for possible kinship relationships that may introduce errors in selection of markers effects. However, the mixed model assumes all effects are additive and so cannot model two-way or epistatic interaction effects (Pressoir et al., 2009). Therefore, the selected effects from the TASSEL output were used as a starting point in PROC GLMSelect for multiple regression analysis to identify both main and two-way marker effects associated with agronomic traits evaluated in this study.

## 2.3 Results and Discussion

### 2.3.1 Agronomic Trait Analysis within and Across Locations

Large variations were observed among the rice lines evaluated for AC, HD, and HR in each location (Table 2.1). AC data was obtained from two locations; HD was collected in five

locations while HR data was obtained in four locations. The AC traits showed that the maximum values were larger than the minimum values by 2 to 3-fold (Table 2.1). The mean values and ranges of AC were at near commercial levels for U.S. elite long and medium-medium grain varieties. Means for each location (AR, TX) and averages across locations were essentially identical. Correlation coefficients were 0.97** between AC in AR and TX. Heritability values were surprisingly low, but were similar within and across locations (0.45-0.46).

Table 2.1 Mean, range and heritability estimates for amylose content (AC), heading date (HD), and head rice (HR) among 192 rice lines evaluated in AR, LA, MO, MS, TX, 2000.

| Trait | Location | N | Mean ± SD** | Range | Heritability |
|---|---|---|---|---|---|
| AC | AR | 187 | 19.26 ± 4.10 | 12.1 -26.6 | 0.48 |
| | TX | 190 | 18.66 ± 4.32 | 9.8 – 25.9 | 0.49 |
| | AVG* | 192 | 18.95 ± 4.17 | 11.0 -26.3 | 0.49 |
| HD | AR | 192 | 83.64 ± 4.39 | 70.0 – 98.0 | 0.36 |
| | LA | 192 | 86.91 ± 3.93 | 72.0 -97.5 | 0.36 |
| | MO | 192 | 91.37 ± 4.02 | 76.5 -105.5 | 0.35 |
| | MS | 192 | 83.41 ± 3.94 | 68.5 -93.5 | 0.36 |
| | TX | 192 | 80.09 ± 4.05 | 66.0 -95.0 | 0.36 |
| | AVG | 192 | 85.08 ± 3.72 | 70.6 – 96.2 | 0.36 |
| HR | AR | 192 | 47.00 ± 10.99 | 14.1 – 66.9 | 0.30 |
| | LA | 192 | 64.74 ± 5.23 | 25.3 -72.3 | 0.30 |
| | MS | 192 | 48.74 ± 6.46 | 31.3 -63.1 | 0.31 |
| | TX | 192 | 52.98 ± 5.70 | 26.5 -62.2 | 0.31 |
| | AVG | 192 | 53.36 ± 5.15 | 39.3 -62.7 | 0.31 |

*AVG-average across locations; ** SD-Standard deviation

For the HD trait, an average of 25 days was noted between the minimum and maximum values. The overall mean heading date of 85 was typical for elite lines of the southern U.S. Rice lines appeared to flower earlier in TX (mean=80) compared to other locations while in MO, rice lines flowered the latest (mean=91). Heritability values were essentially identical both within and across locations (0.35-0.36).

Extensive variation for the important HR trait was observed especially for AR data. The maximum values were greater than minimum values by 2 to 4-fold. Only the LA location produced mean HR values of 0.65 that reached the minimum commercial threshold of 0.60. Heritability values for this trait were low as expected and consistent across locations (0.30-0.31).

Analysis of variance showed that rice lines and locations were significantly different for all three traits analyzed (Table 2.2). Least squared means were used to determine location differences in a pair-wise fashion. Results showed that AR and TX exhibited different mean values for AC. Nevertheless, correlation coefficients for this trait were high (0.97-0.99) between AR and TX, AR and AVG and TX and AVG. Additional data from different location-year combinations are needed to confirm initial observations from this study.

Table 2.2 ANOVA results of amylose content (AC), heading date (HD) and head rice (HR) based on fixed effects model.

| Trait | Effects | df | MS | F value | Pr (F) |
|---|---|---|---|---|---|
| AC (%) | Location | 1 | 32.302249 | 58.19 | <.0001 |
| | Line | 191 | 34.628950 | 62.38 | <.0001 |
| HD (day) | Location | 4 | 3566.49375 | 1027.75 | <.0001 |
| | Line | 191 | 69.19895 | 19.94 | <.0001 |
| HR (%) | Location | 3 | 12236.76389 | 313.55 | <.0001 |
| | Line | 191 | 106.70146 | 2.73 | <.0001 |

For the HD trait, all locations were different except for AR and MO (Pr> |t| 0.62). Correlation coefficients between locations were substantial ranging from 0.64-0.92. This shows a moderate effect of location on the time of flowering of rice lines. For HR, all locations were

found to be different ((Pr> |t| 0.03 – 0.0001) which is not surprising, given that HR is known to respond strongly to environmental conditions. Correlation coefficients of HR were poor (0.07-0.81) indicating that head rice of rice lines grown in different locations tend to generate significantly different HR results. All these results show that relatively high levels of trait variation were present and normally distributed among the elite lines chosen for this study.

The mean (0.37) and range (0.01-0.81) of polymorphism information content (PIC) values for the narrow U.S. germplasm in this study were smaller compared to a diverse collection of 95 Asian and African inbred lines (0.50, range 0.00-0.91) reported by Zhang et al. (2005). Nevertheless, the 192 lines from the current research produced greater PIC value than that of a second collection of 123 U.S. inbred lines (0.27) in the same study by Zhang et al. (2005). Mean allelic diversity of the inbred material in our study at 0.40 was considerably lower when compared to a previous study of hybrid rice with a corresponding mean of 0.52 reported by Xu et al. (2002).

**2.3.2 Analysis of Population Structure and Kinship Relationships**

When the model-based Structure program was used, no population stratification was detected in the narrow elite rice germplasm. This result is consistent with the known pedigrees of the 192 lines that consist almost exclusively of tropical japonica, one of the five major subpopulations previously identified in rice (Garris et al., 2005). Moreover, the Ward's clustering results confirmed that the inbred lines chosen for this study represent a single genetic group or collection (results not shown). Because the Ward's method is less computationally demanding than the Structure software, additional comparisons between the two methods for population stratification are warranted. Kinship relationships among lines were estimated using TASSEL. As expected due to narrow genetic base, the majority of lines (69%) exhibited high

kinship relationship, pair-wise relationships (similarity > 0.3) as shown in Figure 2.1.

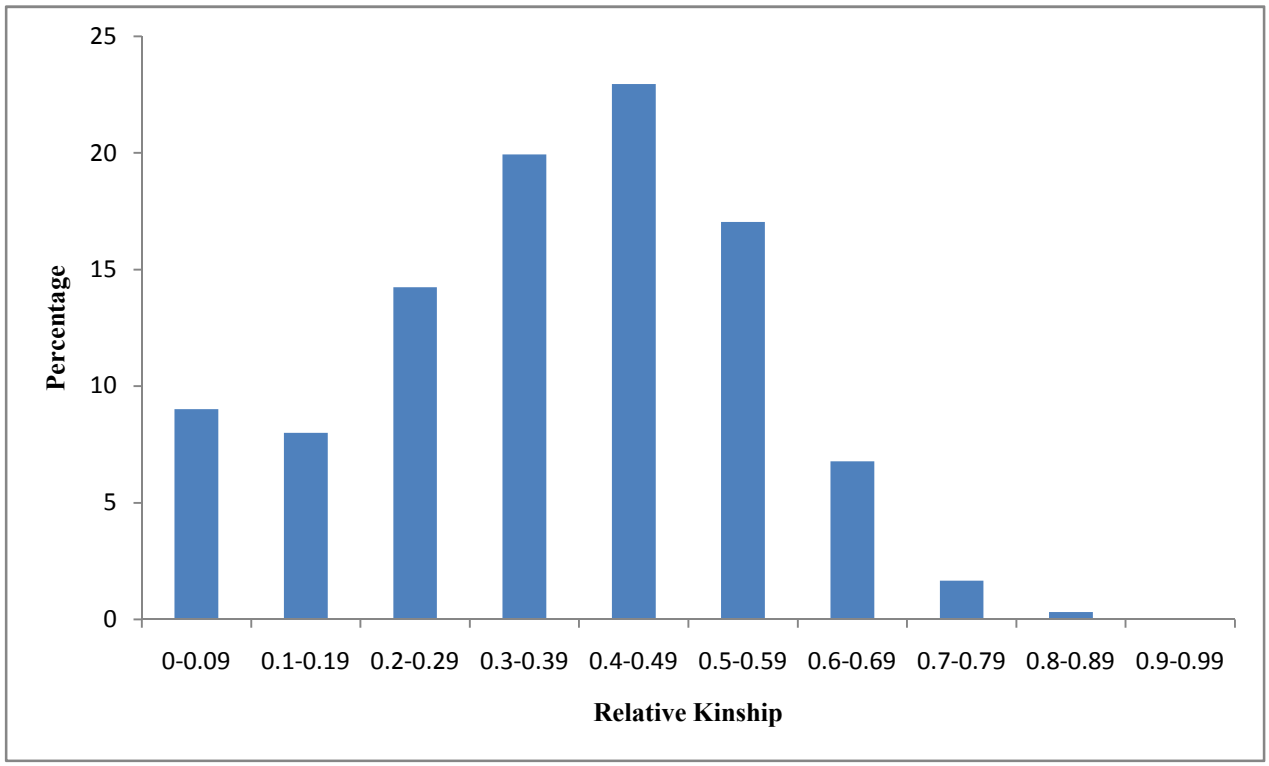



Figure 2.1 Distribution of pairwise relative kinship estimates in the 192 elite rice lines representing a narrow genetic base.

### 2.3.3 Marker-Trait Associations by Mixed Model (TASSEL)

As stated previously, one of our main goals in this study was to evaluate the ability of TASSEL and GLMSelect to identify effects associated with complex traits and to provide genetic insights into factors that control these traits. We chose to study amylose content as a proof-of-concept because much is already known about the genetics and molecular biology of this trait. SSR markers selected by TASSEL to be associated with AC are summarized and presented in Table 2.3. In TX, 35 markers were found to be associated with AC with a total $R^2$ of 0.26, whereas for AR, 29 markers were associated with $R^2$ of 0.23. Across locations, 30 markers were associated with $R^2$ of 0.23. These results demonstrate that TASSEL was poor in selecting markers that explained overall variation observed for AC.

Table 2.3 SSR markers identified by TASSEL mixed model associated with amylose content (AC) for Arkansas (AR), Texas (TX), and across locations (AVG).

| Location | SSR Markers |
|---|---|
| AR | RM190_122, RM169_168, RM408_127, RM510_119, RM214_154, RM116_279, RM202_161, RM435_167, RM3431_150, RM317_161, RM409_091, RM21_157, RM1167_171, RM435_163, RM409_085, RM72_183, RM118_158, RM1189_180, RM420_186, RM5752_126, RM225_138, RM408_119, RM231_181, RM118_162, RM1167_175, RM190_126, RM169_166, RM234_135, RM433_223 |
| TX | RM190_122, RM510_119, RM169_168, RM214_154, RM116_279, RM1167_171, RM408_127, RM190_126, RM21_157, RM409_091, RM3431_150, RM231_181, RM317_161, RM5752_126, RM118_158, RM118_162, RM1189_180, RM435_167, RM437_252, RM225_138, RM169_166, RM72_183, RM202_161, RM420_186, RM409_085, RM72_186, RM149_241, RM3430_211, RM120_184, RM161_181, RM422_389, RM482_192, RM435_163, RM1189_190, RM234_135 |
| AVG | RM190_122, RM510_119, RM169_168, RM214_154, RM408_127, RM116_279, RM1167_171, RM3431_150, RM435_167, RM317_161, RM409_091, RM21_157, RM202_161, RM118_158, RM1189_180, RM231_181, RM5752_126, RM72_183, RM190_126, RM409_085, RM118_162, RM435_163, RM420_186, RM225_138, RM169_166, RM437_252, RM1167_175, RM408_119, RM234_135, RM161_181 |

Several QTL for AC have been reported on almost all of the 12 rice chromosomes, but the majority occurs on chromosome 6. Interestingly, four markers identified both within and across locations by the TASSEL-mixed model on chromosome 6 (RM190, RM435, RM510, RM225) and RM1189 on chromosome 9 mapped to the same regions as previous QTL studies based on the Gramene website (www.gramene.org). These markers were also found by previous association mapping work using discriminant analysis (DA) in the same population (Kadaru, 2006).

Table 2.4 presents the summary of the SSR markers identified by TASSEL associated with HD in each location at P-value < 0.15. LA recorded the most number of significant markers associated with HD with 54 ($R^2$=73%), followed TX at 46 ($R^2$=71%), MS at 42 markers ($R^2$=47%), and AR at 41 markers ($R^2$=58%). MO recorded the least number of 24 significant markers and the lowest $R^2$=0.29. Across locations, 44 markers were associated with HD with a cumulative $R^2$ of 0.57.

Table 2.4 SSR markers identified by mixed model (TASSEL) associated with heading date (HD) for each of the five locations: Arkansas (AR), Louisiana (LA), Missouri (MO), Mississippi (MS), Texas (TX), and across locations (AVG).

| Location | SSR Markers |
|---|---|
| AR | RM517_266, RM403_242, RM517_260, RM132_080, RM214_148, RM403_239, RM477_223, RM3912_195, RM132_083, RM279_164, RM271_086, RM468_266, RM317_161, RM271_098, RM184_215, RM3430_211, RM338_179, RM421_243, RM144_256, RM478_212, RM437_274, RM433_221, RM273_201, RM13_149, RM16_167, RM178_115, RM178_117, RM474_261, RM248_081, RM210_159, RM119_148, RM420_199, RM408_127, RM144_253, RM421_235, RM7_175, RM5_114, RM2_164, RM120_184, RM225_142, RM474_253, RM25_141 |
| LA | RM184_215, RM517_266, RM403_242, RM420_199, RM248_081, RM273_201, RM3912_195, RM132_080, RM251_117, RM190_122, RM403_239, RM468_266, RM421_243, RM144_256, RM3430_211, RM478_212, RM408_127, RM171_328, RM214_148, RM510_119, RM271_086, RM271_098, RM421_235, RM202_176, RM420_186, RM13_149, RM178_115, RM178_117, RM5864_132, RM475_185, RM433_221, RM279_164, RM517_260, RM181_239, RM184_204, RM437_252, RM16_167, RM437_274, RM486_097, RM181_244, RM477_223, RM317_161, RM474_261, RM116_279, RM132_083, RM119_148, RM284_144, RM7_175, RM169_168, RM3431_150, RM229_125, RM348_130, RM413_077, RM112_123 |
| MO | RM214_148, RM13_149, RM190_122, RM144_253, RM486_097, RM144_256, RM478_212, RM420_199, RM510_119, RM403_242, RM403_239, RM475_185, RM149_241, RM315_132, RM248_081, RM517_266, RM132_080, RM317_161, RM431_254, RM184_204, RM181_239, RM184_215, RM25_141, RM234_135 |
| MS | RM184_215, RM403_242, RM420_199, RM3912_195, RM132_080, RM184_204, RM144_256, RM214_148, RM517_266, RM248_081, RM271_098, RM3430_211, RM517_260, RM437_274, RM3431_150, RM120_184, RM317_161, RM420_186, RM403_239, RM433_221, RM202_176, RM459_060, RM273_201, RM5_114, RM477_223, RM171_344, RM119_148, RM623_334, RM178_115, RM178_117, RM468_266, RM181_244, RM190_122, RM459_064, RM132_083, RM171_328, RM478_212, RM421_243, RM25_141, RM5864_129, RM13_149, RM271_086 |
| TX | RM184_215, RM403_242, RM517_266, RM403_239, RM132_080, RM271_086, RM478_212, RM408_127, RM132_083, RM517_260, RM214_148, RM477_223, RM273_201, RM468_266, RM3430_211, RM178_115, RM178_117, RM271_098, RM421_243, RM437_274, RM210_159, RM16_167, RM317_161, RM3912_195, RM421_235, RM420_199, RM119_148, RM474_253, RM181_244, RM251_117, RM171_328, RM3431_150, RM248_081, RM413_079, RM433_221, RM13_149, RM5864_132, RM2_164, RM231_181, RM25_141, RM486_097, RM279_164, RM162_240, RM106_287, RM413_077, RM120_184 |
| AVG | RM403_242, RM184_215, RM517_266, RM132_080, RM214_148, RM403_239, RM420_199, RM478_212, RM248_081, RM3912_195, RM517_260, RM477_223, RM271_086, RM271_098, RM144_256, RM3430_211, RM468_266, RM190_122, RM13_149, RM273_201, RM132_083, RM317_161, RM421_243, RM437_274, RM408_127, RM486_097, RM433_221, RM510_119, RM178_115, RM178_117, RM184_204, RM279_164, RM16_167, RM421_235, RM120_184, RM181_244, RM420_186, RM3431_150, RM171_328, RM5864_132, RM7_175, RM144_253, RM181_239, RM25_141 |

Six markers were common to all locations and are potential markers to differentiate populations based on HD. The common markers were RM214, RM478, RM420 on chromosome 7, RM437, RM13 on chromosome 5, and RM403 on chromosome 1.

Based on known published QTL for HD at the Gramene website (http://www.gramene.org), all six markers mapped within QTL regions for this trait. RM214 on chromosome 7 was also found by DA in the same population (Kadaru, 2006). Individual $R^2$ values of these selected markers associated with HD were low (<5%). These results indicate the potential of mixed model to identify markers associated with this trait, although many markers were required to explain only a moderate level of observed variation.

For head rice (HR), Table 2.5 summarizes the SSR markers associated with the trait in each location. The highest recorded number of marker-HR association was the TX location with 47 ($R^2$=59%) followed by LA with 42 markers ($R^2$=0.58), MS with 30 markers ($R^2$=0.36) and AR with 28 associated markers ($R^2$=0.36). Across locations, 33 significant marker trait-associations were found with cumulative $R^2$ of 0.41. Seven SSR markers (RM5, RM104, RM106, RM112, RM481, RM171, and RM120) associated with HR from the TASSEL analyses were also reported for QTL regions at the Gramene website. In addition, five other markers were located near known QTL regions (RM279, RM171, RM228, RM348, and RM234) for head rice. Four of these markers (RM104, RM106, RM481, and RM279) were also identified by DA in the same population (Kadaru, 2006). Although these markers were significantly associated with HR and mapped to the same region as previous QTL, the low $R^2$ values for individual markers may hamper their immediate use for marker-assisted breeding.

Figure 2.2 summarized the results of plotting observed phenotype data for AC, HD and HR against predicted phenotype data obtained from mixed model output in TASSEL using 194 bi-allelic markers.

Table 2.5 SSR markers identified by mixed model (TASSEL) associated with head rice (HR) for each of the four locations: Arkansas (AR), Louisiana (LA), Mississippi (MS), Texas (TX) and across locations (AVG).

| Location | SSR Markers |
|---|---|
| AR | RM315_137, RM3431_150, RM315_132, RM498_211, RM181_244, RM210_159, RM119_148, RM2_164, RM2_148, RM171_328, RM418_283, RM408_127, RM482_192, RM341_142, RM3912_191, RM120_184, RM112_123, RM228_115, RM1359_162, RM104_238, RM338_179, RM333_165, OSR13_094, OSR13_098, RM234_135, RM162_240, RM341_136, RM250_177 |
| LA | RM475_199, RM116_279, RM132_080, RM468_266, RM341_142, RM112_123, RM112_126, RM116_277, RM181_239, RM459_060, RM475_185, RM279_164, RM420_186, RM403_242, RM517_260, RM517_266, RM248_081, RM181_244, RM420_199, RM3912_191, RM474_261, RM277_114, RM5864_132, RM623_350, RM190_122, RM413_077, RM5_114, RM3430_211, RM284_144, RM118_162, RM421_235, RM437_274, RM16_167, RM225_142, RM16_183, RM271_098, RM477_223, RM273_199, RM2_148, RM413_079, RM279_158, RM202_176 |
| MS | RM437_274, RM181_239, RM3912_191, RM231_191, RM3431_150, RM475_199, RM341_136, RM250_177, RM104_238, RM1359_162, RM475_185, RM403_239 , RM481_156, RM120_182, RM104_222, RM106_293, RM251_119, RM181_244, RM341_142, RM418_283, RM72_186, RM5864_132, RM21_139, RM482_192, RM116_277, RM435_167, RM17_157, RM437_252, RM234_141, RM498_211 |
| TX | RM474_261, RM437_274, RM206_131, RM408_127, RM1167_175, RM106_287, RM119_148, OSR13_094, OSR13_098, RM109_095, RM418_298, RM2_164, RM517_260, RM1189_190, RM16_183, RM296_119, RM413_077, RM481_156, RM277_114, RM420_199, RM623_334, RM475_199, RM181_239, RM142_237, RM149_241, RM316_196, RM116_279, RM5_114, RM437_252, RM348_139, RM316_212, RM112_123, RM116_277, RM21_139, RM296_125, RM413_079 , RM149_240, RM162_240, RM408_119, RM279_164, RM287_103, RM482_186, RM433_223, RM106_293, RM273_201, RM341_142, RM104_222 |
| AVG | RM475_199, RM341_142, RM437_274, RM181_239, RM3912_191, RM408_127, RM481_156, RM21_139, RM315_137, RM112_123, RM482_192, RM104_238, RM3431_150, RM171_328, RM225_142, RM116_279, RM1189_190, RM418_283, RM104_222, RM408_119, RM341_136, RM112_126, RM119_148, RM2_164, RM437_252, RM162_240, RM287_103, RM1359_162, RM315_132, RM316_212, RM475_185, RM234_135, RM474_261 |

Similar trends were observed for AC at both TX and AR where high correlation coefficients ($r^2$=0.92) were observed, although extreme phenotypes tended to be both under- and overestimated. For both graphs, low AC phenotypes were overestimated while high AC phenotypes were underestimated. The results suggest the potential of the selected SSR markers by TASSEL for predicting the amylose content at the TX and AR locations, although this outcome is contrasted to the adjusted $R^2$ values obtained for the markers selected by TASSEL.

Similar trends were also observed for HD (Figure 2.2). Correlation coefficients for AR and LA locations for the observed and predicted values were high (>0.83). Similar $r^2$ values for MO, MS, and TX data were also obtained (data not shown). It was noted that few extreme phenotypes tend to have poor prediction compared to the majority of lines close to the mean. Markers tend to overestimate HD phenotypes for early maturity and tend to underestimate the late maturing lines. Overall the model showed relative good predictive ability, but the selected markers may not be useful for predicting extreme phenotypes.

For the complex HR character, correlation coefficients were moderate ($r^2$<0.80). The data presented in Figure 2.2 for LA and AR represents the overall trend for all locations. It is interesting to note the contrasting trend in the observed and predicted values for LA and AR. While the majority of LA data were detected on the upper section of the graph, AR data were more evenly distributed. It can be noted though that except for a few low HR phenotypes for LA which were poorly predicted, the majority were found near the ideal fitted line. In contrast, AR data showed that overestimation by the markers was evident on low HR phenotypes and underestimation of high HR phenotypes. These low and high HR phenotypes constitute the majority of the 192 lines evaluated resulting in poor prediction for this trait. Either more molecular markers are needed or replicated phenotype data to further improve prediction ability for this trait.

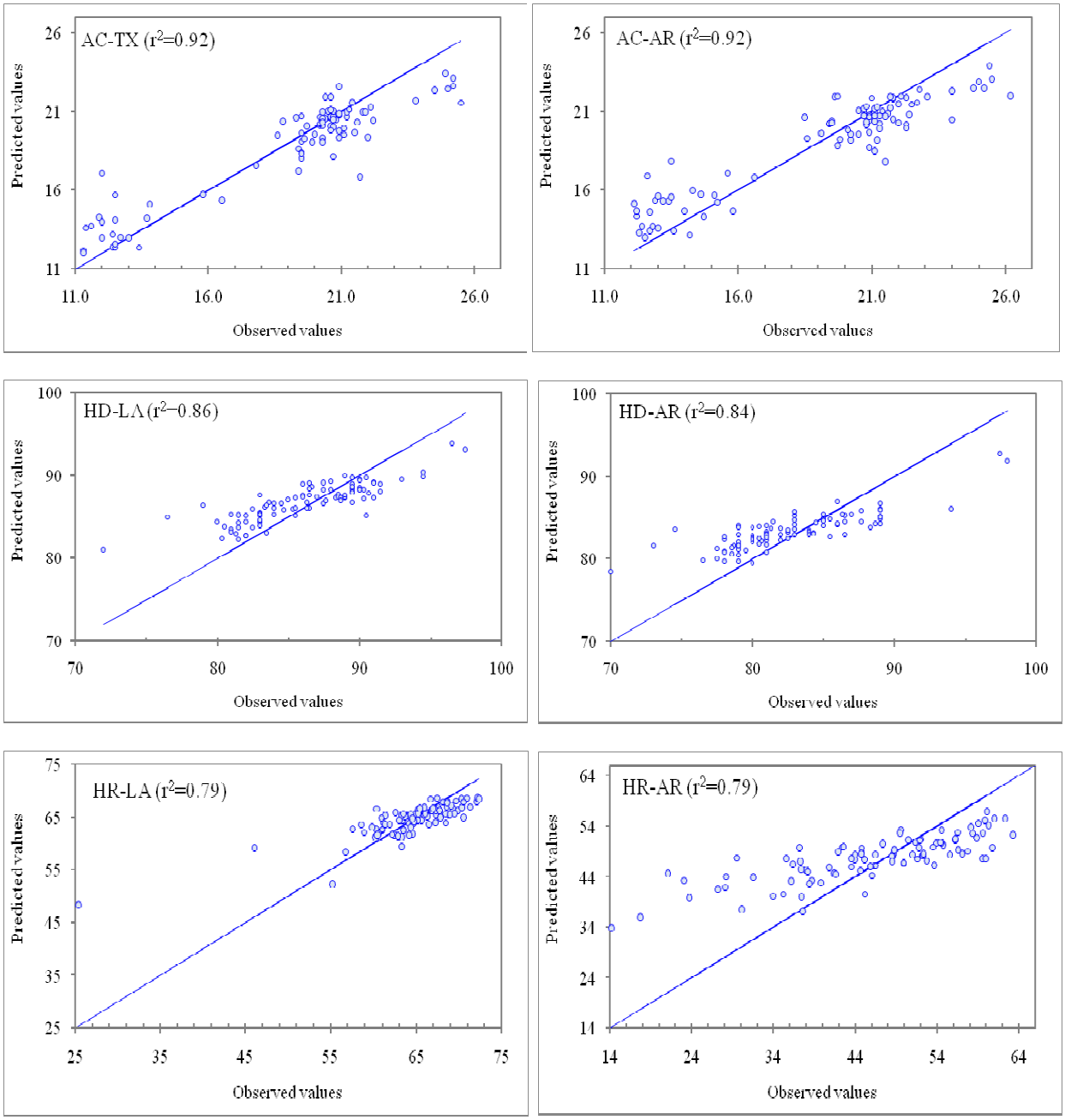



Figure 2.2 Observed and predicted values of amylose content (AC), heading date (HD) and head rice (HR) for Arkansas, (AR), Louisiana (LA) and Texas (TX). Predicted values were based on 194 bi-allelic markers from TASSEL-mixed model analysis.

The overall TASSEL results suggested that while the mixed model could identify individual candidate DNA markers associated with complex traits, the mixed model seemed inadequate to capture total observed phenotypic variation. This may occur because the mixed model only evaluates additive markers one at a time, and ignores multiple regression and two-

way interaction options that may improve identification of markers associated with agronomic traits.

### 2.3.4 GLMSelect Procedure for Association Mapping in Rice

GLMSelect analyses with 172 different combinations of selection criteria, with and without a validation step or consideration of epistasis, revealed two general outcomes. The first was that only moderate amounts of variation could be explained for any of the three traits (adjusted $R^2 \sim 0.50$ to 0.70), no matter the number of selected effects, when validation and epistasis were ignored (results not shown). Second, all variation could be explained (Adjusted $R^2 = 1.00$) for all traits with validation and epistasis where selected effects ranged from 73 to 84, but the large number of effects was impractical for genomics and marker-assisted selection. We therefore considered "optimal models" to be those exhibiting the highest adjusted $R^2$ values with < 30 selected effects.

As shown in Table 2.6, a large proportion of variation in AC could be explained (Adjusted $R^2 = 0.91$) by GLMSelect when validation and epistasis were ignored, but the number of selected effects at 34 was considered too high. When validation was performed and epistasis was ignored, only a moderate Adjusted $R^2$ value of 0.70 could be generated with the maximum number of selected effects which was 13. A high adjusted $R^2$ value of 0.94 with 23 effects was identified as the "optimum" model from GLMSelect when both validation and epistasis were combined with the model options of CHOOSE = Adjusted $R^2$; SELECT = Adjusted $R^2$; stop = PRESS. The smallest values for Root MSE, BIC, AIC, ASE, and PRESS were also obtained when a validation step and epistasis were considered. In contrast, the standard F test implemented in GLMSelect (CHOOSE = none; SELECT = SL (p = 0.15); STOP = none) produced only a moderate adjusted $R^2$ value of 0.62 with a maximum of 16 effects selected.

Analysis of HD by GLMSelect showed that validation and epistasis provided only a small advantage in terms of variation explained and number of effects (Table 2.6). However, a consistent trend for smallest values of BIC and the other criteria was observed with consideration of validation and epistasis. All GLMSelect results with or without validation or epistasis were clearly superior to standard F tests that required 24 effects to explain only 58 % of the observed variation for this complex agronomic trait.

When HR was evaluated, a large number of effects (34) were required to explain a moderate amount of variation (0.84) if validation and epistasis were ignored (Table 2.6). After a validation step was included, the number of selected effects dropped three-fold to 11, but the percent variation explained was poor at 53 %. When both validation and epistasis steps were implemented under selection options of CHOOSE = Adjusted $R^2$, SELECT = Adjusted $R^2$, STOP = PRESS, a high adjusted $R^2$ value of 0.94 was obtained with 29 effects. The same "optimal" model was identified when BIC was chosen for the SELECT option. Similar to results for amylose content and heading date, validation and epistasis steps produced the smallest values for the criteria Root MSE, BIC, AIC, ASE, and PRESS. As was the case for the other two traits, only a moderate amount of variation for HR was accounted for by the standard F test (0.64) that selected 21 effects.

To examine the attributes of the GLMSelect procedure in more detail, coefficient values of selected effects as a function of when they entered the "optimal" regression models for the three traits are shown in Figures 2.3a, b, and c. In the case of AC, the first major finding was that all selected effects shown in Figure 2.3a were epistatic which has implications for the importance of gene interactions in complex traits. Such interactions should therefore not be ignored in association genetics or even standard mapping studies.

Table 2.6 GLMSelect analysis, with and without consideration of validation and epistasis, for Adjusted $R^2$, Root Mean Square Error (MSE), Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC), Average error sum of squares (ASE) and Predicted Residual Sum of Squares (PRESS) associated with amylose content, heading date, and head rice.

| Fit statistics | Amylose content | | | | Heading date | | | | Head rice | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1[a] | 2[b] | 3[c] | 4[d] | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Adjusted $R^2$ | 0.91 | 0.70 | 0.94 | 0.62 | 0.72 | 0.73 | 0.77 | 0.58 | 0.84 | 0.53 | 0.94 | 0.64 |
| Root MSE | 1.28 | 2.36 | 1.10 | 2.85 | 2.06 | 1.97 | 1.74 | 2.37 | 2.32 | 3.77 | 1.41 | 3.30 |
| BIC | 0.68 | 210.42 | -6.22 | 271.56 | 146.75 | 161.51 | 102.31 | 216.37 | 107.28 | 306.90 | 37.51 | 77.29 |
| AIC | 70.68 | 197.04 | 41.78 | 319.23 | 194.23 | 158.84 | 144.31 | 284.56 | 177.28 | 298.01 | 97.51 | 336.30 |
| ASE | 1.01 | 4.85 | 0.94 | 7.01 | 3.35 | 2.96 | 2.47 | 4.36 | 3.29 | 12.64 | 1.43 | 8.67 |
| PRESS | ----[e] | 864.53 | 165.10 | ---- | ---- | 523.15 | 443.67 | ---- | ---- | 1689.96 | 276.65 | ---- |
| No. of Effects | 34 | 13 | 23 | 16 | 18 | 23 | 20 | 24 | 34 | 11 | 29 | 21 |

[a] No validation step, no interaction effects in the model, [b] Validation step performed, no interaction effects in the model, [c] Validation step performed, interaction effects in the model [d] Standard F test implemented in GLMSelect where select option = SL (p=0.15),[e] No data collected

The RM190*RM435 was the first effect identified by GLMSelect to be retained in the model that explained the greatest amount of variation (40 %) for amylose content. Given that RM190 lies within the waxy gene considered to be the major factor contributing to amylose content (Bao et al., 2006), this result demonstrates that GLMSelect can successfully identify effects with a genetic and biological basis. The RM190*RM435 interaction also produced the smallest negative coefficients (-0.60 to-0.70) throughout development of the model that suggested a strong effect that would, in statistical terms, contribute to a reduction in amylose content. However, RM190 produced a positive interaction with RM25 which in turn interacted in a negative manner with RM72 and RM433. These results indicate that genetic control of amylose content in rice is complex with multiple epistatic effects located not only at the waxy locus, but also at other chromosomal regions as reported here for the first time.

When the standard F test with or without validation or epistasis was carried out in GLMSelect, the RM190 marker in the waxy gene was not identified among the top selected effects. Specifically, when forward and stepwise selection were implemented without validation or epistasis, RM190 was identified at the $10^{th}$ and $19^{th}$ selection steps, respectively, each with small Adjusted $R^2$ values. When validation and epistasis were included during forward and stepwise selection, RM190 was identified as an epistatic component only in the $5^{th}$ or $6^{th}$ selection steps. We interpret these results to mean that the standard hypothesis testing completed in this study appears to be inferior to GLMSelect for ability to identify those effects most strongly associated with observed variation for AC.

The relative contribution or "evolution" of coefficients during development of the optimal model for HD is shown in Figure 2.3b. Consistent with amylose content, all selected effects were epistatic except for RM214 and RM273 whose relative contributions to the overall

model were minimal. The RM190*RM296 interaction explained the greatest amount of phenotypic variation (20%) and showed a strong negative coefficient throughout development of the optimal model. Given the known association with amylose content, it is surprising that RM190 may also be associated with heading date, although the marker has been previously mapped within the published QTL QHd6a on chromosome 6 (accession AQEA240, www.gramene.org), and RM296 has been mapped within published QTL QHd9 on chromosome 9 (accession AQEA279, www.gramene.org).

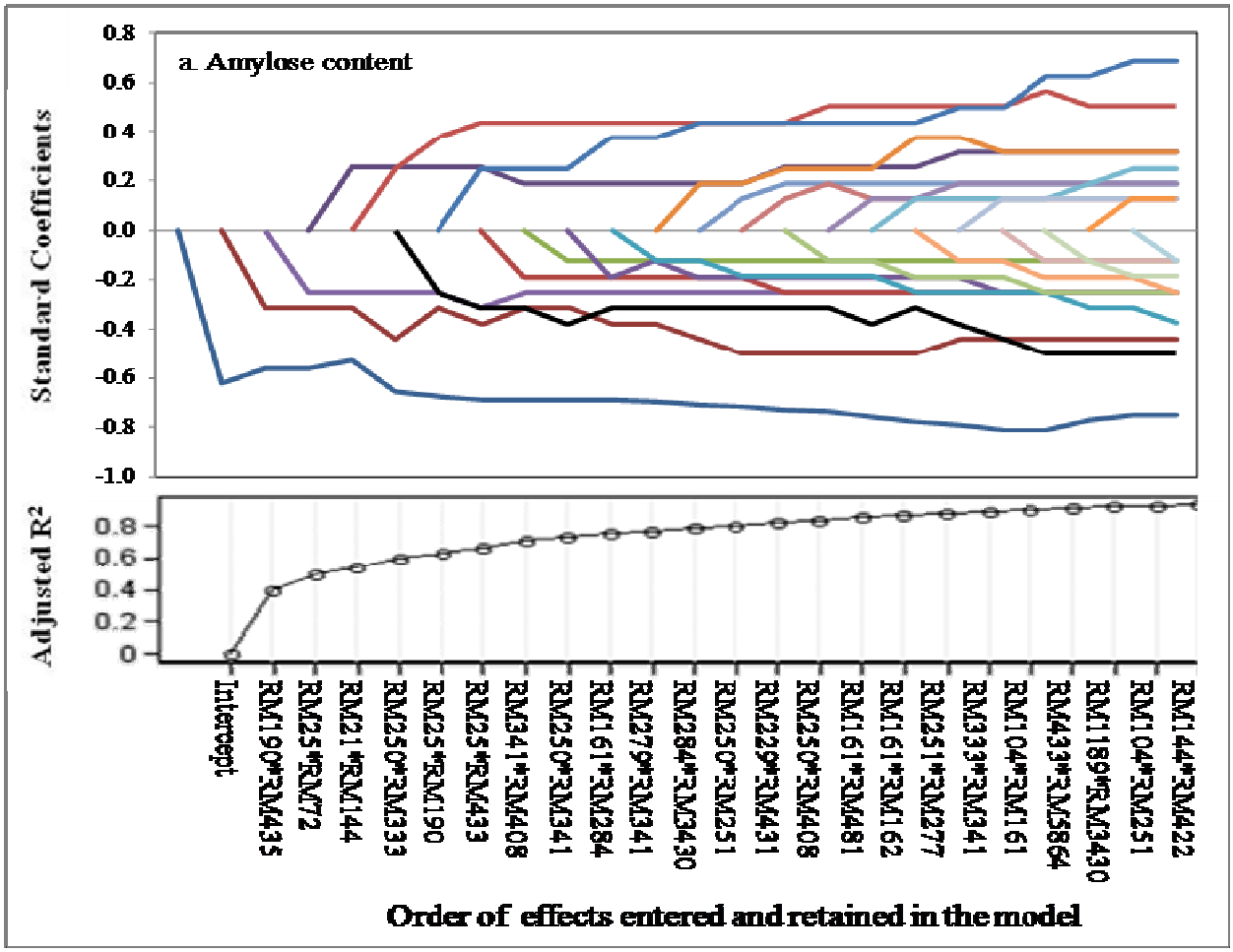



Figure 2.3a Standardized coefficients and adjusted $R^2$ values as a function of when effects are selected and retained by GLMSelect during development of "optimal" model for amylose content.

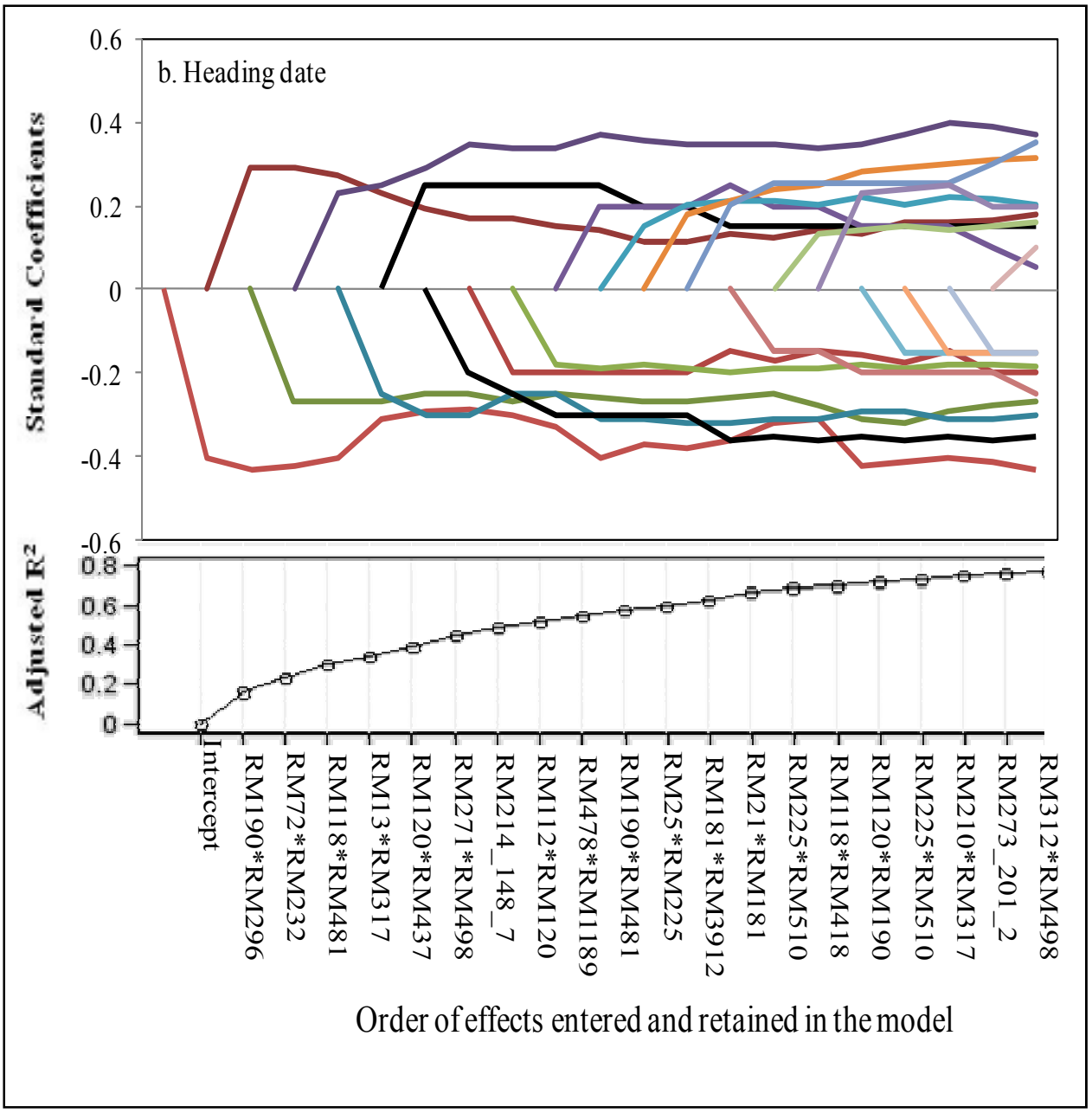



Figure 2.3b Standardized coefficients and adjusted $R^2$ values as a function of when effects are selected and retained by GLMSelect during development of "optimal" model for heading date.

We interpret these results to mean that these loci may lie within regions that interact with each other to affect heading date, but additional studies are needed for confirmation. In addition

to RM190, markers RM25 and RM72 were identified for both heading date and amylose content which suggests that these loci may occur in regions that exhibit pleotrophic effects on these traits. All remaining markers for heading date, except RM21, RM112, RM317, and RM498, mapped within published QTLs that are presented in Figure2.4a and b. The results for HD illustrate the potential value of the GLMSelect procedure for marker-trait analysis where, in this case, a combination of both positive and negative effects contributes to the final predictive model. Selected effects and their relative importance to the optimal GLMSelect model for HR are shown in Figure 2.3c. All effects associated with this trait were epistatic which followed the general trend for the other two traits. These results underscore the need to account for epistatic effects when conducting association studies. The effect that explained the most variation (20%) was identified by GLMSelect as a positive interaction between RM106 and RM144. RM106 on chromosome 2 mapped within 3 cM of a published QTL (accession AQEE014, www.grmene.org) for HR whereas RM144 on chromosome 11 was also detected by the Discriminant Analysis procedure as outlined by Zhang et al. (2005; Oard, unpublished results). It is interesting that RM144 was also identified as epistatic for AC (Fig. 2.3a). The second most influential epistatic effect was composed of RM149 and RM408 that were both found associated with head rice among the 192 lines by the Discriminate Analysis procedure (Oard, unpublished results). RM5 and RM210 comprised the third selected epistatic effect. RM5 mapped within the published QTL hr1 on chromosome 1 that explained 18% phenotypic variation in a study by Aluko et al. (2004) which was consistent with results from the current study. Only 12 out of the 29 selected loci mapped to regions previously cited for head rice, but this may be due to relatively few studies reported for this trait which is a challenge to measure accurately under

field conditions. The 17 remaining loci are therefore considered candidate markers associated with head rice which were not identified in previous studies.

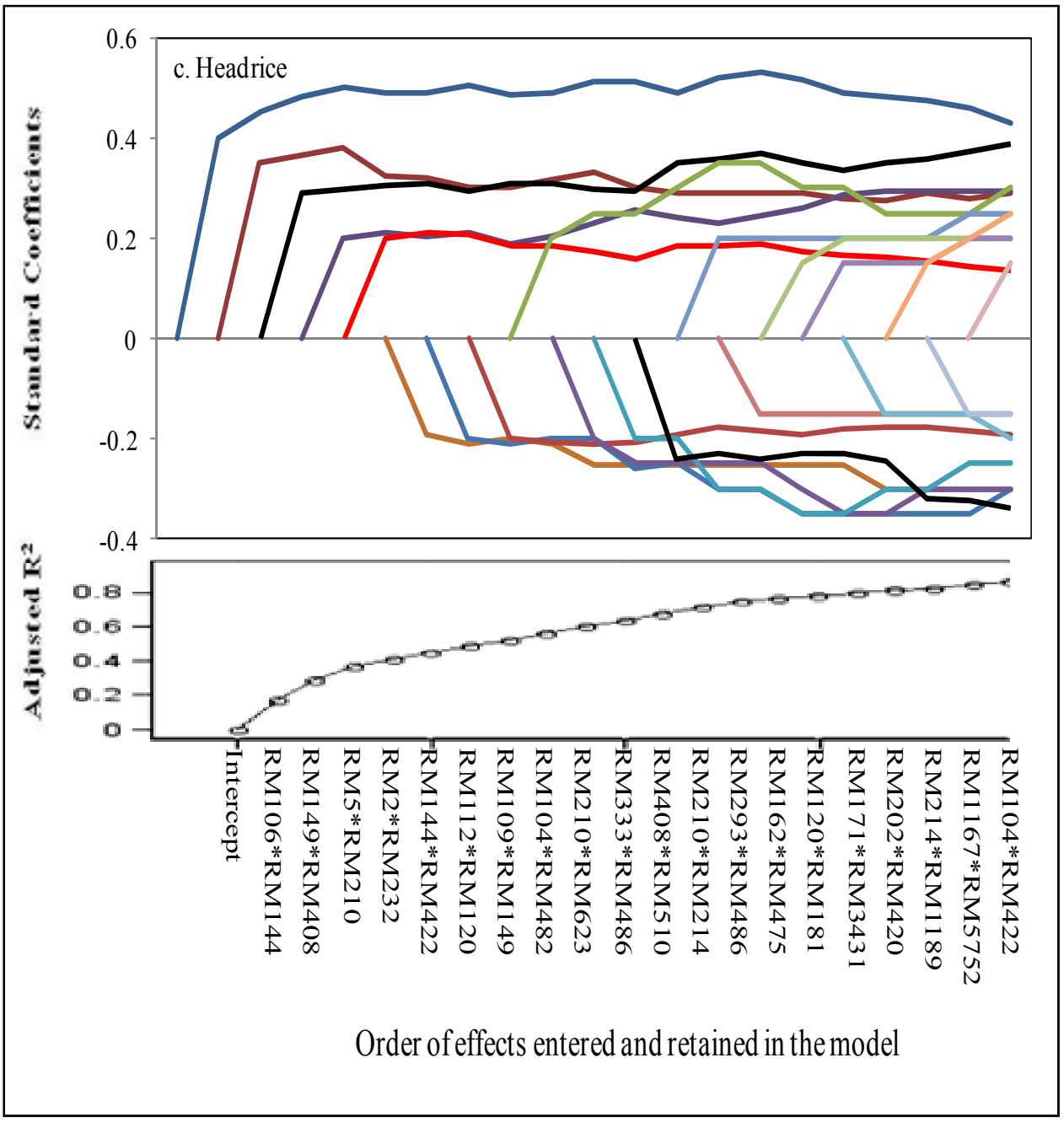



Figure 2.3c Standardized coefficients and adjusted $R^2$ values as a function of when effects are selected and retained by GLMSelect during development of "optimal" model for head rice.

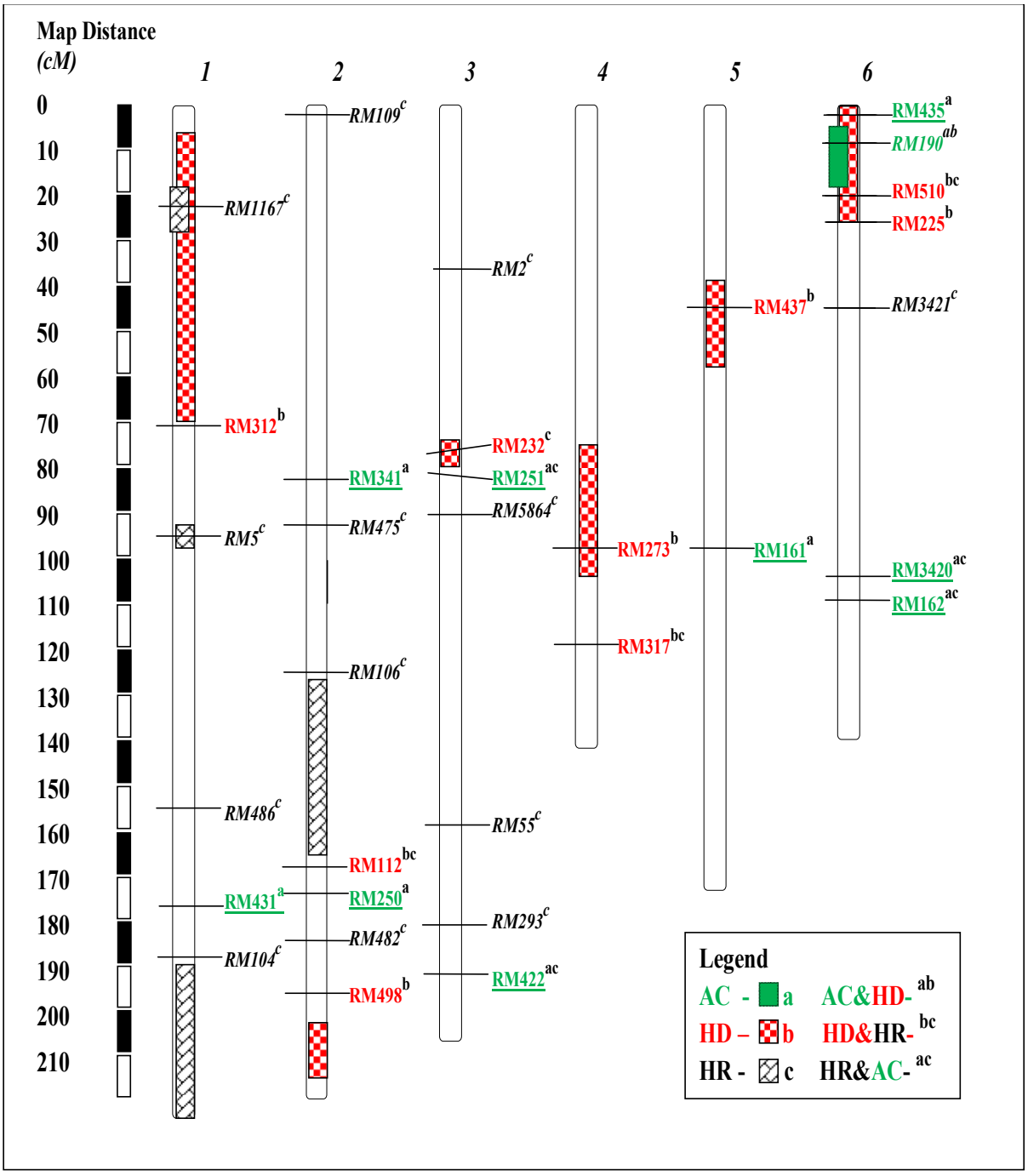



Figure 2.4a Chromosomal locations (1-6) of SSR markers identified by model selection approach for amylose content (AC), heading date (HD), and head rice (HR). Solid and striped boxes inside the chromosomes represent QTL regions detected in previous QTL studies. SSR markers in green and bold with an "a" superscript are amylose content, and red with a "b" superscript are heading date markers, and italics and black with a "c" superscript are head rice markers. Markers labeled with ab, bc or ac superscript combinations are associated with two traits.

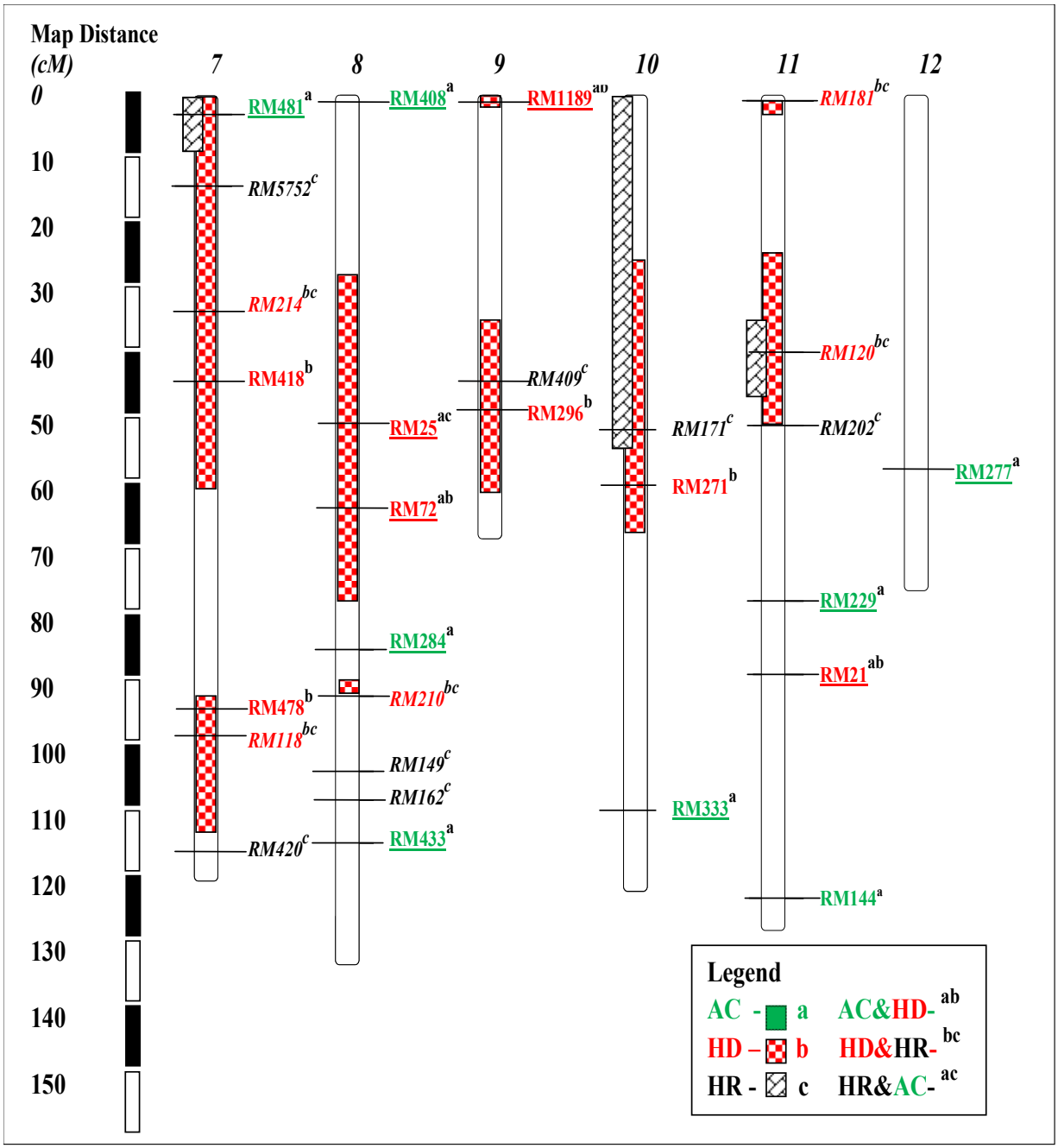



Figure 2.4b Chromosomal locations (7-12) of SSR markers identified by model selection approach for amylose content (AC), heading date (HD), and head rice (HR). Solid and striped boxes inside the chromosomes represent QTL regions detected in previous QTL studies. SSR markers in green and bold with an "a" superscript are amylose content, and red with a "b" superscript are heading date markers, and italics and black with a "c" superscript are head rice markers. Markers labeled with ab, bc or ac superscript combinations are associated with two traits.

### 2.3.5 GLMSelect Analysis for Each Location

Table 2.7 summarizes the results of individual locations for marker-trait association for AC with and without consideration of epistasis. As explained in previous sections on GLMSelect based on data across locations, the addition of an epistatic term in the model improved the ability of markers to explain phenotypic variation in the population. Adjusted $R^2$ increased from 22%-55% for AR data and 60%-94% for TX data. Although the number of selected markers increased from 3 to17, the numbers were still within the acceptable and manageable numbers for applied breeding purposes. Interestingly, the marker effect which is epistatic and found to be the number one marker significantly associated with AC was consistent within and across AR and TX. RM190_122*RM435_167, the best marker for AR accounts for 25% of total variation, while in TX, the same interaction effect accounts for 33%, and across locations it accounts for 40% of total variation explained (Figure 2.3a). RM190 and RM435 are both located on chromosome 6, are known QTLs for AC by previous QTL reports. In addition, 98% of all significant effects in the model for AC were epistatic. These results suggest the potential of multiple regression with selection and validation options and epistasis to identify important QTLs in complex traits like AC.

Table 2.7 Fit statistics of two selection methods (1-GLMSelect without interaction, 2-GLMSelect with two-way interaction) for amylose content within and across locations.

| Fit Statistics/ Location | Arkansas | | Texas | |
|---|---|---|---|---|
| | 1 | 2 | 1 | 2 |
| Adjusted $R^2$ | 0.22 | 0.55 | 0.60 | 0.94 |
| Root MSE | 3.89 | 2.97 | 2.88 | 1.09 |
| BIC | 305.02 | 239.97 | 228.39 | -0.97 |
| AIC | 306.44 | 247.97 | 241.40 | 35.03 |
| ASE | 14.88 | 8.52 | 7.62 | 0.99 |
| PRESS | 1717.24 | 1026.13 | 1164.43 | 163.54 |
| # of Effects | 1 | 3 | 8 | 17 |

HD was analyzed for marker association for AR, LA, MO, MS, and TX locations with and without consideration of epistasis. Table 2.8 presents the summary fit statistics of the analyses. The general trend observed for AC was also observed for HD, *viz.,* adding the epistatic component increased the overall ability of the model to explain trait variation. Almost all associated effects were epistatic (99%) for locations. Adjusted $R^2$ values were > 30% for multiple regression with epistasis compared to multiple regression without epistasis for most of the locations, although more markers were indentified to be associated with HD in all locations. Three selected markers were common across locations (RM214, RM437 and RM13) which were also reported as QTL by Gramene. In addition, RM478 marker was present in all locations except MO and this marker was also identified in Gramene as QTL for HD. Some location-specific markers were also found. For example, RM420 was found only at the MO location, but this marker was also identified in previous work as QTL by Gramene.

Statistics for multiple regression of HR with and without consideration of epistasis are summarized in Table 2.9.Without consideration of epistasis, even with 20 marker effects, adjusted $R^2$ value was still low in AR (adjusted $R^2$=0.57) while in LA no markers were found. Modeling epistatic terms increased overall ability to explain phenotypic variation by >20% with the same number of effects in most locations. It can be noted though that number of effects identified were inconsistent from location to location. For example, LA identified only 3 epistatic effects while AR and MS produced a 10-fold increase and 5-fold more in TX. All effects identified for all locations were epistatic**.** RM5 was found for all locations except LA while RM481 were associated with HR in all locations except MS location. These two markers are known QTLs for HR based on published work (Aluko et al., 2004; www.gramene.org).

Table 2.8 Fit statistics of two selection methods (1-GLMSelect without interaction, 2-GLMSelect with two-way interaction) for heading date within and across locations.

| Fit Statistics/ Locations | Arkansas | | Louisiana | | Missouri | | Mississippi | | Texas | | Average across locations | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| Adjusted $R^2$ | 0.36 | 0.88 | 0.51 | 0.89 | 0.56 | 0.89 | 0.68 | 0.92 | 0.30 | 0.93 | 0.56 | 0.90 |
| Root MSE | 3.60 | 1.69 | 2.68 | 1.30 | 2.73 | 1.36 | 2.22 | 1.13 | 3.35 | 1.06 | 2.47 | 1.18 |
| BIC | 330.75 | 284.49 | 211.62 | 22.68 | 237.93 | 149.69 | 187.98 | -7.05 | 269.49 | -21.37 | 224.83 | 3.22 |
| AIC | 331.43 | 157.84 | 227.04 | 82.68 | 261.09 | 99.72 | 199.43 | 52.95 | 272.29 | 38.63 | 227.72 | 63.22 |
| ASE | 12.03 | 2.17 | 6.45 | 1.23 | 6.11 | 1.39 | 4.03 | 0.94 | 10.82 | 0.82 | 4.94 | 1.03 |
| PRESS | 1754.51 | 471.29 | 860.81 | 191.17 | 1063.68 | 265.70 | 661.72 | 166.06 | 1369.31 | 144.89 | 854.62 | 191.86 |
| # of Effects | 8 | 29 | 10 | 29 | 21 | 29 | 20 | 29 | 3 | 29 | 21 | 29 |

Table 2.9 Fit statistics of two model selection methods (1-GLMSelect without interaction, 2-GLMSelect with two-way interaction) for head rice in each location and across locations.

| Fit Statistics/ Location | Arkansas | | Louisiana | | Mississippi | | Texas | | Average across locations | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| Adjusted $R^2$ | 0.57 | 0.91 | 0.00 | 0.56 | 0.54 | 0.93 | 0.55 | 0.75 | 0.47 | 0.88 |
| Root MSE | 7.52 | 3.41 | 5.79 | 3.84 | 4.45 | 1.76 | 3.64 | 2.70 | 3.89 | 1.89 |
| BIC | 431.07 | 224.97 | 394.29 | 297.08 | 339.75 | 90.07 | 323.81 | 218.79 | 363.33 | 122.46 |
| AIC | 446.28 | 284.98 | 394.30 | 305.08 | 344.46 | 150.07 | 322.15 | 250.79 | 348.90 | 182.46 |
| ASE | 45.33 | 8.35 | 33.20 | 14.19 | 17.31 | 2.25 | 11.45 | 6.29 | 13.15 | 2.71 |
| PRESS | 7444.62 | 1502.63 | 3786.21 | 1639.20 | 2531.54 | 378.59 | 1809.29 | 1007.49 | 2183.87 | 439.67 |
| # of Effects | 20 | 29 | 0 | 3 | 13 | 29 | 15 | 15 | 15 | 29 |

Other known QTLs were location specific. RM104 for example was only detected in LA data while RM171 were significantly associated with HR in MS only. RM120 was found to be associated with HR in MS and AR only, while RM106 and RM104 were not found on any location except when considering across locations.

### 2.3.6 Marker Trait Associations by Combined TASSEL-GLMSelect Procedure

Table 2.10a-c summarizes the common fit statistics of TASSEL-GLMSelect procedure for marker trait association in each location for AC, HD, and HR. The TASSEL-GLMSelect analysis for AC is shown in Table 2.10a. Overall, markers identified at each location could explain the majority of trait variation observed (Adjusted $R^2 > 0.90$). More than 89% of associated effects were found to be epistatic, which highlights the importance of including epistasis in the model as reported by Dudley (2009) in recent maize studies. The list of main and epistatic marker effects is presented in Table 2.11. RM190, previously found to be located within the *waxy* locus and primarily responsible for AC (Bao et al., 2006), was also detected by TASSEL-GLMSelect for the trait in both locations. Six epistatic interactions involving RM190 were found for AR, five for TX and four for AVG. Also, RM435, RM225, and RM1189, identified by TASSEL-GLMSelect and confirmed to be located on known QTL regions by previous Gramene reports, were also found to be associated with the trait in TASSEL. These results indicate the potential of the TASSEL-GLMSelect procedure for association genetics of AC in rice.

For HD, the LA and TX locations identified the greatest number of effects vs. other states (Table 2.10b). Adjusted $R^2$ values for each state ranged from relatively low to moderate with LA and TX producing nearly identical values with an overall average for all states of 0.76. Most of the associated marker effects with HD were epistatic at each location accounting for 98% of total associated effects. Markers for HD found in TASSEL-GLMSelect (RM214, RM478, RM420,

RM537, RM13, and RM403) were also found to be associated with this trait by the TASSEL analysis. RM214 was associated with HD at all locations except AR, while RM13 and RM478 were associated at all locations except MO and MS. Finally, markers RM403 and RM437 were associated with HD in LA and TX.

It is noteworthy that the combined TASSEL-GLMSelect approach for HD produced higher adjusted $R^2$ values with fewer selected effects than the TASSEL method alone. For example, the analysis across locations with TASSEL-GLMSelect selected 24 effects to generate an adjusted $R^2$ value of 0.76 for HD while TASSEL alone required nearly double the number of effects (44) to produce an adjusted $R^2$ value of only 0.57. These results demonstrate the increased power and precision of the combined TASSEL-GLMSelect approach versus exclusive use of the mixed model implemented in TASSEL for marker-trait associations. Results from the remaining statistics in Table 2.10a show the potential value of the combined analysis across locations, in spite of the location effect shown in the ANOVA (Table 2.1). For example, values for Root MSE, BIC, and the remaining statistics were the smallest when the data were analyzed across locations (AVG).

Table 2.10c shows TASSEL-GLMSelect results for HR. Large differences were observed within states to explain observed variation with the fewest number of effects. For example, AR and MS required 2.0 to 2.5-fold greater number of effects to explain the same amount of phenotypic variation compared to LA. In addition, a strong division between TX and the other states was observed for ability of TASSEL-GLMSelect to identify associations between markers and HR. TX produced adjusted $R^2$ values ~ 15% greater than the remaining states that may be due to the greater number of selected effects vs. the other states. Moreover, values of the remaining statistics for TX were smaller than the other states, a trend that showed statistical

consistency for all measures of variation used in this study. In other words, the adjusted $R^2$ values were consistent with those of Root MSE and the other statistics.

Table 2.10a TASSEL-GLMSelect analysis with validation and epistasis within and across locations for Adjusted $R^2$, Root Mean Square Error (MSE), Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC), Average error sum of squares (ASE) and Predicted Residual Sum of Squares (PRESS) associated with amylose content.

| a. Fit Statistics | AR | TX | AVG |
|---|---|---|---|
| Adjusted $R^2$ | 0.91 | 0.92 | 0.92 |
| Root MSE | 1.25 | 1.26 | 1.19 |
| BIC | 98.87 | 90.06 | 82.09 |
| AIC | 89.73 | 86.85 | 80.38 |
| ASE | 1.30 | 1.37 | 1.14 |
| PRESS | 269.48 | 291.55 | 247.12 |
| # of Effects | 24 | 19 | 29 |

AR-Arkansas, TX-Texas, and AVG-Average across locations

Seven SSR markers (RM5, RM104, RM106, RM112, RM481, RM171, and RM120) identified by TASSEL-GLMSelect for HR were previously identified by TASSEL and were previously reported in previous QTL mapping studies (Table 2.11). In addition, four additional SSR markers that were located in the same QTL region for HR (RM279, RM171, RM348, and RM234) were identified by TASSEL-GLMSelect. Most of associated marker effects for HR were found to be epistatic (96%) (Table 2.11). Because few QTL regions have been reported for HR, results from this TASSEL-GLMSelect analysis should provide additional markers and loci worth validating in other populations to increase candidate QTL regions for this trait.

As was shown with the HD results above, the TASSEL-GLMSelect analysis for HR produced greater adjusted $R^2$ values with fewer effects than the TASSEL analysis alone. For example, the TASSEL-GLMSelect method selected 26 effects to produce an adjusted $R^2$ of 0.71 while TASSEL alone required more effects (33) to explain substantially less phenotypic variation (0.41). The HD and HR results demonstrate that while the TASSEL mixed model can

identify individual candidate makers, the approach is insufficient to explain observed overall phenotypic variation for the complex traits evaluated in this study.

Table 2.10b TASSEL-GLMSelect analysis with validation and epistasis within and across locations for Adjusted $R^2$, Root Mean Square Error (MSE), Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC), Average error sum of squares (ASE) and Predicted Residual Sum of Squares (PRESS) associated with heading date.

| b. Fit Statistics | AR | LA | MO | MS | TX | AVG |
|---|---|---|---|---|---|---|
| Adjusted $R^2$ | 0.44 | 0.77 | 0.33 | 0.35 | 0.75 | 0.76 |
| Root MSE | 3.35 | 1.79 | 3.29 | 3.13 | 1.90 | 1.76 |
| BIC | 335.72 | 227.82 | 332.05 | 305.16 | 290.44 | 162.43 |
| AICC | 3.49 | 2.42 | 3.44 | 3.35 | 2.60 | 2.42 |
| AIC | 330.89 | 166.01 | 331.23 | 302.70 | 193.46 | 171.63 |
| ASE | 10.64 | 2.63 | 10.43 | 9.36 | 2.88 | 2.51 |
| PRESS | 1553.28 | 453.09 | 1573.56 | 1318.99 | 585.15 | 453.45 |
| # of Effects | 6 | 21 | 4 | 5 | 26 | 24 |

AR-Arkansas, LA-Louisiana, MO-Missouri, MS-Mississippi, TX-Texas, and AVG-Average across locations within and across locations

Table 2.10c TASSEL-GLMSelect analysis with validation and epistasis within and across locations for Adjusted $R^2$, Root Mean Square Error (MSE), Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC), Average error sum of squares (ASE) and Predicted Residual Sum of Squares (PRESS) associated with head rice.

| c. Fit Statistics | AR | LA | MS | TX | AVG |
|---|---|---|---|---|---|
| Adjusted $R^2$ | 0.63 | 0.65 | 0.68 | 0.71 | 0.71 |
| Root MSE | 6.64 | 3.21 | 3.76 | 2.06 | 2.82 |
| BIC | 610.71 | 332.59 | 428.39 | 306.22 | 312.84 |
| AIC | 580.65 | 336.83 | 426.48 | 216.50 | 324.60 |
| ASE | 35.97 | 9.48 | 12.30 | 3.27 | 6.46 |
| PRESS | 7354.20 | 1583.30 | 2453.78 | 583.13 | 1295.89 |
| # of Effects | 26 | 10 | 19 | 31 | 26 |

AR-Arkansas, LA-Louisiana, MS-Mississippi, TX-Texas, and AVG-Average across locations

Table 2.11 SSR markers identified by TASSEL-GLMSelect associated with heading date (HD), head rice (HR) and amylose content (AC) in each locations: Arkansas (AR), Louisiana (LA), Missouri (MO), Mississippi (MS), and Texas (TX) and across locations (AVG).

| Trait | # of effects | Selected Markers/effect |
|---|---|---|
| AC_AR | 24 | RM190_122*RM214_154 RM116_279 RM190_122*RM435_167 RM169_168*RM317_161 RM214_154*RM409_091 RM21_157 RM317_161*RM435_163 RM409_091*RM409_085 RM1167_171*RM72_183 RM408_127*RM420_186 RM202_161*RM5752_126 RM317_161*RM225_138 RM409_091*RM408_119 RM435_163*RM408_119 RM190_122*RM1167_175 RM1167_17*RM1167_175 RM408_119*RM1167_175 RM435_167*RM190_126 RM72_183*RM190_126 RM225_138*RM169_166 RM234_135 RM435_167*RM234_135 RM190_126*RM234_135 RM225_138*RM433_223 |
| AC_TX | 19 | RM190_122*RM214_154 RM116_279 RM21_157 RM169_168*RM21_157 RM214_154*RM409_091 RM3431_150*RM231_181 RM190_122*RM5752_126 RM214_154*RM118_162 RM190_122*RM435_167 RM118_158*RM202_161 RM420_186 RM225_138*RM149_241 RM116_279*RM161_181 RM21_157*RM161_181 RM190_122*RM435_163 RM116_279*RM435_163 RM1167_171*RM435_163 RM190_126*RM435_163 RM3431_150*RM435_163 |
| AC_AVG | 29 | RM190_122*RM435_167 RM169_168*RM317_161 RM214_154*RM409_091 RM420_186*RM234_135 RM3431_150*RM231_181 RM190_122*RM5752_126 RM409_091*RM408_119 RM408_119*RM234_135 RM408_127*RM21_157 RM3431_150*RM435_163 RM435_163*RM1167_175 RM1189_180*RM435_163 RM116_279 RM1167_171*RM231_181 RM408_127*RM161_181 RM214_154*RM1167_175 RM1167_17*RM1167_175 RM408_127*RM118_158 RM317_161*RM408_119 RM409_091*RM409_085 RM116_279*RM118_162 RM409_085*RM118_162 RM317_161*RM234_135 RM408_119 RM3431_150*RM225_138 RM409_091*RM231_181 RM169_168*RM409_091 RM231_181*RM118_162 RM317_161*RM5752_126 |
| HD_AR | 6 | RM132_080*RM279_164 RM13_149*RM408_127 RM132_083*RM144_253 RM210_159*RM7_175 RM478_212*RM120_184 RM317_161*RM474_253 |
| HD_LA | 21 | RM190_122*RM144_256 RM3912_195*RM478_212 RM420_199*RM171_328 RM3912_195*RM214_148 RM408_127*RM510_119 RM132_080*RM271_098 RM214_148*RM437_252 RM3912_195*RM437_274 RM13_149*RM486_097 RM433_221*RM486_097 RM171_328*RM474_261 RM5864_132*RM116_279 RM184_204*RM132_083 RM251_117*RM119_148 RM403_239*RM284_144 RM202_176*RM7_175 RM251_117*RM229_125 RM3912_195*RM348_130 RM437_252*RM348_130 RM181_244*RM348_130 RM468_266*RM112_123 |
| HD_MO | 4 | RM144_256*RM510_119 RM144_253*RM315_132 RM214_148*RM517_266 RM144_253*RM317_161 |
| HD_MS | 5 | RM3912_195*RM214_148 RM517_260*RM3431_150 RM3430_211*RM120_184 RM184_204*RM420_186 RM144_256*RM190_122 |
| HD_TX | 26 | RM184_215 RM403_239*RM517_260 RM478_212*RM214_148 RM132_080*RM271_098 RM403_239*RM421_243 RM478_212*RM437_274 RM478_212*RM3912_195 RM214_148*RM3912_195 RM132_083*RM119_148 RM3430_211*RM119_148 RM437_274*RM474_253 RM132_083*RM181_244 RM3912_195*RM181_244 RM478_212*RM171_328 RM408_127*RM13_149 RM474_253*RM5864_132 RM171_328*RM231_181 RM273_201*RM25_141 RM210_159*RM486_097 RM231_181*RM486_097 RM171_328*RM279_164 RM210_159*RM162_240 RM474_253*RM162_240 RM433_221*RM106_287 RM251_117*RM120_184 RM248_081*RM120_184 |

Table 2.11 (continued)

| | | |
|---|---|---|
| HD_AVG | 24 | RM214_148*RM478_212 RM132_080*RM271_098 RM478_212*RM144_256 RM214_148*RM3430_211 RM248_081*RM132_083 RM3430_211*RM408_127 RM13_149*RM408_127 RM144_256*RM510_119 RM517_260*RM178_117 RM184_204 RM420_199*RM279_164 RM271_098*RM279_164 RM3912_195*RM120_184 RM317_161*RM181_244 RM3912_19*RM3431_150 RM317_161*RM3431_150 RM420_199*RM171_328 RM433_221*RM171_328 RM181_244*RM7_175 RM478_212*RM144_253 RM3912_195*RM144_253 RM144_256*RM144_253 RM317_161*RM144_253 RM144_253*RM25_141 |
| HR_AR | 26 | RM3431_150 RM315_132*RM498_211 RM181_244*RM210_159 RM181_244*RM119_148 RM181_244*RM418_283 RM315_132*RM482_192 RM181_244*RM482_192 RM181_244*RM341_142 RM408_127*RM3912_191 RM315_137*RM120_184 RM181_244*RM120_184 RM119_148*RM120_184 RM418_283*RM112_123 RM120_184*RM1359_162 RM3431_150*RM104_238 RM482_192*RM104_238 RM1359_162*RM104_238 RM315_132*RM333_165 RM338_179*RM333_165 RM418_283*OSR13_094 RM408_127*RM234_135 RM333_165*RM234_135 RM315_132*RM250_177 RM181_244*RM250_177 RM341_142*RM250_177 RM120_184*RM250_177 |
| HR_LA | 10 | RM132_080*RM279_164 RM116_279*RM403_242 RM420_186*RM623_350 RM341_142*RM190_122 RM277_114*RM190_122 RM3912_19*RM3430_211 RM623_350*RM118_162 RM475_199*RM421_235 RM420_186*RM437_274 RM16_167*RM477_223 |
| HR_MS | 19 | RM475_199*RM250_177 RM3431_15*RM1359_162 RM104_238*RM403_239 RM181_239*RM481_156 RM437_274*RM104_222 RM1359_162*RM104_222 RM341_136*RM106_293 RM341_136*RM251_119 RM481_156*RM341_142 RM3431_150*RM418_283 RM181_244*RM418_283 RM341_142*RM418_283 RM3431_150*RM72_186 RM1359_162*RM72_186 RM120_182*RM72_186 RM418_283*RM72_186 RM418_283*RM482_192 RM475_199*RM435_167 RM104_238*RM234_141 |
| HR_TX | 29 | RM403_242 RM271_086 RM403_239*RM478_212 RM403_239*RM517_260 RM478_212*RM214_148 RM403_239*RM468_266 RM214_148*RM3430_211 RM132_080*RM437_274 RM271_086*RM16_167 RM437_274*RM3912_195 RM132_083*RM420_199 RM437_274*RM474_253 RM3912_195*RM181_244 RM421_243*RM251_117 RM437_274*RM251_117 RM181_244*RM251_117 RM119_148*RM171_328 RM132_080*RM413_079 RM3431_150*RM433_221 RM408_127*RM13_149 RM317_161*RM13_149 RM271_086*RM5864_132 RM171_328*RM231_181 RM273_201*RM25_141 RM420_199*RM279_164 RM271_098*RM162_240 RM433_221*RM106_287 RM3912_195*RM120_184 RM162_240*RM120_184 |
| HR_AVG | 26 | RM481_156*RM171_328 RM225_142 RM437_274*RM225_142 RM3912_191*RM225_142 RM104_238*RM225_142 RM437_274*RM418_283 RM482_192*RM418_283 RM171_328*RM104_222 RM341_142*RM408_119 RM171_328*RM341_136 RM181_239*RM112_126 RM408_127*RM112_126 RM181_239*RM119_148 RM104_238*RM119_148 RM171_328*RM119_148 RM112_123*RM437_252 RM3431_150*RM437_252 RM437_252*RM162_240 RM315_137*RM1359_162 RM181_239*RM315_132 RM104_238*RM315_132 RM1189_190*RM316_212 RM112_123*RM234_135 RM3431_150*RM474_261 RM1189_190*RM474_261 RM104_222*RM474_261 |

The overall results of our study suggest that a multiple linear regression approach such as that carried out in GLMSelect coupled with mixed model effect selection is an appropriate starting point for further research in association genetics of rice. The majority of selected effects mapped to previously published QTLs which increases confidence and value in the combined mixed model-multiple regression strategy. The remaining selected effects point to new candidate regions not detected by previous research. All results demonstrate that methods to detect epistatic interactions will be necessary to identify loci that play pivotal roles in complex agronomic traits in rice. Hypothesis testing by F tests is a common method to establish thresholds for significance in standard QTL mapping and association genetics studies. Our study provides strong evidence that use of selection criteria such as adjusted $R^2$, BIC and AIC will identify fewer effects that explain greater phenotypic variation than standard F tests.

It is now well established in association genetics that stratification must be accounted for if present in selected populations. The model based approach of Pritchard (2000) is often used to detect subpopulations in association genetics studies, but this method assumes random mating that may not be appropriate for inbred species such as rice (Gao et al., 2007). We therefore tested for stratification in our narrow germplasm using both the Structure approach and the Ward's clustering method based on genetic distance. No population stratification was detected either by the Structure or the Ward's method. Future studies are needed to determine if genetic distance clustering techniques with less computational demands would be equally effective in detecting stratification in rice versus model-based strategies.

Results from this study suggest that mixed model-multiple regression approaches that consider epistasis with a validation step may be effective in selecting loci for marker-assisted selection. However, the number of selected effects from this study was too large, and therefore too expensive, for practical breeding programs. Increasing the number of markers for analysis of

this population may increase power and precision to reduce selected markers and associated costs. An important question is which marker platform would be most effective in future association genetic studies in rice. Microsatellites are currently the most popular marker type in rice genetic and evolution studies. Our study showed that while allelic diversity of microsatellites was relatively low in this narrow germplasm, the frequency of rare alleles was very high. Rare alleles were removed from the analysis because they were considered non-informative and as such could only contribute to an increase in Type I errors. The use of bi-allelic SNP markers would help alleviate this problem because they are more prevalent than microsatellite markers, and are amenable to high throughput analysis. We conclude that high density SNP markers coupled with TASSEL-GLMSelect procedures such as those outlined in this study should be further explored for association genetics in rice.

**2.4 References**

Akaike H (1974) A new look at the statistical model identification. IEEE Trans. Automat. Contrl AC-19:716-723.

Aluko, G., Martinez C, Tohme J, Castano C, Bergman C, Oard JH ( 2004) QTL mapping of grain quality traits from the interspecific cross *Oryza sativa* x *O. glaberrima*. Theor Appl Genet 109: 630–639.

Arbelbide M, Yu J Bernardo R (2006) Power of mixed-model QTL mapping from phenotypic, pedigree and marker data in self-pollinated crops. Theor Appl Genet 112: 876-884.

Baker M (2008) Genetics by numbers. Nature 451: 516-518.

Ball RD (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics 159: 1351-1364.

Bao JS, Corke H, Sun M (2006) Microsatellites, single nucleotide polymorphisms and a sequence tagged site in starch-synthesizing genes in relation to starch physicochemical properties in nonwaxy rice (*Oryza sativa* L.) Theor Appl Genet 113:1185–1196.

Beavis WD (1994) The power and deceit of QTL experiments. Proc Annu Corn Sorghum Res Conf 49: 250-266.

Beavis WD, (1998) QTL analyses: power, precision, and accuracy, pp. 145–162 in Molecular Dissection of Complex Traits, edited by A H Paterson, CRC Press, New York.

Beló A, Zheng P, Luck S, Shen B, Meyer Dj, Li B, Tingey S, Rafalski A (2008) Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol genet genomics 279:1-10.

Bogdan M, Doerge RW (2005) Biased estimators of quantitative trait locus heritability and location in interval mapping. Heredity 95:476–484.

Bogdan M, Ghosh JK, Doerge RW (2004) Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. Genetics 167:989-999.

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-77.

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Statist Soc 64:641-656.

Cao G, Zhu J, He C, Gao Y, Yan J, Wu P (2001) Impact of epistasis and QTL x environment interaction on the developmental behavior of plant height in rice (*Oryza sativa* L.). Theor Appl Genet 103:153-160.

Casa AM, Pressoir G, Brown P, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Sci 48:30-40.

Cui Y, Wu J, Shi C , Littell RC, Wu R (2006) Modeling epistatic effects of embryo and endosperm QTL on seed quality traits. Genet Res 87:61-71.

Cui Y, Wu J (2005) Statistical model for characterizing epistatic control of triploid endosperm triggered by maternal and offspring QTLs. Genet Res 86:65-75.

Dudley JW, Johnson GR (2009) Epistatic models improve prediction of performance in corn. Crop Science (49):763-770.

Fan CC, Yu XQ, Xing YZ, Xu CG, Luo LJ, Zhang Q (2005) The main effects, epistatic effects and environmental interactions of QTLs on the cooking and eating quality of rice in a doubled-haploid line population. Theor Appl Genet 110:1445-52.

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Ann Rev Plant Biol 54:357–374.

Garris JA, Tai TH, Coburn J, Kresovich S, McCouch (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631-1638.

Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176:1635–1651.

George AW, Visschler PM, Haley CS (2000) Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. Genetics 156: 2081-2092.

Goodnight C J (1999) Epistasis and heterosis, pp. 59–67 in The Genetics and Exploitation of Heterosis in Crops, edited by J. G. Coors and S. Pandey. American Society of Agronomy, Crop Science Society of America and Soil Science Society of America, Madison, WI.

Hayes P, Szucs P (2006) Disequilibrium and association in barley: Thinking outside the glass. Proc Nat Acad Science 103:18385-18386.

Henderson CR (1984) Application of linear models in animal breeding, Univ. of Guelph, Ontario Kearsey MJ, Farquhar AG (1998) QTL analysis in plants: where are we now? Heredity 80:137-142.

Kearsey MJ, Farquhar AG (1998) QTL analysis in plants: where are we now? Heredity 80:137:42.

Li ZK, Luo LJ, Mei HW, Wang DL, Shu Q Y, Tabien R, Zhong DB, Ying C S, Stansel JW, Khush G S, Paterson AH (2001) Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield. Genetics 158**:** 1737–1753.

Liao CY, P Wu, B Hu, KK Yi (2001) Effects of genetic background and environment on QTLs and epistasis for rice (*Oryza sativa* L.) panicle number. Theor Appl Genet 103:104-111.

Manicacci D, Kulanddaivelu C, Fourman M, et al., (2009) Plant Physiology 150:506-520.

Mather KA, Caicedo AL, Polato NR, Olsen KM, Mccouch S, Purugganan MD (2007) The extent of linkage disequilibrium in rice (*Oryza sativa* L.). Genetics 177:2223-2232.

Mcharo M, Labonte D, Oard JH, Kays SJ, McLaurin WJ (2004) Linking quantitative traits with AFLP markers in sweet potato using Discriminant Analysis. Acta Horticultura Acta Hort 637:285-293.

Mei HW, Luo L J, Ying CS, Wang YP, Yu XQ, Guo LB, Paterson AH, Li Z K (2003) Gene actions of QTLs affecting several agronomic traits resolved in a recombinant inbred rice population and two testcross populations. Theor Appl Genet 107:89–101.

Nagamine Y, Haley CS (2001) Using the mixed model for interval mapping of quantitative trait loci in outbred line crosses. Genet Res 77: 199-207.

Parisseaux B, Bernardo R (2004) In silico mapping of quantitative trait loci in maize. Theor Appl Genet (2004) 109: 508-514.

Piepho HP, Gauch HG (2001) Marker pair selection for mapping quantitative trait loci. Genetics 157:433-444.

Pressoir G, Brown PJ, Zhu W, Upadyahula N, Rochefrod T, Buckler ES, Kresovich S (2009) Natural variation in maize architecture is mediated by allelic differences at the PINOID co-ortholog *barren inflorescence2*. The Plant Journal 58:618-628.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Genetics 103:18656–18661.

Schön CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE (2000) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485-498.

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461-464.

Simko I, Haynes KG, Jones RW (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 73(4): 2237–2245.

Sokal RR and Sneath PHA (1973) Principles of Numerical Taxonomy. W.H. Freeman and Company, San Francisco, 1973. 359 p.

Takashi M et al. (2005) The map-based sequence of the rice genome. Nature 436: 793-800.

Utz HF, Melchinger, Schön CC (2000) Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics 54:1839–1849.

Wan XY, Wan JM, Wan L, Jiang JK, Wang HQ, Zhai JF, Weng HL, Wang CL, Lei JL, Wang X, Zhang X, Cheng ZJ, Guo XP (2006) QTL analysis for rice grain length and fine mapping of an identified QTL with stable and major effects. Theor Appl Genet 112: 1258–1270.

Ward, JH (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58: 236-244.

Weber A, Clark RM, Vaughn L, Sánchez-Gonzalez de J, Yu J, Yandell BS, Bradbury P, Doebley J (2007) Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp. Parviglumis). Genetics 177(4):2349-59.

Wen W, Mei H, Feng F, Yu S, Huang Z, Wu J, Chen L, Xu X, Luo L (2009) Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (*Oryza sativa* L). Theor Appl Genet (published online May 21, 2009).

Xu W, Virmani SS, Hernandez JE, Sebastian LS, Redona ED, and Li Z (2002) Genetic diversity in the parental lines and heterosis of the tropical rice hybrids. Euphytica 127:139-148.

Yu SB, Li JX, Xu CG, Tan YF, Gao YG (1997) Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. Proc Natl Acad Sci USA 94: 9226–9231.

Yu SB, Li JX, Xu CG, Tan YF, Li XH, Zhang Q (2002) Identification of quantitative trait loci and epistatic interactions for plant height and heading date in rice. Theor Appl Genet 104:619-625.

Yu, J, Pressoir G, Briggs WH, Vroh BI, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208.

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136: 1457-1468.

Zhang N, Xu Y, Akash M, McCouch S, Oard JH (2005) Identification of candidate markers associated with agronomic traits in rice using Discriminant Analysis. Theor Appl Genet 110: 721-729.

Zhao K, Aranzana MJ, Kim S, Lister C, Schindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. Plos Genet 3: e4.

# CHAPTER 3 EVALUATION OF SUPPORT VECTOR REGRESSION FOR ACCURACY AND POWER OF CANDIDATE MARKERS ASSOCIATED WITH COMPLEX TRAITS IN RICE

## 3.1 Introduction

### 3.1.1 Association Genetics in Plants

Association genetics has been reported recently as a powerful tool to dissect and identify markers and genes associated with agronomic traits of economic importance. Recent examples include markers associated with iron chlorosis in soybeans (*Glycine max*) that were validated in two separate inbred populations (Wang et al., 2008), and identification of a candidate gene for oleic acid content among maize inbreds corroborated in progeny from a separate biparental cross (Belo et al., 2008). Additional studies have reported selection of candidate markers in association mapping of maize (Yu et al., 2006), Arabidopsis (Zhao et al., 2009), potato (Simko et al., 2006; Malosetti et al., 2007) and barley (Rostoks et al., 2006). The statistical basis for association genetic studies of complex traits in plants has been the general linear model (GLM) that assumes continuous response variables are linearly associated with one or more fixed categorical variables such as DNA marker alleles.

### 3.1.2 Support Vector Regression (SVR)

The general linear regression approach is known nevertheless to be sensitive to noisy data, leading to poor predictive accuracy of new information. The support vector regression (SVR) method was developed by Vapnik (1995) to increase model accuracy and power by approximating the unknown nonlinear relationship between the continuous response variables and corresponding predictors. SVR has gained broad popularity due to its robustness to noise, computational efficiency, and simplicity of the method. Implementation of SVR to study the relationship between maize hybrid and inbred lines has been previously investigated (Maenhout et al., 2007; De Baets et al., 2008).

### 3.1.3 SVR Attributes and Model

Because SVR is not commonly used in plant research, the main features of the model are briefly described. SVR is often referred to as the e-insensitive regression or $\varepsilon$-SVR that is used to find a fitting function that deviates at most $\varepsilon$ from the quantitative response value *y* for each training sample $1 \le i \le n$. For a given training data set, let $x = (x_1, \ldots\ldots x_k)$ with *k* predictors and *y* a continuous response. SVR assumes, like multiple linear regression, that the relationship between the predictors and response variable is given by a deterministic function, denoted by *f*, plus the error (the difference between *y* and *f*); that is, $y = f(x)$ + the error. The function *f* is referred to as the basis function. The task is to find a functional form for *f*, using training data to predict the response of new data. The smaller the error, the more accurate is the prediction of *y*. The simplest form of *f* is the linear function $y = f(x) = w'x + b$, where $w'$ stands for the transpose of the column vector *w* resulting in a row vector, and $w'x$ is the sum of their cross-products. The *w* coefficients are often referred to as weights.

For a given training data set with one input *x* as shown in Figure 3.1, *f* is depicted as the dark straight line that fits the data, where *w* and *b* are the slope and the intercept, respectively. This line is referred to as the decision boundary. SVR allows for errors less than $\varepsilon$, a small positive number. The upper dashed black line is defined as function $y_i = \boldsymbol{w}'\boldsymbol{x}_i + b + \varepsilon$ and the lower dashed line function is defined as $y_i = \boldsymbol{w}'\boldsymbol{x}_i + b - \varepsilon$. The two dashed lines have the same distance from the decision boundary. The value of $\varepsilon$ is shown with the two double-headed vertical arrows. Errors within the space defined by the arrows can be tolerated and so are ignored. All the points within the two dashed lines are not included in SVR analysis and only the points outside are considered. The two dashed lines define a buffer zone or fence for data analysis. The vector *w* is orthogonal to the decision boundary and the two dashed lines. The margin, defined as

the distance between the decision boundary and either of the two dashed lines, is equal to $\frac{\varepsilon}{\sqrt{\|\boldsymbol{w}\|^2+1}}$ and a positive function of $\frac{1}{\|\boldsymbol{w}\|^2}$, where $\|\boldsymbol{w}\|^2 = w'w$ is the norm of the vector *w*. The total width of the buffer zone is shown in Figure 3.1 with the two double-headed arrows that are orthogonal to the decision boundary. Because wider margins yield smaller generalization errors, the two dashed lines are anticipated to be as far apart as possible. The primary goal of SVR is to maximize the margins while allowing for deviations at most $\varepsilon$ from the response value *y* for each training sample $1 \le i \le n$. The idea of SVR can be formulized into the following mathematical optimization problem: Maximize $\frac{1}{\|\boldsymbol{w}\|^2}$ or minimize $\|\boldsymbol{w}\|^2$, subject to $\begin{cases} y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - b \le \varepsilon \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b - y_i \ge \varepsilon \end{cases}$, *i*=1,..*n*. Minimizing the norm is equivalent to minimizing values of the coefficients, which forces the margin to be as large as possible. Given a specified tolerable value of $\varepsilon$, optimizing the above functions results in optimum values of *w* and *b* corresponding to the largest margin.

The above optimization method approximates *y* values with $\varepsilon$ precision. However, in most practical situations, it is impossible to find a linear combination for each observation such that the prediction error is 0. In such cases, errors greater than $\varepsilon$ are tolerated. To cope with larger errors, we define a set of "slack" variables $\zeta = (\zeta_1,...,\zeta_k),\ \zeta^* = (\zeta_1^*,...,\zeta_k^*),\ \zeta_i, \zeta_i^* \ge 0$ as a measure of prediction accuracy for the training examples. Hence, when $\zeta_i = 0$ the training example is predicted with absolute accuracy. The slack variables lead to a revised setting in which the weights *w* are chosen so that *C* is maximized where $C > 0$ is the tuning parameter that determines the trade-off between the small values of *w* (flatness of *f*) and the magnitude of errors larger than $\varepsilon$ that can be tolerated (Vapnik, 1995). In Figure 3.1 the values of $\zeta$ and $\zeta^*$ are shown with the

two double-headed vertical arrows. The optimization problem then becomes: Minimizing

$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*)$, subject to $\begin{cases} y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - b \le \varepsilon + \zeta_i \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b - y_i \le \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \ge 0 \end{cases}$ .

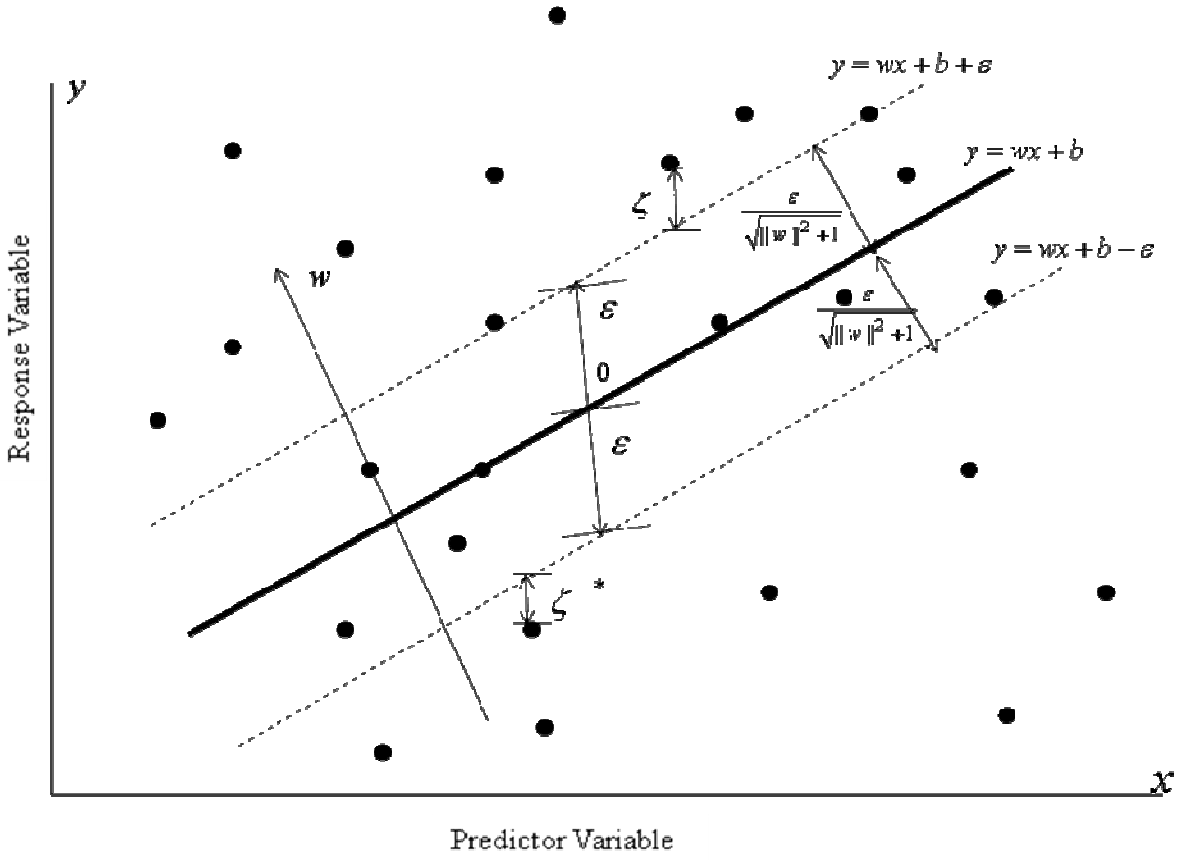



Figure 3.1 Depiction of SVR analysis for training data prediction. The variable *y* is the continuous response variable, and *x* is the explanatory variable. Each data point represents a training sample *i* (*i*=1, …*n*) in the training data set with the observed values as $(x_i, y_i)$. For each training sample, the predicted value, *f*, of the response variable *y* may deviate from the actual value of *y* by some error; that is, *y* = *f*(*x*) + the error. This function for the solid line (decision boundary) is $y = f(x) = wx + b$. The coefficient *w* is the weight and *b* is the intercept. Each of the data points within the two dashed lines has a distance from the decision boundary less than $\varepsilon$; that is, the resulting errors from these points can be tolerated and so are ignored. The margin or distance between the decision boundary and either of the two dashed lines is $\frac{\varepsilon}{\sqrt{\|w\|^2 + 1}}$, where $\|w\|^2 = w'w$ is the norm of the vector *w*. The variables $\zeta = (\zeta_1, ..., \zeta_k)$, $\zeta^* = (\zeta_1^*, ..., \zeta_k^*)$, $\zeta_i, \zeta_i^* \ge 0$ are slack variables to cope with errors larger than $\varepsilon$.

Two primary characteristics of SVR are evident from the above optimization procedure. The first is that SVR is robust to statistical noise due to ignorance of errors less than $\varepsilon$. Second, SVR preserves prediction accuracy and flatness via minimizing weights *w* or maximizing the margin.

The above optimization is derived when the basis function *f* is linear. When *f* is nonlinear, SVR uses specific algorithms referred to as kernel functions for efficient computation. Once the kernel function is specified then *f* is chosen. The kernel function *f* transforms predictors $\boldsymbol{x}_i$ to a higher dimensional space, called the feature space, and then solves the linear optimization problem in the feature space. Linear operation in the feature space, where $\boldsymbol{x}_i$ are located, is equivalent to non-linear operation on $\boldsymbol{f}(\boldsymbol{x}_i)$ in the input space. Three commonly used kernel functions are polynomial, radial basis, and sigmoid. Details of the optimization based on kernels are described by Vapnik (1998).

### 3.1.4 Power and Effect Size Estimation in SVR

The power of a statistical test is the probability that one will reject a false null hypothesis given that the null hypothesis is really false at the chosen significance level. As the type 2 error rate is the probability of failing to reject a false null hypothesis, power also is defined as 1 minus the type 2 error rate.

To compute power, one needs to decide on the effect size (ES), defined as the effect that would be considered to be significant. In the general regression context, the inference about an individual predictor variable $x_k$ centers around whether the corresponding slope $\beta_k$ is 0; that is, the null and alternative hypotheses are $\mathrm{H}_0 : \beta_k = 0, \quad \mathrm{H}_a : \beta_k \neq 0$. The "effect" in this context refers to the deviation of $\beta_k$ from 0 (or the presence of linear association between *y* and $x_k$). Therefore, the ES measures how much $\beta_k$ deviates from 0. The most convenient measure of ES for $x_k$ is based on the correlation coefficient between *y* and $x_k$ denoted by $\rho$. Cohen (1988) considered a $\rho$

value of 0.1 to be a small, 0.3 medium, and 0.5 large effect size. Given the significance level $\alpha$ (e.g. $\alpha$=0.05) and the sample size, the greater the ES the larger the power of a hypothesis test. In common practice, power of 0.80 is generally considered acceptable (Cohen, 1998). We carried out our power analysis based on the ES of $\rho$.

The objective of our research is to evaluate the non-linear SVR method for ability to generate high accuracy and power for candidate markers associated with three agronomic traits in rice. We found that SVR generated relatively high levels of accuracy and power that warrants further investigation for association genetics of complex traits in rice. Moreover, new algorithms for SVR were developed during this study to identify marker variables and epistatic interactions. Finally, validation of the SVR approach was inferred as judged by selected markers corresponding to genetic regions that were identified in previous QTL mapping studies.

## 3.2 Materials and Methods

### 3.2.1 Plant Material and Phenotypic Data Collection

A total of 192 inbred rice lines were planted in 2000 by U.S. public rice breeders as part of the "Uniform Regional Rice Nursery" at Crowley, Louisiana; Beaumont, Texas; Stuttgart, Arkansas; Stoneville, Mississippi; and Cape Girardeau, Missouri. The inbred lines consisted of 52 entries from Arkansas, one from California, 55 from Louisiana, 25 from Mississippi and 58 from Texas. Based on grain length, 162 were long grain types, 24 were medium grain and 6 were short grain. The lines were planted from March to April, 2000 in each of the five states listed above in two to four replicated six-row plots, 2.0 m x 1.4 m, in a randomized complete block design. At each location standard agronomic practices were carried out for maximum grain yield and weed and insect control. The center four rows of each plot were used to collect data for heading date (HD = days from seedling emergence to panicle emergence from swollen stem or boot**)** and head rice (HR = whole grains/whole grains + broken grains) x 100). Amylose content

data (AC = percentage of starch in rice grain composed of the polysaccharide amylose) were collected in 2000 from the Texas and Arkansas locations. Trait means across replications at each location were obtained from the University of Arkansas Rice Research and Extension Center, Stuttgart, AR to compute variances, and correlations between traits using PROC MIXED, SAS Institute, v. 9.1.3.

### 3.2.2 Molecular Marker Analyses

Microsatellite (SSR) marker data for the 192 lines were obtained from Dr. Thomas Tai, USDA-ARS, UC-Davis, Davis, CA. A total of 97 SSR markers, evenly spaced over the 12 chromosomes at ~ 20 cM intervals, generated a total of 579 alleles with an average of six alleles/locus. Rare alleles at < 0.07 percent were removed from homozygous loci, but heterozygous loci were retained to provide 194 marker alleles at 97 bi-allelic loci for the final analysis. Detection of potential population structure was carried out by the "Structure" software program, v. 2 (http://pritch.bsd.uchicago.edu/structure.html). Heritability values for each marker variable were determined and averaged for each trait by the TASSEL software (http://www2.maizegenetics. net/index.php?page=bioinformatics /tassel/index.html).

### 3.2.3 SVR Procedure

The SVR procedure was carried out in the R software package using the *e*1071 library (http://cran.r-project.org/web/packages/e1071/index.html). Optimized values of parameters C and ε were determined by trial and error. To assess the generalization ability of SVR, we considered the *m*-fold cross-validation in which we divided training data into *m* subsets of equal size. Of the *m* subsets, a single subset was retained as the validation data for testing models, and the remaining $m-1$ subsets were used as training data. The cross-validation process was then repeated *m* times (the folds), with each of the *m* subsets used exactly once as the validation data. The *m* results, i.e. the *m* measurements of cross-validation accuracy, were then averaged to

produce a single estimation of cross-validation accuracy. We used four criteria to measure cross-validation accuracy, *viz.* the mean squared error (MSE), R squared ($R^2$), normalized root mean square error (NRMSE) and squared correlation coefficient ($r^2$), computed as follows:

$MSE = \frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2$, the average of squared deviations between predicted value $\hat{y}$ of the response variable and actual value $y$ for the *i*th training example. The smaller the MSE, the better is the prediction accuracy. $R^2 = 1 - \sum_{1}^{n}(y_i - \hat{y}_i)^2 \Big/ \sum_{1}^{n}(y_i - \bar{y})^2$, the percentage of the total trait variation explained by fitting the model. The greater the $R^2$, the better is the accuracy. The normalized root mean square error is computed as $NRMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$; the smaller the NRMSE, the greater is the prediction accuracy. The correlation coefficient between $\hat{y}$ and $y$ is computed as $r^2 = \left[\sum_{1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right]^2 \Big/ \left[\sum_{1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{1}^{n}(y_i - \bar{y})^2\right]$. The larger the $r^2$, the greater is the prediction accuracy.

The $r^2$ values were automatically generated from the SVR package in R. When the *m*-fold cross-validation was specified in the SVR package in R, MSE values were produced to assess cross-validation accuracy on the training data. $R^2$ and NRMSE values were computed by using other functions in the R software package.

It is important to note that the SVR methods used for the previous maize studies (Maenhout et al., 2007; De Baets et al., 2008) did not allow for identification of individual predictor variables or selection of specific interaction terms that may be associated with complex traits. Therefore, we developed a variable selection procedure in SVR to select main and epistatic marker effects associated with continuous or complex traits (see Appendix 1). The SVR variable selection procedure using $R^2$ as the selection criteria is described as follows: (1) The SVR

procedure was carried out for each marker variable, using linear, radial basis nonlinear, polynomial, and sigmoid kernel functions separately. (2) A 10-fold cross validation procedure was conducted to assess cross-validation accuracy. The MSE and $r^2$ values were computed from this step. (3) The $R^2$ value was computed for each individual marker variable in the R software package. (4) Marker variables were ranked according to their $R^2$ values. (5) Steps 1 to 4 were repeated for the combined individual marker variables and their pairwise interactions. (6) Forward variable selection was conducted on the combined selected marker variables from step 5 as follows: First, the predictor variable with the largest $R^2$ value was chosen as the candidate to enter the SVR model. The outside variable with the largest $R^2$ value was the next candidate for entry. The forward selection routine then fitted an SVR model with two marker variables, where the first candidate and the new marker variable were both in the model. The $R^2$ value of the new model then was obtained. If the new $R^2$ value exceeded the previous value by a threshold of 0.0001 which was established to allow selection of up to 35 significant variables, then the second marker was added to the model. Otherwise, the second marker was dropped, and the program proceeded to test for the next marker variable. The forward selection process was terminated when no more predictor variables could be added. The variables selected in the final model were considered as candidate markers associated with the corresponding agronomic trait. In a separate procedure, power calculations for SVR results were carried out using the pwr.r.test in the R package (http://cran.r-project.org/ web/ packages/ pwr/index.html).

### 3.2.4 GLMSelect Procedure

SAS PROC GLMSelect (SAS Institute) was used to identify candidate markers associated with amylose content, heading date, and head rice in a general linear framework. The following steps were carried out (1) All 194 bi-allelic markers were modeled as fixed effects with each agronomic trait considered as a continuous response variable in a multiple linear

regression. We used the following options: selection=forward; choose=ADJRSQ (adjusted $R^2$); select=ADJRSQ (adjusted $R^2$); stop=30. These options identified the top 30 marker effects based on adjusted $R^2$. (2) The selected variables in step 1 were again fitted in a multiple regression form in a forward selection that included main effects and pairwise interactions. The GLMSelect options were identical to those in step 1 except for the stop option that terminated selection at 33 effects for amylose content, 25 effects for heading date and 27 effects for head rice. This action was carried out to establish the same number of selected effects for each trait in GLMSelect as in the SVR approach.

### 3.3 Results

#### 3.3.1 Phenotypic Characterization of the Rice Population

The three agronomic traits in our study produced means, variances, and a range of values that were typical for southern U.S. elite inbred lines. Specifically, mean amylose content values = 19.1%, variance = 13.8, range = 11.0 – 26.2%; heading date mean = 85.1, variance = 14.1, range = 71-96; head rice mean = 53.4, variance = 26.8, range = 39-63. Heading date was moderately associated with amylose content ($\hat{\rho}$ = 0.35, $p < 0.001$), where $\hat{\rho}$ is the estimated correlation coefficient between heading date and amylose content. A small negative association was detected between amylose content and head rice ($\hat{\rho}$ = -0.21, p = 0.002) while no association was detected between heading date and head rice ($\hat{\rho}$ = -0.05, p = 0.224). Mean heritability values calculated by TASSEL across all selected markers were the highest for amylose content (0.48), intermediate for heading date (0.35) and the lowest for head rice (0.29). Population structure analysis using the Structure software revealed that the 192 tropical japonica breeding lines in our study belonged to a single group. This outcome was consistent with a previous structure analysis of rice that showed the japonica group was identified as one of five distinct subpopulations (Garris et al., 2005).

### 3.3.2 Accuracy and Precision of SVR and GLMSelect Procedures

Accuracy of SVR in terms of MSE, $R^2$ and NRMSE values with different kernel functions are presented in Table 3.1. The sigmoid kernel function produced the greatest model variation or lowest accuracy with MSE values for amylose content that were 1.3 to 3.3-fold greater than those of the remaining three functions. This trend of greater MSE values using the sigmoid function was also observed for heading date and head rice. The linear kernel produced the second largest MSE values for all three traits that were 2.5-fold greater than values generated by the polynomial or radial basis method. Considering all four SVR functions, the radial basis kernel produced the greatest accuracy with MSE for heading date and head rice while generating identical precision as the polynomial method for amylose content. Across all kernel functions, the largest MSE values were detected with head rice, and the smallest values were associated with amylose content which is consistent with heritability values computed in this study using the TASSEL software (see above). The three remaining measures of SVR model precision, namely $R^2$, NRMSE, and $r^2$ were also computed. The radial basis and polynomial kernels increased $R^2$ values, consistent with MSE values, by a wide range of 29 to 66 percent across all traits compared to the linear and sigmoid kernels (Table 3.1). Following the same trend, the smallest NRMSE values were observed with radial basis and polynomial kernels. The computed $r^2$ values were virtually identical to those of the corresponding $R^2$ values for both SVR and linear methods across all traits shown in Table 3.1 (data not shown).

We also analyzed the data using a multiple linear regression approach implemented in SAS GLMSelect (see methods). As shown in Table 3.1, accuracy for GLMSelect as measured by MSE was greater than SVR across all traits when computed with linear or sigmoid kernel functions. Markers effects identified were 25 for both AC and HD traits in both SVR and multiple linear regression method while 27 marker effects were found for HR in both analyses.

MSE values with the linear multiple regression approach were, however, 1.8 to 3.5-fold greater across all traits versus SVR with the polynomial or radial basis function.

Table 3.1 MSE, $R^2$ and RMSE values obtained by SVR (using linear, polynomial, sigmoid, and radial basis kernel functions) and multiple linear regression with epistasis for amylose content (AC), heading date (HD), and head rice (HR) across five locations in AR, LA, MO, MS, TX, 2000.

| Trait | Support Vector Regression | | | | | | | | | | | | Multiple Linear Regression | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Linear | | | Polynomial | | | Sigmoid | | | Radial basis | | | SAS GLMSelect | | |
| | MSE | $R^2$ | NRMSE | MSE | $R^2$ | NRMSE | MSE | $R^2$ | NRMSE | MSE | $R^2$ | NRMSE | MSE | $R^2$ | NRMSE |
| AC | 3.78 | 0.72 | 0.53 | 1.50 | 0.89 | 0.33 | 5.03 | 0.63 | 0.61 | 1.50 | 0.89 | 0.33 | 2.68 | 0.85 | 0.39 |
| HD | 8.52 | 0.39 | 0.78 | 1.53 | 0.89 | 0.33 | 9.60 | 0.31 | 0.83 | 1.54 | 0.91 | 0.30 | 5.46 | 0.61 | 0.62 |
| HR | 16.00 | 0.40 | 0.77 | 2.94 | 0.78 | 0.47 | 16.90 | 0.36 | 0.80 | 2.93 | 0.91 | 0.30 | 7.98 | 0.71 | 0.54 |

Similarly, linear multiple regression exhibited a greater ability to explain trait variation as measured by $R^2$ compared to the sigmoid or linear functions. However, prediction of variation for heading date and head rice were reduced substantially by 20 to 30 percent for the linear GLMSelect versus SVR when using the polynomial and radial basis kernels. For amylose content, $R^2$ values of GLMSelect versus the polynomial or radial basis method were essentially identical. NRMSE values from GLMSelect were similar to those using the radial basis kernel for AC, but were substantially larger for HD and HR compared to values obtained by the radial basis kernel.

### 3.3.3 Power Estimation in SVR

The power of SVR as a function of ES, measured by the correlation coefficient $\rho$, was determined for amylose content, heading date and head rice. Shapes of the power curves for all traits were very similar, so results for amylose content are shown as an example in Figure 3.2. In general the results showed power increasing in a strong linear manner when the effect size $\rho$ was set at small to medium values of 0.1 to 0.3, a result that was consistent with "Cohen's rule"

(1988). The SVR models for all three traits produced an "acceptable" power ≈ 0.80 when the ES of $\rho$ was set at 0.20. An increase in $\rho$ > 0.30 had little impact as the power values approached a maximum plateau of ~ 1.0.

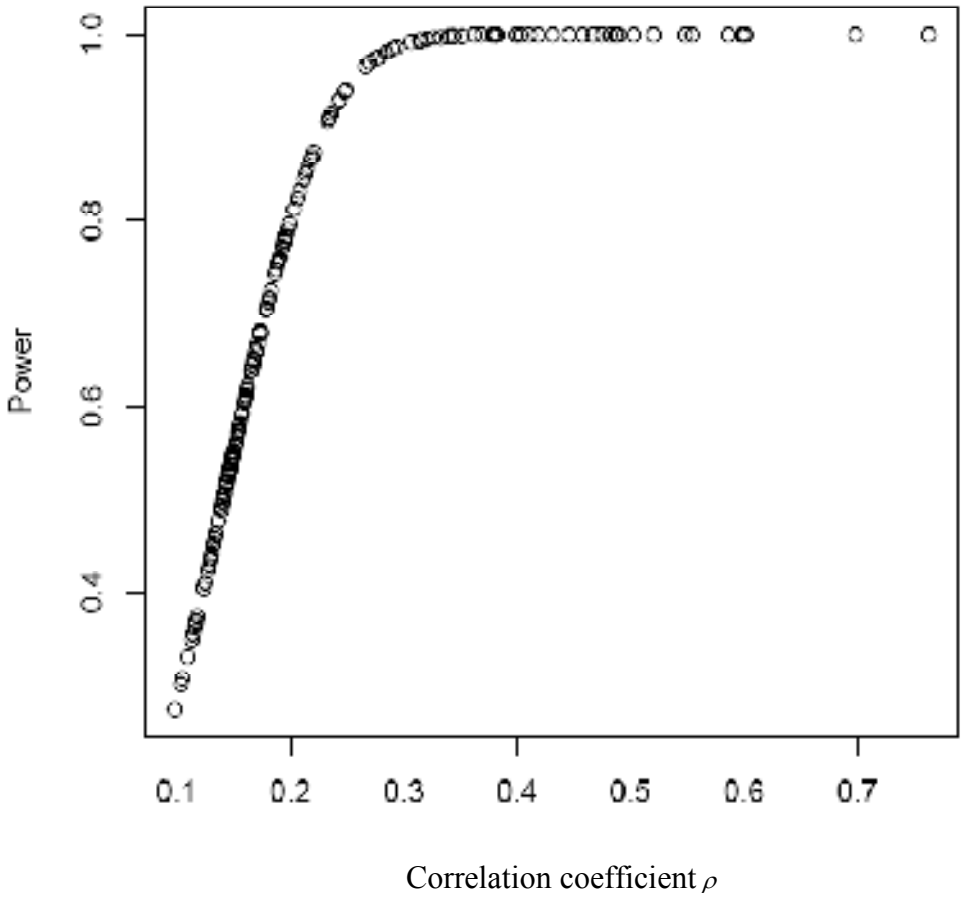
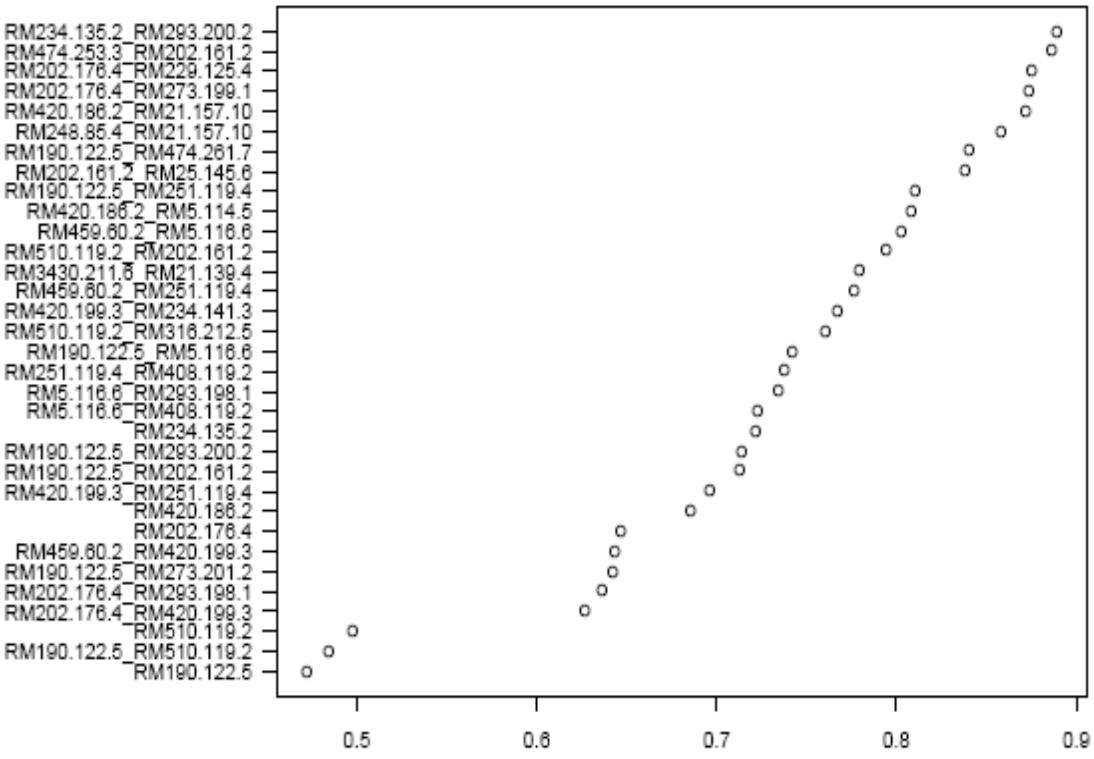



Figure 3.2 Plot of power for optimized Support Vector Regression as a function of correlation coefficient $\rho$ for amylose content (AC).

### 3.3.4 Identification of Marker-Trait Associations

Selected marker alleles from optimized SVR models and their corresponding sequential $R^2$ values for amylose content, heading date, and head rice are presented in Figure 3.3 a-c. All optimized SVR models exhibited high $R^2$ values of ~ 0.90 with 25 to 33 selected main and epistatic effects. As shown in Figure 3.3a, the RM190.122.6 allele was identified as a single main effect that explained the greatest amount of variation for amylose content ($R^2$ = 0.47). The

RM190 locus, located within the granule-bound starch synthase gene, is a well-known microsatellite marker used to classify different levels of amylose content in rice (Ayres et al., 1997). The second selected variable consisted of the RM190.122.6 allele epistatic to RM510.119.2 which mapped within a second QTL for amylose content on chromosome 6 (QTL accession AQBN001, www.gramene.org). The remaining selected markers showed small, incremental effects on total observed variation ($R^2 = 0.91$) which was consistent with previous studies of this trait (McKenzie et al., 1983).




Cumulative $R^2$

Figure 3.3a Selected marker effects from optimized SVR models and corresponding sequential $R^2$ values on the horizontal axis for amylose. The number of total selected variables is 25 and $R^2 = 0.90$.

A total of 25 variables were selected by SVR that explained 90 percent of observed variation for heading date as shown in Figure 3.3b. All marker alleles that comprised the top four epistatic terms were found in previous QTL studies to be associated with heading date. Selected

marker allele RM279.164.6 on chromosome 2 mapped within QTL dth2.1 (accession AQED001, www.gramene .org) and was epistatic to four different alleles (RM132.80.3, chromosome 3; RM3431.150.2, chromosome 6; RM437.274.5, chromosome 5; RM208.147.5, chromosome 2). RM190.122.6 was detected in this study to be associated with both heading date and amylose content which may not be unexpected, given the moderate correlation of 0.35 found in this study between heading date and amylose content. In addition, the RM190 locus mapped within 5 cM of the Hd3a locus, reported to be a major activator of flowering under short day conditions (Tamaki et al., 2007).

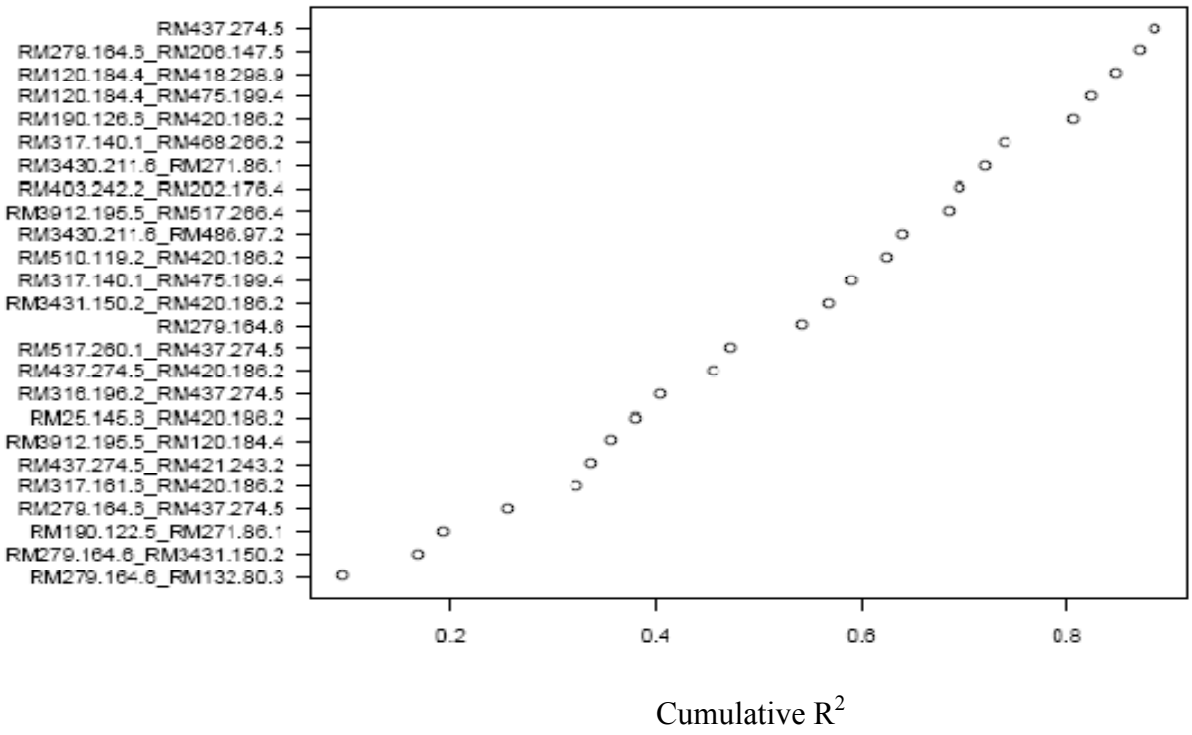



Figure 3.3b Selected marker effects from optimized SVR models and corresponding sequential $R^2$ values on the horizontal axis for heading date. The number of total selected variables is 25 and $R^2 = 0.90$.

The markers associated with head rice by SVR analysis are shown in Figure 3.3c. Selected markers for this important grain quality character were also associated with milling traits in previous QTL mapping studies. For example, RM315.137.2 and RM3912.191.3 were

detected as epistatic alleles contributing the most to head rice ($R^2 = 0.17$). The RM315 locus is located within QTLs previously reported on chromosome 1 for head rice (Septiningsih et al., 2003) and brown rice (Aluko et al., 2004) from interspecific crosses of *O. sativa* x *O. rufipogon*. RM3912 was not detected in previous studies and therefore represents a new candidate locus for head rice. RM476.199.4 identified by SVR on chromosome 1 was mapped previously within the QTL hr1 for head rice (Aluko et al., 2004). RM481.156.6 on chromosome 7 was detected by SVR that also mapped within the QTL mr7 for percent milled rice (Aluko et al., 2004). All but one of the 27 selected variables were epistatic with final $R^2 = 0.89$.

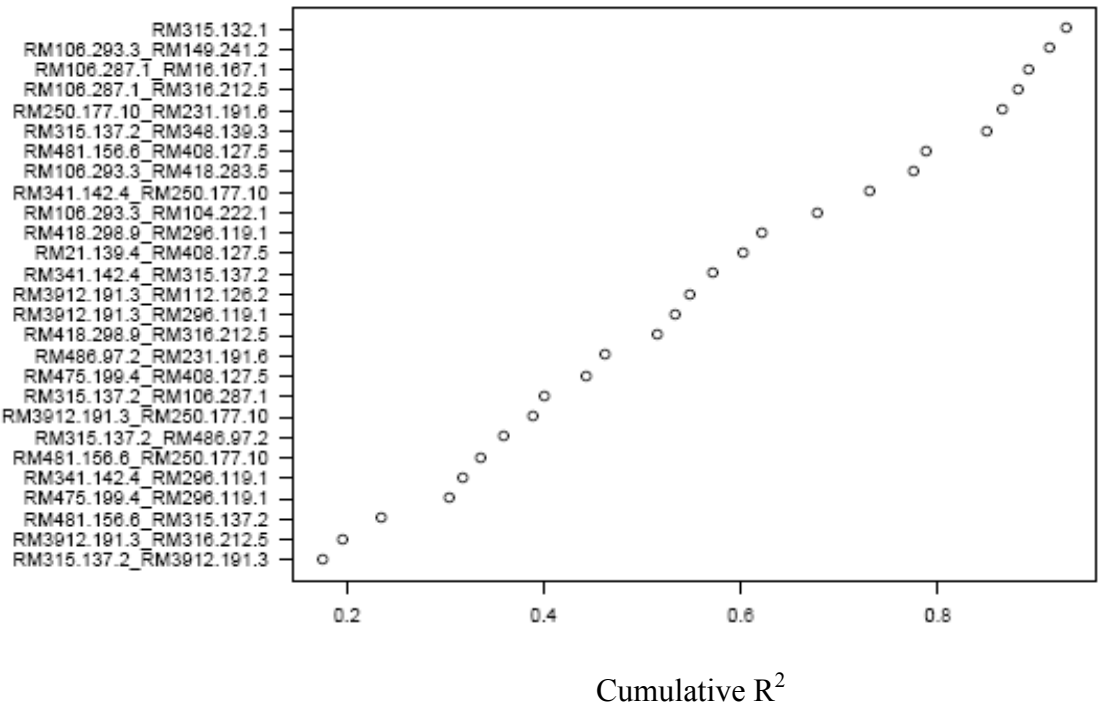



Figure 3.3c Selected marker effects from optimized SVR models and corresponding sequential $R^2$ values on the horizontal axis for and head rice. The number of total selected variables is 27 and $R^2 = 0.89$.

**3.4 Discussion**

Association genetics is a relatively new approach for the plant research community that has the potential to identify and characterize molecular markers associated with traits in ways not

possible by standard QTL mapping techniques. Recent reports have illustrated the potential value of association genetics to select markers for agronomic traits (Wang et al., 2008; Belo et al., 2008), but the best statistical strategy to address issues such as non-linear effects on trait variation remains an open question. The statistical foundation for all plant association genetics to date has been the general linear model that assumes factors or variables that influence agronomic characters will act in a linear manner. Because genes for complex traits and the corresponding phenotypes interact in non-linear relationships, we evaluated non-linear SVR for prediction accuracy and power of DNA markers to be associated with economically important agronomic traits in rice. The results showed that our modified SVR procedure produced high levels of accuracy using the radial basis kernel which is consistent with previous studies of maize inbred lines (De Baets et al., 2008). High levels of power were detected with the SVR procedure for all three complex traits that warrants further investigation. Our SVR approach for marker selection was supported by previous QTL mapping studies that identified the same genetic regions for the three traits evaluated in this study. The outcome and procedures developed during this study could provide insights and guidance for development of model simulations and design of future validation experiments.

The power analysis showed SVR produced high levels of power for the three traits over a wide range of effect sizes for all optimized models. All results suggest that the SVR procedure used during this study allowed for high accuracy levels and robust ability to detect candidate genotype-phenotype associations among rice inbred lines. The ability to account for population structure has been reported as a necessary cornerstone for association genetics in maize and other crops (Yu et al., 2006). The model-based "Structure" software program that assumes outcrossing is a popular approach to detect population stratification. Population structure analysis in the

current study revealed that gross population differentiation was absent among the elite rice breeding lines with a narrow germplasm base.

SSR markers were utilized in this study, but SNP markers will most likely become the marker of choice in the future for marker-assisted breeding due to high abundance, facile genotyping, and high throughput capabilities. One advantage of SVR under those circumstances may be the ability to obtain "sparse" solutions with relatively few variables versus other methods involving large datasets (Vapnik, 1995). Another advantage of SVR may be an internal validation step to estimate parameters that gives rise to high power and precision. All results obtained from this study suggested that SVR exhibited desirable features for association genetics in rice and other inbred species that should be further explored and developed for optimum power and prediction accuracy of marker-trait relationships.

**3.5 References**

Aluko G, Martinez C, Tohme J, Castano C, Bergman C, Oard JH (2004) QTL mapping of grain quality traits from the interspecific cross *Oryza sativa* x *O. glaberrima*. Theor. Appl. Genet. 109: 630–639.

Ayres NM, McClung AM, Larkin PD, Bligh HFJ, Jones CA, Park WD (1997) Microsatellites and a single nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germplasm. Theor. Appl. Genet. 94: 773–781.

Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B, Tingey S, Rafalski A (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Mol. Genet. Genom. 279:1–10.

Cohen J (1988) Statistical power analysis for the behavioral sciences, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 1-576.

De Baets B, Haesaert G, Van Bockstaele E (2008) Marker-based screening of maize inbred lines using support vector machine regression . Euphytica 161:123-131.

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631-8.

Maenhout S, De Baets B, Haesaert G, van Bockstaele E (2007) Support vector machine regression for the prediction of maize hybrid performance. Theor. Appl. Genet.115: 1003-13.

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175: 879-89.

McKenzie KS, Rutger JN (1983) Genetic analysis of amylose content, alkali spreading score, and grain dimensions in rice. Crop Sci. 23:306-313.

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc. Natl. Acad. Sci. USA 103:18656-61.

Septiningsih EM, Trijatmiko KR, Moeljopawiro S, McCouch SR (2003) Identification of quantitative trait loci for grain quality in an advanced backcross population derived from the *Oryza sativa* variety IR64 and the wild relative O. rufipogon. Theor. Appl. Genet. 107:1433-1441.

Simko I, Haynes KG, Jones RW (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173: 2237-45.

Tamaki S, Matsuo S, Wong HL, Yokoi S, Shimamoto K (2007) Hd3a protein is a mobile flowering signal in rice. Science 316:1033-1036.

Vapnik V (1995) The nature of statistical learning theory, Springer-Verlag, New York, pp. 1-338.

Vapnik V (1998) Statistical Learning Theory, John Wiley and Sons, New York, pp. 1-733.

Wang J, McClean PE, Lee R, Goos RJ, Helms T (2008) Association mapping of iron deficiency chlorosis loci in soybean (Glycine max L. Merr.) advanced breeding lines. Theor. Appl. Genet. 116: 777–787.

Yu J, Pressoir G, Briggs WH, Vroh VI, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38:203-208.

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2009) An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet. 3 (2007): e4 doi: 10.1371/journal.pgen.0030004, (verified April 27, 2009).

# CHAPTER 4 EVALUATION OF DNA MARKERS TO FACILITATE BREEDING FOR AROMA AND COOKING QUALITY IN LOUISIANA RICE

## 4.1 Introduction

### 4.1.1 Importance of Rice

Rice (*Oryza sativa* L.) is one of the most important food crops in the world serving as the principal source of calories for more than half of the world's population (Singh and Khush, 2000). Asia produces and consumes approximately 90% of the rice on earth. It is estimated that by the year 2025 nearly 4 billion people, mostly poor, will consume rice as a basic food. Global production is projected at 417 million tons of milled rice for 2007, but global consumption continues to outpace production which is expected at 423.2 million tons of milled rice (Grain: World Markets and Trade May 2006).

### 4.1.2 Rice Industry in the United States and Louisiana

Rice production and marketing in the United States is a multibillion dollar industry. At the farm level alone, rice generates more than $1.5 billion in revenues. In 2007, rice was planted on more than 1.1 million hectares in the United States with production estimated at 8.6 M MT (http://www.usarice.com /index. php?option=com_content&view=article&id=671&Itemid=386). U.S. rice production takes place in six states—Arkansas, California, Louisiana, Mississippi, Missouri, and Texas. The U.S. produces high-quality varieties of short, medium and long grain rice, as well as specialty rice including jasmine and basmati types. The U.S. rice farmers produce two percent of the world's annual rice supply and represent the world's fourth largest rice exporting country. Approximately half of the annual U.S. rice production is used domestically. Americans consume ~ 11 kg of rice per year which is substantially below world consumption levels of 85.9 kg per capita (http://www. unctad.org/infocomm/anglais/rice/market.htm#conso).

Louisiana ranks third in terms of rice total production following Arkansas and California. The rice industry in Louisiana accounts for $235 M in 2006, from 350,000 acres with average yields of 5,820 lbs/acre for a total of 20.1 M cwt (Louisiana Farm Reporter; http://www.lsuagcenter.com/agsummary/progressreport.aspx). Louisiana rice planting for 2007 was 360,000 acres, up 3 percent from a year earlier, but still the lowest acreage planted since 1914 (http://www.aragriculture.org/agfoodpolicy/radio/ may2007/ 042_05082007_audio.htm). For 2008, the area planted was 464,000 acres with average yield of 5,830lbs. The area planted in 2008 increased by more than 20% compared to 2007 (http://usda.mannlib. cornell.edu/ usda/ current/CropProdSu/CropProdSu-01-12-2009.pdf).

### 4.1.3 Status of Specialty Rice Breeding and Demand Worldwide

The demand for high quality or special purpose aromatic and basmati rices in the U.S. and elsewhere has increased during the past two decades (Cordeiro et al., 2000; Jin et al., 2003). Economic value of "*Jasmine*" aromatic rice for Thailand alone was $840 M and "*Basmati*" was $960 M for India and Pakistan in 2003 (http://basmati.com/ aromatic/ index.shtml). Thailand is the number one exporter of "*Jasmine"* aromatic rice to the U.S. In the U.S., ~ 12% of the total rice consumed is aromatic, primarily imported and consumed by the Asian-American community (Sha, 2005). Aromatic rice contains natural chemical compounds which give it a distinctive "popcorn" scent. Jasmine rice is sought for its aroma, flavor, slender kernels, and soft-cooking characteristics (Singh et al., 2000). With the huge market and increasing demand for high quality aromatic rice in the U.S. and worldwide, breeding for special purpose aromatic rice is imperative. Aromatic rice creates the option of securing higher returns over the conventional rice due to higher price (Jin et al., 2003).

However, using the traditional method of breeding, i.e. crossing and then selection, is tedious and labor intensive. Moreover, recessive traits such as aroma may be lost through selfing

and the selection process, given that grain evaluation through taste to determine aroma is often difficult, time consuming and unreliable at times.

### 4.1.4 Molecular Markers for Crop Improvement

The applications of molecular markers as a tool for crop improvement have improved efficiency in breeding new and improved rice lines in the last decade (Collard and Mackill, 2008). Molecular markers allow selection for particular characters or traits on the basis of a simple laboratory test on a small amount of leaf or grain tissue, rather than direct measurement of the character itself. There are several types of molecular markers available for use. Among them are restricted fragment length polymorphisms (RFLP), random amplified polymorphic difference (RAPD), amplified fragment length polymorphisms (AFLP), simple sequence repeats (SSR) and single nucleotide polymorphisms (SNP) that can detect a single nucleotide difference in the DNA sequence between two individuals. The utility of SNP markers has been reported in several crops with great success (Issiki et al., 1998; Bundock et al., 2004; Till et al., 2004), as well as its potential for plant genomic research (Feltus et al., 2004). In rice, *indica* and *japonica* genome sequences have been published and are publicly available (Feltus et al., 2004; Shen et al., 2004; Takashi M, 2005) that allow development of PCR-based SNP markers for efficient marker- assisted breeding.

### 4.1.5 Molecular Markers for Fragrance (Aroma), Amylose Content, and Gelatinization Temperature in Rice

The fragrance in Jasmine and Basmati rice has been associated with increased levels of 2-acetyl-1-pyroline (2AP), the chemical compound whose levels are controlled primarily by a single recessive gene (*fgr*) (Yoshihashi, 2002). The single recessive fragrance gene (*fgr*) has been linked to the RFLP marker RG28 on chromosome 8 at a genetic distance of 4.5 cM (Ahn et al., 1992). This RFLP marker was identified as a candidate to produce a PCR-based marker

capable of discriminating between fragrant and non-fragrant rice cultivars (Garland et al., 2000). PCR-based marker systems are desirable because of their simplicity and requirement for small quantities of tissue. The marker developed, SCU-Rice-SSR-1, relied on the use of capillary electrophoresis to discriminate the single base pair difference in a (T)n repeat between many fragrant and non-fragrant varieties. However, when using less sensitive DNA separation systems such as agarose gels, the marker failed to discriminate between the varietal types. Also, some important parent combinations were not polymorphic for this marker. In 2002, Cordeiro et al. developed an SSR marker that is linked to the fragrance gene. This SSR marker, however, could not predict fragrance status and was not completely accurate. Recently, Bradbury et al. (2005) identified an eight base pair deletion and three SNPs in exon 8 of the gene encoding betainealdehyde dehydrogenase 2 (*BAD2*) on chromosome 8 in rice as the most likely cause of fragrance in Basmati and Jasmine rices. Non-fragrant rice possess the fully functional copy of the gene encoding *BAD2*, while the fragrant rice possess the deletion and SNPs in *BAD2*, resulting in a frame shift that generates a premature stop codon to disable the *BAD2* enzyme. The crippled *BAD2* gene results in an increase of the previous compound 2-acetyl-1-pyroline (2AP) in the pathway.

Eating and cooking qualities of rice are primarily due to amylose content (AC) that is governed largely by the *waxy* locus (Zhou et al., 2003; Bao et al., 2004; Yamanaka et al., 2004). Low AC is usually associated with tender, cohesive, and glossy cooked rice (Juliano, 1971). Four QTLs on chromosomes 3, 4, 6, and 7 have been identified by Lanceras et al. (2000) for AC in a mapping population that accounted for 80% of the phenotypic variation in the population. Ramalingan (2002) used an SSR marker 484/W2R located on chromosome 6 that successfully differentiated high and low AC in a marker-assisted selection (MAS) program. RFLP markers C688 and C952 on chromosome 6 were designed and used for MAS to improve quality traits in

rice hybrids (Zhou, 2003). SSR marker RM 190, which is known to be located within the *waxy* locus on chromosome 6 (Temnykh et al., 2000; McCouch et al., 2002), is commonly used to classify rice into different AC classes (Bao et al., 2006). Yamanaka et al. (2004) identified SNPs in the *waxy* gene among glutinous cultivars. They developed a "derived cleaved amplified polymorphic sequence" marker (dCAPS) for detection of the one-base splicing mutation without the need for sequencing.

Another important cooking quality trait in rice is gelatinization temperature (GT) which was found to be controlled by the *alk* locus (Bao et al., 2006). Molecular markers for the *alk* gene have been cloned and developed (Gao et al., 2003; Jiang et al., 2004). Nucleotide substitutions in the coding sequence of starch synthase IIa (SSIIa) have been reported to cause alterations in GT (Gao et al., 2003). Chen et al. (2003) identified two separate mutations in the SSIIa gene that were associated with low GT, and Fjellstrom et al. (2004) detected two additional SNPs in exon 8 of the *alk* gene associated with cooking quality . Nakamura et al. (2005) analyzed in detail the effect of amino acid replacement caused by these SNPs on the enzyme activity, amylopectin structure and variation in GT. The results indicated that two of the SNPs (4,198 and 4,229/ 4,330 bp) were essential for SSIIa activity and granule association. Waters et al. (2006) identified two SNPs in the exon in the SSIIa gene that could differentiate between high and low GT. The A/G SNP at base 2412 determined whether a methionine or valine was present at the corresponding amino acid residue in SSIIa, while two adjacent SNPs at bases 2543 and 2544 coded for either leucine (GC) or phenylalanine (TT). Rice varieties with high GT starch exhibited a combination of valine and leucine at these residues. In contrast, rice varieties with low GT starch possessed a combination of either methionine and leucine or valine and phenylalanine at these same residues. Bao et al. (2006) found similar SNPs for differentiating between high and low GT. Their study provided further support for the utilization

of the GC/TT polymorphism in the SSIIa gene. GC/TT SNP could differentiate rice with high or intermediate GT from those with low GT in about 90% of cases among 509 rice samples (Bao et al., 2006). Thus, this SNP polymorphism may be very useful in marker-assisted selection for the improvement of GT and other physicochemical properties of rice.

### 4.1.6 SNP Marker Development for Marker-Assisted Breeding in Rice

Kadaru et al. (2006) developed a modified procedure based on standard Ecotilling (Comai et al., 2004) for rice SNP discovery and genotyping referred to as Alternative Ecotilling (AE). Four previously reported and 14 new SNPs in the *alk* and *waxy* genes among 57 rice accessions based on comparisons with sequencing results were characterized by AE for GT and AC, respectively. In addition, new SNP markers for haplotype-specific markers in exon 7 of the *BAD2* gene for marker-assisted identification and introgression of the aroma gene in U.S. rice were developed in Prof. Oard's laboratory. These SNPs can distinguish aromatic and non-aromatic phenotypes and were consistent with corresponding marker haplotypes for all progeny tested. Allele-specific PCR assays were developed in Prof. Oard's laboratory for aroma (fragrance), *alk*, and *waxy* that can distinguish and differentiate between homozygous and heterozygous SNP alleles of rice that could lead to efficient marker-assisted breeding. The primary objective of this research is to evaluate the potential of selected DNA markers to facilitate rapid introgression of aroma and cooking quality traits into elite Louisiana breeding lines. A second objective is to combine desirable alleles for aroma and cooking quality in elite LA breeding lines with acceptable agronomic traits.

## 4.2 Materials and Methods

### 4.2.1 Plant Material

Breeding lines used in this study were obtained in cooperation with Prof. Xueyan Sha, Rice Breeder at the Rice Research Station, Crowley, LA. The first set of lines consisted of 228

plants, designated Batch 1, in 2007 from 13 different segregating populations, four of which were parents (96 INT/ARNT, AC969, JSMN/DLLA/LLEAH/DLLA, L202/LEAH//TORO/3/IR67016), and nine were derived from 2-way and 3-way backcrossed (BC) and $F_2$ lines . The second set of materials, designated Batch 2, consisted of 58 individual plants from 19 different selected lines that were grown in the greenhouse in 2007. Seeds of the second batch were provided by Prof. Sha as part of the collaborative work on SNP markers.

The 286 lines form Batch 1 and Batch 2 were genotyped for presence of desirable alleles for aroma, AC, and GT in 2007 using SNP markers developed by Prof. Oard's laboratory (aroma, *waxy*) and USDA researchers in Beaumont, TX (*alk*). Out of the 286 lines, 78 individual plants carried desirable alleles either in a homozygous or heterozygous state. Forty eight plants out of 78 were further selected based on overall agronomic plant type. Using the SNP marker data, these 48 plants were crossed among themselves based on maturity and grain type to combine desired combinations of *aroma*, low AC, and low GT alleles. Seventy-six new cross combinations were generated during the summer of 2007. A total of 33 out 76 $F_1$'s together with 32 segregating parents were grown in the greenhouse during the Fall of 2007 for additional crosses and selection of improved plant type. A second round of 91 crosses were generated that constituted the material for genotyping, designated Batch 3, and field evaluation during the Summer of 2008. In addition, populations totaling 67 individual plants (Batch 4 and 5) from Prof. Sha's program were genotyped for GT to confirm and validate the presence of favorable alleles in the selected breeding lines.

### 4.2.2 Hybridization and Pyramiding of Quality Traits in Aromatic Rice Breeding Populations

As stated in previous sections, 35 lines were selected from among the 65 individual plants from Batch 1 that possess desired alleles in either homozygous or heterozygous state and

13 plants were selected from Batch 2 in the same manner. The 35 selected plants were crossed among themselves in 2007, taking into consideration plant height and grain type to develop long-grain type aromatic lines. A total of 65 new crosses were developed by intercrossing 12 lines used solely as male parents and 23 lines used either male or female. In addition, 11 crosses were completed from 9 of the 13 greenhouse selections from Batch 2.

A total of 32 segregating parents, 21 $F_1$'s from Batch 1, and 11 $F_1$'s from Batch 2 were planted in the greenhouse during the Fall of 2007. Evaluations on the greenhouse plants were carried out based on overall phenotype that includes grain type and maturity in addition to previous marker data in planning the hybridization scheme. Crossing schemes were: $F_1$ x $F_1$, $F_1$ x Segregating Parentals (SP), and SP x SP in an attempt to improve overall phenotype and increase frequency of favorable allele. A total of 91 new crosses were generated which were planted and evaluated in the field and genotyped for aroma, AC, and GT SNP alleles during the summer of 2008.

**4.2.3 Leaf Collection and Genomic DNA Extraction**

Leaf samples for 705 individual plants were collected in the field in 2007 and 2008 at the Rice Research Station (Crowley, LA) and in the greenhouse (Table 4.1). All samples were stored at -20 °C prior to DNA extraction which was carried out using the Qiagen DNeasy 96 Plant Kit. A modified DNA extraction procedure was carried out as follows: frozen leaf samples of 30-40 mg were cut into small pieces and placed into microtube collection racks. Two stainless steel beads (2.3mm) (Biospec products) were added to each microtube along with 400 μL working lysis solution. To ensure complete disruption of the plant tissue, racks of collection microtubes were placed in a Mini-Bead Beater (Biospec products) for grinding twice, each at 2.5 minutes. All remaining steps were followed as indicated in the Qiagen DNeasy Plant Handbook. A sample

volume of 200 µL was collected at the final step. DNA concentrations were determined using the NanoDrop (spectrophotometer) and diluted to 10 ng/µL as working stocks for SNP genotyping.

Table 4.1 Summary of plant materials genotyped for Aroma, AC, and GT SNP alleles

| Batch | Collection Site | No. of populations/ Generation | No. of plants |
|---|---|---|---|
| 1 | Rice Research Station | 13/ Parents, $F_2$, BC | 228 |
| 2 | Baton Rouge Greenhouse | 19/ Parents, $F_1$, $F_2$ | 58 |
| 3 | Rice Research Station | 1/ $F_5$:$F_7$ | 54 |
| 4 | Rice Research Station | 1/ $F_5$:$F_7$ | 13 |
| 5 | Rice Research Station | 91/ $F_1$'s | 352 |

### 4.2.4 Polymerase Chain Reaction (PCR), SNP Genotyping, and Scoring

SNP markers developed by Prof. Oard's laboratory (aroma, *waxy*) and USDA researchers in Beaumont, TX (*alk*) were used to obtain genotypes. For the *BAD2* aroma gene, primer sequence for the fragrance allele was 5'-CTGGTATATATTTCAGCTGATC-3' and the non-fragrance allele was 5'-AAAGATTATGGCTTCAGTGATC-3' with a common reverse primer of 5'-CCAGTGAAACAGGCTGTCAA-3'. For the *waxy* gene, primer sequence of the high AC allele was 5'-CAGGAAGAACATCTGCACGG-3' and the low AC allele was 5'-CAGGAAGAACATCTGCACGT-3' with a common reverse primer 5'-TTTCCAGCCCAACACCTTAC-3'. The last primer combination was the *alk* gene where primer sequence for the high GT allele in the *alk* gene was 5'-TGCCGCGCACCTGGAGC-3' and the low GT allele was 5'-CATGCCGCGCACCTGGAAA-3' and a common reverse primer of 5'-CGCCGAGCCGCACAAGC-3'.

PCR was performed using 2.0 µL template DNA (20 ng) in a 10 µL PCR reaction containing 1 µL of 10X buffer (Applied Biosystems, Inc.), 0.8 µL of dNTP mix (Applied Biosystems, Inc.), 0.2 µL of each primer (20 µM), 0.08 µL polymerase enzyme (Applied Biosystems, Inc.) and 5.72 µL distilled water . PCR reactions were performed using the iCycler (Bio-Rad, CA). The thermocycle profile used to amplify the 237 bp fragment of the aroma/*BAD2*

gene was 95 °C- 2 min, 28 cycles of (95 °C - 12 s, 60 °C - 12 s, 72 °C - 12 s) and 72 °C - 5 min. PCR amplifications for the 186 bp fragment of the *waxy* gene was carried out using 95 °C - 3 min, 28 cycles of (95 °C - 20 s, 60 °C - 20 s, 72 °C - 20 s) and 72 °C - 5 min. Finally, PCR amplification profile for the 90 bp *alk* gene fragment was 95 °C - 3 min, 28 cycles of (95 °C - 20 s, 63 °C - 20 s, 72 °C - 20 s) and 72 °C - 5 min.

PCR products were resolved in a 2% agarose gel treated with 0.05 µg/mL ethidium bromide solution and visualized under UV light using Gel Logic 200 Imaging System. SNP genotypes were scored as band present (1) or absent (0) for each individual sample. Data were recorded and entered into an Excel spreadsheet.

**4.2.5 Field Experiment, Phenotypic Data Collection, and Analysis**

The 91 different $F_1$ crosses that consisted of 352 individual plants were grown in the greenhouse during the Fall of 2007. Seedlings were grown in the greenhouse from April 4, 2008 until they were ready for transplanting a month later at the Crowley Rice Research Station. Individual $F_1$ plants were transplanted at approximately 20 cm distances within rows and 30 cm distance between rows. Normal cultural and management practices were followed for fertilizer, herbicide, and insecticide applications. Standard water management was carried out to ensure normal and healthy growth of plants.

Phenotypic data collected for eight traits in this study were the following: (1) plant height (PLTH) which was measured from the base of the rice plant to the tip of the highest panicle in cm, (2) plant maturity (PLTM) was scored as intermediate (i) (100-115d), late (l) (116-130d), and very late (vl) (>130d) maturity (3) panicle number (PANN) was determined by counting the number of seed-bearing panicles in each plant, (4) panicle length (PANL) was determined by measuring the main tiller/panicle and expressed in cm, (5) spikelet number (SPKN) was determined by counting the filled and unfilled spikelets of the main tiller, (6) seeds per panicle

(SDPN) was determined by counting the filled grains of the main tiller in each plant in the population, (7) spikelet fertility (SPFT) expressed as percentage was determined by the formula SDPN/SPKN *100 (8) presence of pubescent or glabrous leaves (PBGL) was determined by physically examining leaves on each plant. Data were gathered and entered into an Excel spreadsheet. The agronomic data were analyzed using SAS software package 9.1.3. Descriptive statistics were computed using PROC UNIVARIATE. PROC CORR was used to calculate linear correlations between the different agronomic traits. Frequency distribution figures were generated in Microsoft Excel 2007.

## 4.3 Results and Discussion

### 4.3.1 Molecular Profiles of Breeding Populations for Aromatic Rice

Table 4.1 summarizes plant material genotyped for aroma, AC, and GT alleles that were placed into five Batches based on population and/or generation. Molecular analysis of Batch 1 revealed that 28% (65/228) carried desired allelic combinations in either a homozygous or heterozygous state (Table 4.2). Specifically, 68% (44/65) of the plant selections were homozygous for aroma while a majority (64/65) of the lines were heterozygous for AC, and all selections were heterozygous for GT. Approximately one-half (54%, 35/65) of these individuals were selected in the field based on overall plant type and utilized in crosses to combine desired alleles and increase their frequency in elite breeding lines.

Table 4.3 summarizes the genotyping results for Batch 2 consisting of $F_2$ material grown in the greenhouse in 2007. Approximately one-quarter (22%, 13/58) carry desired alleles for aroma, AC, and GT in the heterozygous state. These $F_2$ plants were then utilized in crosses to combine desired alleles and increase allele frequency in the population. In addition, separate genotyping for GT were carried out on 54 and 13 advanced lines ($F_5$:$F_7$) from Batch 3 and Batch 4, respectively.

Table 4.2 Molecular profiles of 65 selected rice plants from Batch 1 that contain the desired alleles for Aroma, AC, and GT.

| Pedigree | Plant # | Aroma | AC | GT | |
|---|---|---|---|---|---|
| 96 INT/ARNT/4/9502008-A//AR1188/CCDR/3/CPRS/LGRU//97KDM X2-1 | 2-5 | H | h | h | |
| 96INT/ARNT/3/96 INT/ARNT//CCDR | 4-4 | H | h | h | ** |
| | 4-5 | h | h | h | ** |
| | 4-8 | H | h | h | ** |
| | 4-14 | h | h | h | |
| | 4-18 | h | h | h | |
| | 4-29 | h | h | h | ** |
| | 4-32 | H | h | h | ** |
| | 4-36 | H | h | h | |
| 96INT/ARNT/3/97 KDM X2-1/WELLS//AC969 INT/ARNT//CCDR | 5-6 | H | h | h | ** |
| | 5-7 | H | H | h | ** |
| | 5-8 | H | h | h | |
| | 5-9 | H | h | h | ** |
| | 5-10 | H | h | h | ** |
| | 5-11 | H | h | h | |
| | 5-12 | H | h | h | ** |
| | 5-13 | H | h | h | ** |
| | 5-14 | H | h | h | ** |
| | 5-15 | H | h | h | ** |
| | 5-16 | H | h | h | ** |
| | 5-17 | H | h | h | ** |
| | 5-18 | H | h | h | ** |
| | 5-19 | H | h | h | ** |
| | 5-20 | H | h | h | ** |
| CPRS/KDM 105/3/JSMN/DLLA/LEAH/DLLA | 7-1 | H | h | h | |
| | 7-2 | h | h | h | |
| | 7-6 | H | h | h | |
| | 7-8 | h | h | h | ** |
| | 7-11 | h | h | h | |
| | 7-14 | H | h | h | |
| | 7-15 | H | h | h | |
| | 7-16 | h | h | h | ** |
| | 7-18 | h | h | h | |
| | 7-26 | h | h | h | |
| | 7-29 | H | h | h | ** |

Table 4.2 (continued)

| Pedigree | Plant # | aroma | AC | GT | |
|---|---|---|---|---|---|
| CPRS/KDM 105//96 INT/ARNT | 8-2 | h | h | h | ** |
| | 8-3 | H | h | h | ** |
| | 8-5 | H | h | h | ** |
| | 8-6 | h | h | h | |
| | 8-7 | h | h | h | ** |
| | 8-9 | H | h | h | |
| | 8-12 | h | h | h | ** |
| | 8-18 | h | h | h | ** |
| | 8-22 | H | h | h | |
| | 8-25 | H | h | h | |
| | 8-26 | h | h | h | |
| | 8-28 | H | h | h | |
| | 8-37 | h | h | h | |
| | 8-38 | h | h | h | |
| JSMN/DLLA/LEAH/DLLA/4/9502008-A//AR 1188/CCDR/3/CPRS/LGRU | 10-1 | H | h | h | ** |
| | 10-4 | H | h | h | ** |
| | 10-6 | h | h | h | |
| | 10-7 | h | h | h | ** |
| | 10-9 | h | h | h | |
| | 10-10 | h | h | h | ** |
| | 10-19 | H | h | h | |
| | 10-20 | H | h | h | |
| L202/LEAH//TORO/3/IR67016/4/97 KDM X2-1/WELLS//AC969 | 13-1 | H | h | h | |
| | 13-2 | H | h | h | |
| | 13-3 | H | h | h | ** |
| | 13-4 | H | h | h | |
| | 13-5 | H | h | h | |
| | 13-6 | H | h | h | ** |
| | 13-7 | H | h | h | ** |
| | 13-8 | H | h | h | ** |
| | 13-9 | H | h | h | |

** Plant selected for overall appearance and additional crosses, H-homozygous for allele of interest, h-heterozygous for allele of interest, AC - amylose content, GT-gelatinization temperature

The results again showed that approximately 25% of the lines for Batch 3 (14/54) and Batch 4 (3/13) possessed the desired low GT allele. These results suggest that using markers at a late stage of selection such as $F_5$ or $F_7$ for a recessive trait such as GT would result in a relative small chance (~25%) of selecting lines that possess the desired allele(s). Nevertheless, separate laboratory tests from Prof. Sha's laboratory confirmed and validated the marker results from this study (data not shown). Overall, these results illustrates the potential and value of the GT (*alk*) SNPs for marker-assisted breeding to develop aromatic rice as it could enrich for desired GT alleles early in the breeding process. Moreover, if breeding objectives entail simultaneous selection of different quality traits that are difficult or expensive to phenotype, markers would play a significant role.

Table 4.3 Molecular profiles of 13 selected $F_2$ rice plants from Batch 2 that contain alleles for Aroma, AC, and GT.

| Pedigree | Plant # | Aroma | AC | GT |
|---|---|---|---|---|
| L202/LEAH/TORO/3/IR67016/TAU CAURI//KBNT/LCSN | 3-1 | h | h | h |
| | 3-2 | h | H | h |
| | 4-3 | h | h | h |
| | 5-1 | h | H | h |
| | 6-1 | H | h | h |
| | 6-2 | h | h | H |
| JSMN/DLLA/96SP287/3/952008-A/DREW | 14-2 | h | H | h |
| | 15-2 | h | h | h |
| | 16-1 | H | h | H |
| | 17-1 | H | h | H |
| | 17-2 | h | H | H |
| | 18-2 | h | H | h |
| | 19-2 | h | h | h |

H-homozygous for allele of interest, h-heterozygous for allele of interest, AC - amylose content, GT-gelatinization temperature

Genotyping results for the 452 $F_1$'s (Batch 5) for aroma, AC, and GT were summarized in Table 4.4. The majority of $F_1$'s (78%) were heterozygous for aroma and only 1% of the population was found to be homozygous. Genotype result for AC revealed that 26 % of the aromatic lines also carried the desired low AC allele. Furthermore, 44% of the $F_1$'s were found to be heterozygous for AC-possessing both low and high AC alleles. GT genotypes revealed that 17% of the total $F_1$'s produced a low GT genotype -the desired outcome for the development of aromatic lines. Nearly half (41%) of the lines were also found to be heterozygous for both the low and high GT alleles. Gene combinations for the desired genotypes are summarized in Table 4.4.

Table 4.4 Aroma, AC and GT allele-genotyping in one and two-gene combinations of the 452 $F_1$'s (Batch 5) evaluated at Rice Research Station, Crowley, LA, 2008

| Genotype | Aroma | AC | GT | Genotype | Aroma/AC | Aroma/ GT | AC/ GT |
|---|---|---|---|---|---|---|---|
| $H^+$ | 6 | 118 | 78 | $H^+/H^+$ | 6 | 1 | 37 |
| h | 353 | 201 | 186 | $H^+$/h or h/$H^+$ | 86 | 66 | 54 |
| $H^-$ | 87 | 105 | 145 | h/h | 174 | 161 | 135 |
| Nd | 6 | 28 | 43 | $H^-$/h or h/$H^-$ or $H^-/H^-$ | 152 | 175 | 145 |
| Total | 452 | 452 | 452 | Total | 418 | 403 | 371 |

$H^+$-homozygous for desired allele; h-heterozygous; $H^-$-homozygous for undesirable allele; Nd-no data

A small portion of the population had either the aroma/AC or aroma/GT combinations in the homozygous state. However, a small percentage (8%) of the populations had both the AC and GT combinations in the desired combinations. Two gene combinations that are heterozygous for both genes are 38%, 36% and 30%, respectively, for Aroma/AC, Aroma/GT and AC/GT combinations. These results showed the potential of markers to rapidly identify individuals possessing the desired two-gene gene combinations for breeding and selection. Moreover, individuals lacking the gene(s) of interest could be discarded in early generations. This would

improve selection efficiency and breeding for specialty jasmine rice where quality traits are difficult and expensive to phenotype.

### 4.3.2 Descriptive Statistics and Correlation Analysis of Agronomic Traits for $F_1$'s of Aromatic Breeding Lines

Table 4.5 summarizes the six agronomic data generated from 452 $F_1$'s (Batch 5) evaluated at the Rice Research Station, Crowley, LA, 2008. The results show high levels of variation for all traits measured in this study. For example, the range for plant height (PLTH) was extensive at 60 cm. Nevertheless, a majority 75% of the $F_1$ plants (340/452) were <100 cm in height, an acceptable level for elite breeding material. For panicle number (PANN), the range was also wide (38 panicles), indicating the potential of identifying high tillering plants in the populations. For spikelet fertility (SPKN), only 37% of the $F_1$'s produced values >85%.

Table 4.5 Descriptive statistics of six agronomic traits of 452 $F_1$'s (Batch 5) evaluated at Rice Research Station, Crowley, LA, 2008.

| Trait | Unit | N | Mean | Std Dev | Min | Max |
|---|---|---|---|---|---|---|
| PLTH | cm | 451 | 94.27 | 9.13278 | 69 | 129 |
| PANN | - | 447 | 14.87 | 5.55304 | 3 | 41 |
| PANL | cm | 452 | 23.83 | 2.35814 | 17.8 | 33.1 |
| SPKN | - | 452 | 228.42 | 57.30926 | 81 | 361 |
| SDPN | - | 452 | 175.20 | 50.98544 | 26 | 331 |
| SPFT | percent | 452 | 77.52 | 15.02513 | 9 | 100 |

PLTH-plant height, PANN-panicle number, PANL-panicle length,
SPKN-spikelet number per panicle, SDPN-filled seeds per panicle, SPFT-spikelet fertility

The low percentage of fertile $F_1$'s may be due to environmental conditions in 2008 or to inter-subspecific hybridizations that affect reproductive potential. For the remaining four traits (PANL, SPKN, SDPN, SPFT), similar trends of high genetic variations were observed. Nearly half (57%) exhibited the desired glabrous leaf trait for U.S. rice breeding. One half (52%) of the

lines showed intermediate maturity, 38% exhibited late maturity while the remaining lines were very late in maturing (data not shown). The overall observed high levels of variation in these lines were consistent with hybrids derived from *indica* and *japonica* germplasm. The majority of the $F_1$ plants showed agronomic traits similar to other advanced LSU breeding lines that warrant further selection combined with the molecular profiling methods described above.

Figure 4.1 shows the phenotypic distribution of the six agronomic traits collected from the 452 $F_1$ plants. All traits showed a normal or close to normal distribution except for SPFT that was substantially skewed toward high fertility for most of the $F_1$'s. The results suggest that high levels of variation present in this material will provide ample opportunity to identify new phenotypic combinations in the next generation that are superior to those of the parents evaluated in this study.

Correlation analysis of the six quantitative traits using Pearson Correlation Coefficients revealed interesting results (Table 4.6). For example, PLTH was found to be moderately associated with multiple yield related traits that includes panicle number (PANN), panicle length (PANL), number of spikelets per panicle (SPKN), and seeds per panicle (SDPN), indicating the possibility of using PLTH as indirect selection for the above mentioned traits that are more difficult to phenotype than PLTH. On the other hand, PLTH was negatively associated with SPFT indicating that taller plants tended to have lower spikelet fertility. Similarly, PANL was found to be associated with SPKN, and SDPN in the $F_1$'s evaluated. This result would also suggest the potential of PANL as indirect selection for SPKN and SDPN. As expected, SDPN was found to be associated with SPFT. Clearly, a more thorough experiment in multiple locations is needed to confirm and validate these results. Shi (1995) found high association between panicle density and grains per panicle as well as spikelet fertility. These associations found in this study may be due to high phenotypic variability observed in the populations

brought about by the *indica x japonica* crosses used in the development of the populations. These warrant further verification in other populations under different environmental conditions.

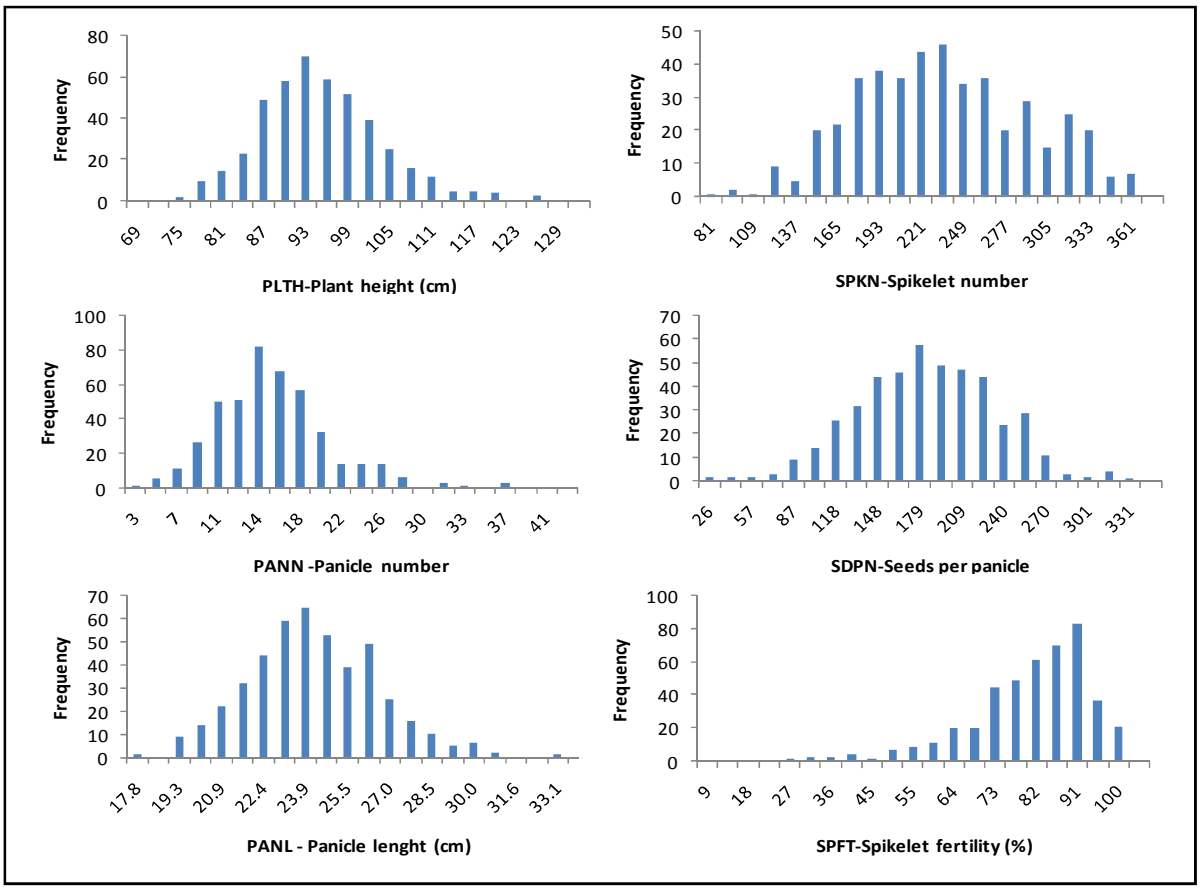



Figure 4.1 Frequency distribution of six agronomic traits of $F_1$'s (n=452) of aromatic rice breeding populations evaluated at Rice Research Station, Crowley, LA, 2008.

### 4.3.3 Marker and Phenotype Profiles of Selected Aromatic Lines

Table 4.7 summarizes the individual plant selections based on marker data and their corresponding phenotypes. It was interesting to note that one $F_1$ plant (324-7) possessed desired alleles for presence of aroma, low AC, and low GT. The corresponding phenotype data for this selection were encouraging. For example, PANN, SPKN, SDPN were above the population mean and all other traits were within the acceptable range for an elite breeding line except for spikelet fertility (<85%).

Table 4.6 Correlation analyses of six agronomic traits of 452 $F_1$'s of aromatic rice breeding populations evaluated at Rice Research Station, Crowley, LA, 2008.

| | Pearson Correlation Coefficients<br>Prob > \|r\| under H0: Rho=0 | | | | |
|---|---|---|---|---|---|
| | PLTH | PANN | PANL | SPKN | SDPN |
| PANN | 0.33416<br>(<.0001) | | | | |
| PANL | 0.33911<br>(<.0001) | 0.03645<br>(0.4420) | | | |
| SPKN | 0.32410<br>(<.0001) | -0.01460<br>(0.7581) | 0.42164<br>(<.0001) | | |
| SDPN | 0.16387<br>(0.0005) | 0.02132<br>(0.6530) | 0.30613<br>(<.0001) | 0.71940<br>(<.0001) | |
| SPFT | -0.17245<br>(0.0002) | 0.04457<br>(0.3471) | -0.09768<br>(0.0379) | -0.21860<br>(<.0001) | 0.50281<br>(<.0001) |

PLTH-plant height, PANN-panicle number, PANL-panicle length,
SPKN -spikelet number per panicle, SDPN-filled seeds per panicle, SPFT-spikelet fertility

Plant 324-7 is a candidate for advancement and selection of better improved agronomic traits in succeeding generations. In addition, 5 other $F_1$'s (324-1, 324-4, 324-6, 325-1, 364-2) were homozygous for low AC and low GT and heterozygous for aroma with good phenotypic data including high spikelet fertility (>85%). Except for one of the five selections, plant height was <100 cm. Moreover, the five selected $F_1$'s produced >12 panicles along with high density spikelets. Overall, the six selected $F_1$'s were considered as candidates for advancement as elite aromatic lines with superior agronomic traits. The remaining 28 selections could be advanced to $F_2$ and succeeding generations.

Results from the molecular and agronomic analysis in this study clearly showed that the SNP marker approach was able to enrich for frequency of desired alleles in lines in only two generations that also exhibited desired agronomic characteristics. The strong implication is that

marker-assisted methods for certain traits such as aroma, AC, and GT could speed up and increase efficiency in development of new Louisiana aromatic rice varieties.

Table 4.7 Molecular and phenotypic profiles of selected $F_1$'s (n=34) derived from selected backcrosses and advanced generation lines of aromatic rice breeding populations evaluated at Rice Research Station, Crowley, LA, 2008.

| Phenotype | | | | | | | | Genotype | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Plant# | PLTH | PLTM | PANN | PANL | SPKN | SDPN | SPFT | Aroma | AC | GT | |
| 324-1 | 99 | i | 18 | 23.1 | 280 | 252 | 90 | h | H | H | ** |
| 324-2 | 87 | l | 4 | 23.9 | 173 | 153 | 88 | h | H | H | |
| 324-3 | 102 | i | 18 | 26.6 | 234 | 179 | 77 | h | H | H | |
| 324-4 | 98 | i | 22 | 21.9 | 197 | 180 | 91 | h | H | H | ** |
| 324-6 | 99 | i | 20 | 23.2 | 157 | 147 | 94 | h | H | H | ** |
| 324-7 | 105 | i | 25 | 23.2 | 234 | 172 | 73 | H | H | H | *** |
| 325-1 | 91 | i | 13 | 24.6 | 311 | 262 | 84 | h | H | H | ** |
| 325-2 | 78 | i | 13 | 23.7 | 321 | 245 | 76 | h | H | H | |
| 328-2 | 111 | i | 16 | 24.4 | 203 | 146 | 72 | h | H | H | |
| 328-6 | 114 | l | 19 | 23.9 | 268 | 192 | 71 | h | H | H | |
| 329-2 | 111 | l | 22 | 23.7 | 162 | 133 | 82 | h | H | H | |
| 329-3 | 109 | i | 14 | 25.3 | 191 | 137 | 72 | h | H | H | |
| 329-6 | 110 | l | 12 | 27.5 | 289 | 216 | 75 | h | H | H | |
| 329-7 | 110 | l | 10 | 23.9 | 277 | 193 | 70 | h | H | H | |
| 336-1 | 100 | l | 16 | 22.0 | 213 | 148 | 69 | h | H | H | |
| 336-7 | 100 | l | 10 | 22.3 | 231 | 133 | 57 | h | H | H | |
| 336-8 | 100 | l | 14 | 19.6 | 170 | 104 | 61 | h | H | H | |
| 337-6 | 100 | i | 13 | 25.8 | 254 | 136 | 53 | h | H | H | |
| 337-7 | 88 | l | 9 | 22.9 | 256 | 120 | 47 | h | H | H | |
| 337-9 | 104 | l | 18 | 25.1 | 326 | 246 | 75 | h | H | H | |
| 343-1 | 83 | i | 17 | 22.5 | 181 | 143 | 79 | h | H | H | |
| 343-2 | 95 | nd | 8 | 25.4 | 237 | 195 | 82 | h | H | H | |
| 343-3 | 94 | nd | 10 | 24.1 | 196 | 146 | 75 | h | H | H | |
| 345-1 | 88 | l | 17 | 23.4 | 112 | 83 | 74 | h | H | H | |
| 347-1 | 69 | l | 4 | 22.5 | 202 | 162 | 80 | h | H | H | |
| 347-2 | 97 | l | 8 | 22.8 | 207 | 143 | 69 | h | H | H | |
| 347-3 | nd | nd | nd | nd | nd | nd | nd | h | H | H | |
| 347-4 | 83 | l | 3 | 22.4 | 249 | 172 | 69 | h | H | H | |
| 347-6 | 94 | l | 13 | 25.7 | 279 | 187 | 67 | h | H | H | |
| 347-9 | 93 | l | 13 | 25.0 | 252 | 203 | 81 | h | H | H | |
| 359-1 | 97 | vl | 10 | 26.9 | 168 | 93 | 56 | h | H | H | |
| 364-1 | 95 | l | 12 | 23.0 | 249 | 183 | 73 | h | H | H | |
| 364-2 | 87 | l | 16 | 22.2 | 220 | 194 | 88 | h | H | H | ** |
| 365-1 | 100 | l | 24 | 26.0 | 185 | 142 | 77 | h | H | H | |

PLTH-plant height, PLTM-plant maturity (i-intermediate, l-late, vl-very late), PANN-panicle number, PANL-panicle length, SPKN-spikelet number per panicle, SDPN-filled seeds per panicle, SPFT-spikelet fertility, h-heterozygous, H-homozygous for the trait of interest, ***-selection based on genotype and phenotype data, nd-no data collected

The primary method to identify aromatic rice primarily involves cooking rice grains to distinguish aromatic rice from non-aromatic types (Sha et al., 2000). But the use of this method is limited by the need of technical expertise due to inconsistencies of results from person to person and has low throughput in addition to the fact that phenotyping is completed after grain harvest. Recent advances in molecular biology paved the way for the development of a more precise method of genotyping. Allele-specific PCR amplification assay for the *BAD2* gene has been developed (Bradbury, 2005; Kadaru, 2006) for use in marker-assisted breeding. This technology has been demonstrated to be reliable and efficient in Australian temperate *japonica* aromatic and non-aromatic germplasm (Bradbury, 2005) and in U.S. aromatic rice varieties which are mainly derived from tropical *japonica* and *indica* germplasm (Kadaru, 2007). AC and GT traits are primary determinants to rice cooking and eating qualities that are also difficult to phenotype. Modern AC determination for example requires an Automatic Recording Titrator (ART-3, Hirama Laboratories, Kanagawa, Japan) using the iodine titration method (Tan et al., 1999; Tian et al., 2005) while GT determination involves a long and rigorous process as proposed by Little et al. (1958) and Wang et al. (2007). Recent developments in allele-specific PCR assays for AC (*waxy*) and GT (*alk*) genes opens the avenue for a more precise and accurate method to determine grain quality (Kadaru et al., 2006) and expedite breeding efficiency.

The MAS strategy holds promise for traits that are difficult to phenotype in early generations, particularly after grain harvest which can be costly and labor-intensive. In a classical $F_2$ population where genes/traits of interest are recessive, theoretically 25% of the population carries the recessive trait in homozygous state. With the use of efficient and reliable molecular markers, 75% of the $F_2$ plants in each population could be discarded. This approach would not only reduce the number of plants to evaluate, but also increase the frequency of favorable alleles in succeeding generations. Moreover, the power of MAS is perhaps best

exemplified when selection of more than one trait is involved. Zhou et al. (2003) demonstrated the ability of MAS to simultaneously improve four quality traits in hybrid rice germplasm.

The results from this study showcased the ability of molecular markers to screen and select individuals that possess the quality traits of interest in rice. It is interesting to note the power of SNP markers in discriminating individual plants with high success. Without the use of markers such as those used in this study, identification of individual plants or lines that carry all favorable alleles for aroma, low AC and low GT would be labor-intensive, costly, and time consuming. Results from this study open the avenue of routinely using SNP markers for breeding aromatic and cooking quality traits. However, further validation in the next generation of the best plant selections would further confirm the stability and accuracy of these SNP markers for marker-assisted breeding.

Although MAS offers enormous potential, the cost of genotyping is still a barrier for wider application of MAS in applied plant breeding programs (Collard and Mackill, 2008). A cost-benefit analysis of using markers for specific traits will determine the efficient use of this technology. In maize for example, it was established that using markers to select for *opaque2*-the gene associated with quality protein in maize, is more economical than conventional screening methods (Dreher, et al., 2003).

In case of quality traits in rice, the use of SNP markers mentioned in this study as an integral part of rice breeding program for aromatic rice is highly suggested. Selection by SNP markers combined with phenotypic selection can reduce the amount of material to evaluate and shorten the time of breeding over conventional methods of phenotyping and selection. Finally, from an agronomic standpoint, it is interesting to note the high correlations of most of the quantitative traits measured as it opens an avenue for indirect selection of traits that are difficult to measure.

## 4.4 References


Ahn SN, Bollich CN, Tanksley SD (1992) RFLP tagging of a gene for aroma in rice. Theor Appl Genet 84: 825--828.

Bao J, Kong X, Xie T, Xu L (2004) Analysis of genotype and environmental effects on rice starch. 1. Apparent amylose content pasting, viscosity, and gel texture. J Agri Food Chem 52: 6010-6016.

Bao JS, Corke H, Sun M (2006) Microsatellites, single nucleotide polymorphisms and a sequence tagged site in starch-synthesizing genes in relation to starch physicochemical properties in nonwaxy rice (*Oryza sativa* L.) Theor Appl Genet 113: 1185-1196.

Bradbury LMT, Henry RJ, Jin Q, Reinke RF, Waters DLE (2005) A perfect marker for fragrance genotyping in rice. Mol Breeding 16: 279-283.

Bundock PC, Henry RJ (2004) Single nucleotide polymorphism, haplotype diversity and recombination in the Isa gene of barley. Theor. Appl. Genet. 109(3):543-51.

Chen MH, Bergman CJ, Fjellstrom RG (2003) SSSIIa locus genetic variation associated with alkali spreading value in international rice germplasm. In: Proceedings of the plant and animal genomes conference, vol XI. p 314.

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society. 363(1491) 557-572.

Comai L, Young K, Till BJ, Reynolds SH, Greene EA, Codomo CA, Enns LC, Johnson JE, Burtner C, Odden AR, Henikoff S (2004) Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. Plant J. 37(5):778-86.

Cordeiro GM, Christopher MJ, Henry RJ, Reinke RF (2002) Identification of microsatellite markers for fragrance in rice by analysis of the rice genome sequence. Mol Breed. 9:245-250.

Dreher K, Khairallah J, Ribaut M, Morris M (2003) "Money matters (I): Cost of field and laboratory procedures associated with conventional and marker-assisted maize breeding at CIMMYT". Molecular Breeding 11(3):221-234.

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Res. 14(9):1812-9.

Fjellstrom RG, Chen M, Bergman CJ,, McClung AM (2004) Single nucleotide polymorphism markers at the rice *alk* locus controlling alkali spreading value. In: Rice Technical Working Group Meeting Proceedings, February 29-March 4, 2004, New Orleans, LA.

Gao ZY, Zeng DL, Cui X, Zhou YH, Yan M, Huang D, Li JY, Qian Q (2003) Map-based cloning of the *alk* gene, which controls the GT of rice. Sci China C-Life Sci 46:661–668.

Garland S, Lewin L, Blakeney A, Reinke R (2000) PCR-based molecular markers for the fragrance gene in rice (*Oryza sativa* L.). Theor Appl Genet 101: 364-371.

Grain: World Markets and Trade May 2006.

Issiki M, Morino K, Okagaki RJ, Wressler SR, Izawa T, Shimamoto K (1998) A naturally occurring functional allele of the rice *waxy* locus has a GT to TT mutation at the 5'splice site of the first intron. Plant J. 15:133-138.

Jiang HW, Dian WM, Liu FY, Wu P (2004) Molecular cloning and expression analysis of three genes encoding starch synthase II in rice. Planta 218:1062–1070.

Jin Q, Waters DLE, Cordeiro GM, Henry RJ, Reinke RF (2003) A single nucleotide polymorphism (SNP) marker linked to the fragrance gene in rice (*Oryza sativa* L.). Plant Sci. 165: 359-364.

Juliano BO (1971) A simplified assay for milling rice amylose. Cereal Sci Today 16:334–336.

Kadaru SB, Yadav AS, Fjellstrom RG, Oard JH (2006) Alternative ecotilling protocol for rapid, cost effective single-nucleotide polymorphism discovery and genotyping in rice (*Oryza sativa* L.). Plant Molecular Biology Reporter 24:3-22.

Kadaru SB (2007) Identification of Molecular Markers and Association Mapping of Selected Loci Associated with Agronomic Traits in Rice. Ph.D. Dissertation. Louisiana State University and A&M College, Baton Rouge, LA.

Lanceras JC, Huang ZL, Naivikul O, Vanavichit A, Ruanjaichon V, Tragoonrung S (2000) Mapping of genes for cooking and eating qualities in Thai jasmine rice (KDML105). DNA Res. 7(2):93-101.

Little RR, Hilder GB, Dawson EH (1958) Differential effect of dilute alkali on 25 varieties of milled white rice. Cereal Chem 35:111–126.

McCouch SR, Teytelman L, Xu Y, Lobos KB, Clare K, Walton M, Fu B, Maghirang R, Li Z, Xing YZ, Zhang Q, Kono I, Yano M, Fjellstrom R, Declerck G, Schneider D, Cartinhour S, Ware D, Stein L (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). DNA Res 9:199–207.

Nakamura Y, Francisco Jr PB, Hosaka Y, Sato A, Sawada T, Kubo A, Fujita N (2005) Essential amino acid of starch synthase IIa differentiate amylopectin structure and starch quality between japonica and indica rice varieties. Plant Mol Biol 58:213–227.

Ramaligam J, HS Basharat, G Zhang (2002) STS microsatellite marker-assisted selection for bacterial blight resistance and *waxy* gene in rice, *Oryza sativa* L. Euphytica 127(2):255-260.

Sha XY, Linscombe SD, Bearb KF, Howard AM, Theunissen BW, Hoffpauir HL, Cramer SW (2000) Evaluation of specialty rice progenies for aroma. 92th Annual Research Report: Rice Research Station. Crowley: Louisiana Agricultural Experiment Station 55-58.

Sha XY (2005) Researchers make progress on new aromatic rice varieties. Rice Research Station News. Crowley: Louisiana Agricultural Experiment Station 2: 4.

Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, Wang G, Wang C, Qian L, Li X, Yu QB, Liu HJ, Chen DH, Gao JH, Huang H, Shi TL, Yang ZN (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol. 135(3):1198-205.

Shi C, Shen Z (1995) Genetic correlation analysis of agronomic traits in rice. Rice Genetics Newsletter 12:34.

Singh RK, Khush GS, Singh US, Singh AK, Singh S (2000) Breeding aromatic rice for high yield, improved aroma and grain quality. P71-106. In Aromatic rices. Singh RK, Singh US, Khush GS ed. Oxford and IBH Publishing Co. Pvt. LTD. New Delhi. 289 p.

Takashi M (2005) The map-based sequence of the rice genome. Nature 436: 793-800.

Tan YF, Li JX, Yu SB, Xing YZ, Xu CG, Zhang QF (1999) The three important traits for cooking and eating quality of rice grains are controlled by a single locus in an elite rice hybrid, Shanyou 63. Theor Appl Genet 99:642–648.

Temnykh S, Park WD, Ayres N, Cartinour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (2000) Mapping and genome organization of microsatellite sequence in rice (*Oryza sativa* L.). Theor Appl Genet 100:697–712.

Tian R, Jiang GH, Shen LH, Wang LQ, He YQ (2005) Mapping quantitative trait loci underlying the cooking and eating quality of rice using a DH population. Mol Breed 15:117–124.

Till BJ, Reynolds SH, Weil C, Springer N, Burtner C, Young K, Bowers E, Codomo CA, Enns LC, Odden AR, Greene EA, Comai L, Henikoff S (2004) Discovery of induced point mutations in maize genes by TILLING. BMC Plant Biol. 4(1):12.

Wang LQ, Liu WJ, Xu Y, He YQ, Luo LJ, Xing YZ, Xu CG, Zhang Q (2007) Genetic basis of 17 traits and viscosity parameters characterizing the eating and cooking quality of rice grain. Theor Appl Genet 115:463–476.

Waters DE, Henry RJ, Reinke RF, Fitzgerald MA (2006)Gelatinization temperature of rice explained by polymorphisms in starch synthase. Plant Biotechnology J 4:115–122.

Yamanaka S, Nakamura I, Watanabe KN, Sato Y (2004) Identification of SNPs in the *waxy* gene among glutinous rice cultivars and their evolutionary significance during the domestication process of rice. Theor Appl Genet 108: 1200-1204.

Yoshihashi T ( 2002) Quantitative analysis on 2-acetyl-1-pyrroline of an aromatic rice by stable isotope dilution method and model studies on its formation during cooking. J. Food Sci. 67:619-622.

Zhou PH, Tan YF, He YQ, Xu CG, Zhang Q (2003) Simultaneous improvement for four quality traits of Zhenshan 97, an elite parent of hybrid rice, by molecular marker-assisted selection. Theor Appl Genet 106: 326-331.

# CHAPTER 5 GENETIC ANALYSIS OF POLLEN STERILITY IN LINES DERIVED FROM A NATURAL OUTCROSS BETWEEN A LOUISIANA RED RICE BIOTYPE AND COMMERCIAL RICE

## 5.1 Introduction

### 5.1.1 Red Rice

Red rice (*Oryza sativa* L.) is considered a noxious weed of cultivated rice (*Oryza sativa* L.) in most temperate growing regions of the world (Fisher and Ramirez, 1993). Known in Italy since the last century, and for some time considered as a pathologic strain of the crop, red rice exhibits wide variability for numerous anatomical, biological and physiological features. In addition to being a noxious pest for rice production in the Southern United States (Oard et al., 2000), outcrossing between red rice biotypes and cultivated rice has posed a major challenge to effective weed management. Zhang (2006) documented that as high as 3.2% of outcrossed hybrids were observed in one field in Louisiana in 2003.

In spite of the aforementioned disadvantages, red rice can offer potential for diversification and improvement of the rice gene pool in the U.S. through introgression of pollen sterility and other traits. For example, Xiong et al. (1999) performed a mapping study between *indica* and photoperiod-sensitive wild rice. One QTL was detected accounting for 52% of the variation, suggesting that wild rice possesses a major gene for photoperiod sensitivity. Gealy (2006) hypothesized that difference in flowering time in U.S. red rice populations were associated with two independent homozygous complementary dominant photoperiod sensitive genes.

### 5.1.2 Male Sterility in Rice

Male sterility is a characteristic found widely in plants (Zuo et al., 2008) with more than 100 different male sterile mutants reported in rice (Bruskiewich et al., 2003). Male sterility prevents self-fertilization, but represents tremendous value for basic research on plant

reproduction and commercial exploitation of heterosis (Zhang et al., 2008). Recently, several male sterile rice mutants from different sources (Zuo et al., 2008, Zhang et al., 2008) have been phenotypically characterized and mapped. Inheritance studies on these different mutants showed a single recessive nuclear gene (Zuo et al., 2008, Zhang et al., 2008) controlling the expression of this trait.

### 5.1.3 Cytoplasmic Male Sterility in Rice

There are generally two kinds of sterility classes in rice: cytoplasmic male sterility (CMS) and nuclear male sterility (NMS) (Zhang et al., 2008). The CMS system is controlled by the interaction of cytoplasmic and nuclear genes (Virmani, 1994). The genetic factor(s) present in the cytoplasm have been reported to occur in mitochondrial DNA (Levings and Pring, 1976; Forde and Leaver, 1980; Kadowaki et al., 1986). This phenomenon in rice was first reported by Weeraratne in 1954 (Li et al., 2007). Shinjyo and Omura reported in 1966 the first CMS observed in elite rice cultivars. The CMS line was designated CMS-BT, being the product of an inter-subspecific cross between *indica* Chinsurah Boro II and *japonica* Taichung native 65 (Shinjyo, 1975). As early as 1964, Yuan Long Ping discovered male sterility in the *indica* variety Dong-Ting-Wan-Xian, but the breakthrough came in 1970 when he discovered a spontaneous-male-sterile plant referred to as CMS-WA in a wild population in Hainan Island, China (Yuan, 1977). Four years later, the first hybrid rice combination Nanyou-2 was released, showing higher yield potential as compared to inbred varieties. Since then, several CMS lines have been developed through interspecific, inter-subspecific, and intervarietal modes of hybridization (Li et al., 2007).

There are few reports on the molecular basis of CMS in rice (Liu et al., 1989).These reports provide evidence that rice CMS is controlled by variation in mitochondrial (mt) DNA (Virmani, 1994; Mignouna et al., 1987; Wang et al., 1987). Molecular properties of mt DNA in

CMS rice were reported by Kadowaki et al. (1988) who showed variation in mt DNA from ten strains of rice with male sterile cytoplasm. Restriction fragment length polymorphism (RFLP) was observed among mt DNAs analyzed, and eight different patterns among ten CMS lines were observed. These results suggested substantial variation in mt DNA which may be associated with CMS in rice. To further understand the molecular mechanism of CMS, Kadowaki and Oh-Fuchi (1989) cloned a DNA fragment (POSB 376) uniquely observed in CMS-BT type cytoplasm. Transcription of POSB 376 was altered by introduction of a fertility restorer gene, which suggested a role of mitochondrial RNA processing enzyme(s) in CMS. In addition, Liu et al., (1989) analyzed the translation products of the rice mitochondria *in vitro*, using electrophoresis and autoradiography. They found that a 22-kDa polypeptide was absent in the BT-CMS lines Nonghu 26A and Fengjin A compared to the fertile cytoplasm in Nonghu 26B and Fengjin B. These results suggested that mitochondrial genomes associated with fertility were mutated or rearranged. Huang et al. (2006) characterized the diversity of rice CMS cytoplasm and the mechanism of CMS using RFLP. They analyzed the sterile (A) and maintainer lines (B) of nine CMS sources that have been widely used in commercial production in China. The results showed that mitochondrial genomic differences were detected between A and B lines and within A lines in many functional genes.

**5.1.4 Hybrid Rice in China, Asia, and the U.S.**

CMS is broadly categorized into three types namely, CMS-WA, CMS-HL, and CMS-BT based on inheritance, morphology of abortive pollens and restoration-maintenance relationships (Li et al., 2007). However, commercial hybrid rice is almost exclusively based on CMS-WA, accounting for 90% of three-line hybrids in China (Yuan and Peng, 2005) and 100% outside China (Sattari et al., 2008). Moreover, the International Rice Research Institute relies heavily on

CMS-WA in the development of rice hybrids. This scenario opens the vulnerability of the rice hybrids to narrowing genetic base due to one common CMS background.

Hybrid varieties were first released in China in the 1970's. Chinese hybrids currently produce grain yields of 10% to 20% over the best local inbreds and occupy ~50% of the total rice acreage. Hybrid varieties are developed by breeders using two main approaches (Li et al., 2007). The "three-line method" uses cytoplasmic male sterile (CMS or A), "maintainer" (B), and fertility restoration (R) lines in crosses to develop new varieties. The "two-line" method uses photoperiod or temperature-sensitive male sterile and R lines for variety development. After early success with the first generation of hybrids in China, a yield plateau was first observed in the late 1990's (Cheng et al., 2007). Chinese breeders subsequently developed the so-called "super rice" varieties that increased yields over previous hybrids by 10-30% in some environments. The increased yield from super rice hybrids was attributed to the use of wide *indica/japonica* crosses and increased attention to improvement of agronomic traits of A, B, and R materials. For U.S. commercial hybrid rice, the company RiceTec is currently the major commercial entity that creates and markets hybrid rice varieties to U.S. rice producers. RiceTec states that their varieties enjoy a ~ 20 to 30% yield advantage over inbreds in LA and AR (RiceTec.com). Yield trials conducted by breeding programs in LA and AR indicate yield advantage of RiceTec hybrids versus inbreds at ~ 10%. Bayer Crop Science is currently establishing a research station near El Campo, Texas with emphasis on the U.S. hybrid rice market. Bayer has considerable experience in development of hybrid rice in Asia (www.bayercropscience.com). Under the umbrella name of "Arize", Bayer has recently commercialized different varieties in India, the Philippines, Indonesia, Vietnam, Bangladesh, Pakistan, and Brazil.

### 5.1.5 Initial Characterization of Red Rice–Clearfield Hybrid

Red rice biotypes carry the same AA genome as cultivated rice, so red rice weeds can hybridize with adapted commercial strains (Oard, 2005; Gealy et al., 2006). Weiqiang Zhang in Prof. Oard's laboratory conducted a survey in 2002 and 2003 to monitor the potential and consequences of hybridization between Newpath-resistant Clearfield rice and weedy rice in Louisiana commercial fields (Zhang et al., 2006). Seeds were collected in 2003 from potential outcrosses of Clearfield variety CL161 and red rice biotypes. In 2004 seeds from the putative outcrosses were planted at the Ben Hur Farm near Baton Rouge, LA and treated with Newpath herbicide. Survival of herbicide treatment along with whole-plant and molecular information provided strong evidence that a natural outcross between CL161 and a red rice biotype had occurred. One interesting putative $F_1$ plant that survived the herbicide treatment was substantially shorter (66 cm) versus other treated plants (~110 cm) and flowered ~ 10-14 days earlier than the other red rice-Clearfield hybrids. This plant was also partially sterile at ~ 30% seed set. Due to these unique characteristics and overall plant type, ~ 100 $F_2$ seeds from this plant were collected, dried to 12% moisture, and stored at room temperature. These $F_2$ materials were the foundation of this study. The objective of this research was to conduct a genetic analysis of pollen sterility/male sterility in a single $F_2$ population derived from a natural outcross of the red rice biotype described above with the commercial Louisiana variety Cleafield161.

## 5.2 Materials and Methods

### 5.2.1 Plant Material Characterization and Generation of $F_1$ and $F_2$ Populations

The starting plant material for this study consisted of 63 $F_2$ individuals originally derived from a natural outcross in 2003 between a Louisiana red rice biotype and the Clearfield variety CL161 (Zhang, 2006). Each $F_2$ plant was characterized in the greenhouse during the summer of 2007 for fertility/sterility traits: pollen viability and morphology by $I_2$-KI staining, and number of

fertile seeds from bagged panicles. Based on pollen evaluation, 58 out of 63 $F_2$ plants were completely pollen-sterile and the remaining plants were partially sterile (95% -99% pollen sterility). $I_2$-KI staining of pollen grains from the 63 $F_2$ plants showed that pollen varied from irregular and unstained, spherical and stained, and combinations of both. A majority of bagged panicles (59/63) yielded no selfed-seeds, indicating that the $F_2$ hybrid population was indeed sterile. To restore fertility and perform genetic analysis of pollen sterility, 32 100% pollen-sterile $F_2$ plants were crossed with Louisiana varieties Cocodrie and Trenasse. A total of 159 $F_1$ plants from 27 red rice x Cocodrie crosses and 156 $F_1$ plants from 25 red rice x Trenasse crosses were planted in the greenhouse during the Fall of 2007. Selection of the "best" $F_1$ plants among 315 individual plants for genetic analysis were based on overall all phenotype-maturity, plant height, panicle length and number, spikelets fertility. Thirty-three $F_1$ plants were selected for $F_2$ segregation and genetic analysis in the field. Phenotypic traits measured and collected for the $F_1$ plants were plant height, panicle number, and panicle length, filled and unfilled grains of the main tiller. $F_2$ seeds from the 33 $F_1$ plants were harvested and dried to 12% moisture for storage at 4°C until planting in the field during the summer of 2008.

### 5.2.2 Genetic Analysis and Characterization of Pollen Sterility and Additional Agronomic Traits in $F_2$ Populations

Segregation analyses for pollen sterility and two agronomic traits were performed in one $F_2$ population derived from a cross of red rice x Cocodrie. $F_2$ seeds produced from the population grown in the greenhouse were planted and evaluated in 2008 in field plots at the Crowley Rice Research Station. The planting date was April 16, 2008 and the study was terminated on August 18, 2008. Individual $F_2$ seeds were planted at 25 cm distances within rows and 38 cm distance between rows with a "Hege Vacuum Planter". The Cocodrie male parent was also planted for comparison. $F_2$ population sizes that ranged from 200-500 were dependent on the amount of

seeds harvested from each plant in the greenhouse. Normal cultural and management practices were strictly followed for fertilizer, herbicide, and insecticide applications. Standard water management was carried out to ensure normal and healthy growth of plants.

The data collected for eight traits in this study were the following: (1) plant height (PLTH) which was measured from the base of the rice plant to the tip of the highest panicle in cm, (2) panicle number (PANN) was determined by counting the number of seed-bearing panicles in each plant, (3) panicle length (PANL) was determined by measuring the main tiller/panicle and expressed in cm, (4) spikelet number (SPKN) was determined by counting the filled and unfilled spikelets of the main tiller, (6) seeds per panicle (SDPN) was determined by counting the filled grains of the main tiller in each plant in the population, (7) spikelet fertility (SPFT), expressed as percentage was determined by the formula SDPN/SPKN *100; individuals with SPFT values of < 20 were scored as sterile while those with SPFT values > 20 were scored as fertile. This scoring system was based on the seed set of $F_2$ ratooned plants brought in the greenhouse, where plants with at least 20% seed set in the field were sterile in the greenhouse, (8) presence of pubescent or glabrous leaves (PBGL) was determined by physically examining leaves on each plant, (9) pollen sterility (PNST) was determined by collecting at least 10 spikelets from each sterile plant and ~ 1000 pollen grains were stained with $I_2$-KI and viewed with a standard compound light microscope at 100X. A visual estimate of the percentage of $I_2$-KI-treated sterile pollen for each $F_2$ plant was determined in three separate microscopic fields and recorded as average percent pollen sterility. Individuals with PNST values < 2% were scored as fertile while those with values > 98% were scored as sterile (Sattari et al., 2008).

The agronomic data were analyzed using SAS software package 9.1.3. Descriptive statistics were computed using PROC UNIVARIATE. PROC CORR was used to calculate linear correlations between the different agronomic traits. The chi-square test was conducted to

test the goodness-of-fit using PROC FREQ in SAS. Frequency distribution figures were generated in Microsoft Excel 2007 program.

### 5.3 Results

#### 5.3.1 Pollen Sterility of $F_2$ Red Rice-Clearfield Outcross

The 63 $F_2$ plants derived from the red rice-CL161 outcross were generally short (<70cm), leaves were pubescent and tillers ranged from 5-22 under greenhouse conditions. The 63 $F_2$ plants exhibited a high degree of pollen sterility. For example, 92% of the plants were completely pollen sterile with overall mean sterility of 99.73%. The majority (94%) of the $F_2$ plants produced no seeds while the remaining plants each produced only a single seed under controlled condition (bagged panicle). These data confirmed that the 63 $F_2$ plants from the outcross between red rice and CL161 were male sterile.

#### 5.3.2 Phenotypic Characterization of $F_1$ Hybrids Derived from Red Rice-Clearfield 161 x Cocodrie or Trenasse

A total of 315 $F_1$ plants were produced from crosses of red rice-CL 161 x Cocodrie and red rice-CL161 x Trenasse. The plants were grown under greenhouse conditions during the spring of 2008 with 33 selected for advancement and characterization of six agronomic characters. Twenty-four selected $F_1$'s were red rice - CL161 x Cocodrie and 9 $F_1$'s were red rice - L161 x Trenasse. Genetic variability existed for most of the traits measured. For example, plant height varied substantially with a range of 22 cm. Plant height values were generally lower versus most commercial rice varieties that range in height from ~ 90 - 95 cm. The male parents for these crosses (Cocodrie and Trenasse) which are popular LA varieties exhibited mean heights of 93 and 94 cm, which were at least 15 cm greater than the average of the $F_1$'s. This result was not unexpected because the original male sterile parent of the outcross measured ~ 66 cm in height. For panicle number and length, the ranges were 11 panicles and 7 cm, respectively.

Panicle length as very similar to the male parents, but not for tiller number due to differences in growing conditions. The range for percent seed fertility for the bagged and unbagged samples were 84% and 90.8%, respectively. These results indicated the high phenotypic variation for this trait. Means for the bagged panicles for SDPN and SPFT data were significantly lower than the unbagged samples by 47 seeds and 27%, respectively. A difference of 9 seeds on the average was noted for SPKN. These results may be due to environmental differences that affected the panicle growth and seed setting in the bagged versus. the unbagged conditions (Sattari et al., 2008).

Table 5.1 Descriptive statistics for the seven agronomic traits of 33 selected $F_1$'s derived from natural outcrosses between red rice-clearfield 161 x Cocodrie or Trenasse, greenhouse, LSU Baton Rouge, Fall 2007.

| Trait | Unit | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| PLTH | cm | 33 | 78 | 5.64 | 69 | 91 |
| PANN | - | 33 | 9 | 2.74 | 5 | 16 |
| PANL | cm | 33 | 22 | 1.84 | 19 | 26 |
| SPKN[a] | - | 33 | 141 | 35.83 | 58 | 203 |
| SPKN[b] | - | 33 | 150 | 27.04 | 107 | 232 |
| SDPN[a] | - | 33 | 62 | 40.99 | 0 | 145 |
| SDPN[b] | - | 33 | 109 | 38.88 | 1 | 171 |
| SPFT[a] | percent | 33 | 45 | 26.31 | 0 | 84 |
| SPFT[b] | percent | 33 | 72 | 22.47 | 0.77 | 92 |

PLTH-plant height, PANN-panicle number, PANL-panicle length, SPKN-spikelet number per panicle, SDPN-filled seeds per panicle, SPFT-spikelet fertility samples, SD-standard deviation
[a]data from bagged panicle, [b]data from unbagged panicle

Table 5.2 Correlation analysis for the seven agronomic traits of 33 selected $F_1$'s derived from natural outcrosses between red rice-clearfield 161 x Cocodrie or Trenasse, greenhouse, LSU Baton Rouge, Fall 2007.

| | Pearson Correlation Coefficients, Prob > \|r\| under H0: Rho=0 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | PLTH | PANN | PANL | SPKN[a] | SPKN[b] | SDPN[a] | SDPN[b] | SPFT[a] |
| PANN | 0.08648 | | | | | | | |
| | (0.6323) | | | | | | | |
| PANL | 0.69051 | 0.37805 | | | | | | |
| | (<.0001) | (0.0301) | | | | | | |
| SPKN[a] | 0.30692 | 0.33797 | 0.30308 | | | | | |
| | (0.0823) | (0.0544) | (0.0864) | | | | | |
| SPKN[b] | 0.61210 | 0.38303 | 0.72409 | 0.26061 | | | | |
| | (0.0002) | (0.0278) | (<.0001) | (0.1430) | | | | |
| SDPN[a] | 0.25620 | 0.45760 | 0.09147 | 0.36195 | -0.03434 | | | |
| | (0.1501) | (0.0074) | (0.6127) | (0.0385) | (0.8495) | | | |
| SDPN[b] | 0.61569 | 0.32221 | 0.45304 | 0.01467 | 0.60452 | 0.45245 | | |
| | (0.0001) | (0.0674) | (0.0081) | (0.9354) | (0.0002) | (0.0082) | | |
| SPFT[a] | 0.10030 | 0.33243 | -0.02351 | -0.07598 | -0.18029 | 0.87240 | 0.38796 | |
| | (0.5787) | (0.0587) | (0.8967) | (0.6743) | (0.3154) | (<.0001) | (0.0257) | |
| SPFT[b] | 0.34139 | 0.17998 | 0.08769 | -0.14595 | 0.12239 | 0.57865 | 0.85690 | 0.59848 |
| | (0.0519) | (0.3162) | (0.6275) | (0.4177) | (0.4974) | 0.0004) | (<.0001) | (0.0002) |

PLTH-plant height, PANN-panicle number, PANL-panicle length, SPKN-spikelet number per panicle, SDPN-filled seeds per panicle, SPFT-spikelet fertility, SD-standard deviation
[a]- bagged panicle sample, [b]-unbagged panicle sample

Linear correlation analysis was performed between the nine phenotypic traits of the 33 $F_1$'s using Pearson's coefficient. Table 5.2 presents the summary of the analysis. Plant height was highly associated with panicle length, spikelet number and seeds per panicle of the

unbagged sample, but not correlated with the other traits. Similarly, panicle length was highly correlated with spikelet number, but only moderately correlated with seeds per panicle, suggesting maximum seed production for this population requires long panicles. The bagged and unbagged samples were not correlated with spikelet number or seeds per panicle. This result might be due to potential effect of bagging on seed production (Sattari et al., 2008). Because only a few fertile pollen grains are sufficient for normal fertilization and seed set in rice (Sattari et al., 2008), bagging of panicles may have affected pollen fertility and seed set of the plant examined in this study.

### 5.3.3 Descriptive Statistics and Correlation Analysis of Agronomic Traits in $F_2$ Population

Table 5.3 shows the means, standard deviation, minimum and maximum values for each of six agronomic traits examined in 478 $F_2$ individuals derived from the red rice-CL161 x Cocodrie. High levels of phenotypic variations were observed for all six traits as would be expected from a "wide cross" between an adapted elite commercial cultivar and a weedy, red rice biotype. For example, PLTH exhibited a range of 57 cm. In comparison to the male parent, 92% of the $F_2$ plants had the same height or were shorter than Cocodrie. PANN also exhibited high phenotypic variation with a range of 44 panicles, and the mean PANN value is greater by 6 tillers than the male parent Cocodrie. It is interesting to note that 76% of $F_2$ population had more panicles than Cocodrie, indicating a potential for heterosis in the original cross. Similarly, for PANL and SDPN, 71% and 59%, respectively, of the $F_2$ populations produced larger panicles and more seeds per panicle compared to the male parent Cocodrie. Both traits also showed high phenotypic variability. Approximately one-quarter (26%) of the population exhibited a glabrous leaf texture, the ideal leaf texture for rice varieties in the U.S. For pollen sterility evaluation, 37 plants were essentially sterile (99%), and 24 $F_2$ plants were partially sterile (80%) based on spikelet fertility.

Table 5.3 Descriptive statistics of six agronomic traits in one $F_2$ population derived from a red rice–Clearfield 161 x Cocodrie cross and the male parent Cocodrie, evaluated at Rice Research Station, Crowley, LA, 2008.

| Trait | Unit | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| $F_2$ population | | | | | | |
| PLTH | cm | 478 | 80 | 9.92 | 50 | 107 |
| PANN | - | 478 | 21 | 7.91 | 2 | 46 |
| PANL | cm | 478 | 22 | 2.20 | 15.2 | 27.5 |
| SPKN | - | 478 | 176 | 47.40 | 69 | 338 |
| SDPN | - | 478 | 128 | 58.45 | 7 | 294 |
| SPFT | percent | 478 | 71.60 | 25.51 | 4.50 | 98.85 |
| Parent-Cocodrie | | | | | | |
| PLTH | cm | 10 | 93 | 4 | 88 | 101 |
| PANN | - | 10 | 15 | 3 | 12 | 22 |
| PANL | cm | 10 | 21 | 1 | 20 | 23 |
| SDPN | - | 10 | 164 | 22 | 121 | 203 |

PLTH-plant height, PANN-panicle number, PANL-panicle length,
SPKN-spikelet number per panicle, SDPN-filled seeds per panicle, SPFT-spikelet fertility

Frequency distributions of the six traits examined among the 478 $F_2$ individuals derived from the red rice biotype described above and the Clearfield cultivar CL161 are shown in Figure 5.1. PANN and SPKN followed a normal distribution and PLTH and PANN were close to normal. SPFT showed a skewed distribution toward high fertility levels. The arrow in Fig 5.1 points to the mean value of the male parent Cocodrie for PLTH, PANL, PANN, and SDPN. As shown in the graphs, a majority of the lines produced greater values than the male parent Cocodrie, an indication of high genetic variability and potential heterosis or hybrid vigor in the population. For SDPN, it is interesting to note a bi-modal distribution similar to the pattern observed by Sattari et al. (2008) in an $F_2$ population derived from the Gambiaca - CMS source for pollen fertility.

Correlation analysis of the six quantitative traits using Pearson Correlation Coefficients revealed interesting results (Table 5.4). For example, PLTH was found to be associated with multiple traits including panicle length (PANL), number of spikelets per panicle (SPKN), seeds per panicle (SDPN) and moderately associated with panicle number (PANN) and spikelet fertility (SPFT). Similarly, PLTH was highly associated with PANL, SPKN, and SDPN in the selected $F_1$'s. This would suggest the possibility of using PLTH as indirect selection for the other quantitative traits that are more labor-intensive and time-consuming. Separate data from other populations and lines are needed to further confirm these initial observations and findings. Most correlation analyses involved only the yield and yield-related traits. Shi (1995) for example found high association between density of panicle versus grains per panicle as well as spikelet fertility versus density of panicle. No report has been published on correlation of plant height vs. yield-related traits similar to what was found in this study. These results may be due to high phenotypic variability in the population as brought about by the wide cross of red rice-CL 161 x Cocodrie. Moreover, PANL was strongly associated with SPAN indicating that the longer the panicle, the more spikelets are expected, but is only moderately correlated with SDPN due to the sterility of some spikelets. These association results warrant further verification in other populations under different environmental conditions.

### 5.3.4 Genetic Analysis of Pollen Sterility and Selected Agronomic Traits

Chi-squared analyses were carried out to determine the goodness-of-fit for segregation of PBGL, SPFT, and PNST observed in the 478 $F_2$ individuals derived from the natural outcross of the red rice biotype and the CL161 Clearfield cultivar (Table 5.5). PBGL values were recorded in this study to determine if this trait followed a Mendelian segregation pattern in the $F_2$ population. As expected, pubescent to glabrous leaves satisfactorily fit a Chi-square goodness-

of-fit ratio of 3:1 pubescent: glabrous, indicating that PBGL in this population was governed by a single dominant nuclear-encoded gene as reported by Gealy (2006).

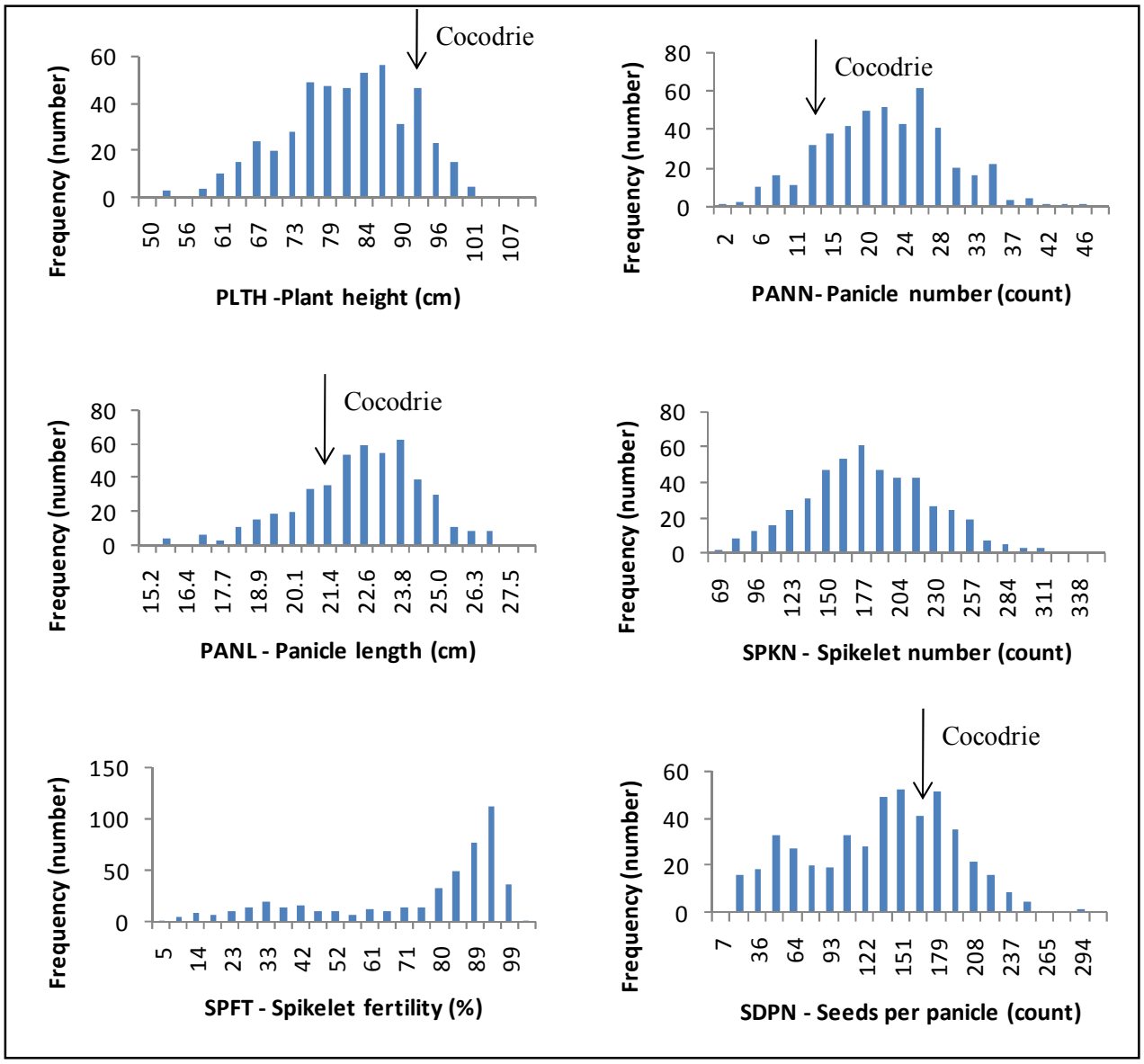



Fig 5.1 Frequency distribution of the six quantitative traits in the $F_2$ population (n=478) of a red rice –Clearfield 161 x Cocodrie cross, evaluated at Rice Research Station, Crowley, LA, 2008.

*arrow indicates the mean of the male parent, Cocodrie

Table 5.4 Correlation analysis of the six quantitative traits in the $F_2$ population of a red rice-clearfield 161 x Cocodrie cross, evaluated in Crowley, LA, 2008.

| Pearson Correlation Coefficients, N = 478<br>Prob > \|r\| under H0: Rho=0 | | | | | |
|---|---|---|---|---|---|
| | PLTH | PANN | PANL | SPKN | SDPN |
| PANN | 0.38403 | | | | |
| | (<.0001) | | | | |
| PANL | 0.53409 | 0.37139 | | | |
| | (<.00010) | (<.0001) | | | |
| SPKN | 0.54621 | 0.30840 | 0.60741 | | |
| | (<.0001) | (<.0001) | (<.0001) | | |
| SDPN | 0.64486 | 0.22466 | 0.35814 | 0.67405 | |
| | (<.0001) | (<.0001) | (<.0001) | (<.0001) | |
| SPFT | 0.46246 | 0.05905 | -0.00375 | 0.12429 | 0.79013 |
| | (<.0001) | (0.1975) | (0.9349) | (0.0065) | (<.0001) |

PLTH-plant height, PANN-panicle number, PANL-panicle length, SPKN-spikelet number per panicle, SDPN-fertile seeds per panicle, SPFT-spikelet fertility

The SPFT values indicated that seed fertility in this population was dominant over sterility when the female parent was pollen-sterile. SPFT satisfactorily fit a 15F:1S chi-square ratio indicating that seed fertility for this population was controlled by a two dominant Mendelian genes. These results are consistent with previous reports on CMS fertility restoration studies that indicate digenic inheritance with dominant duplicate gene action for CMS-Dissi, Gambiaca (Sattari et al., 2008), and CMS-WA (Sattari et al., 2008; Virmani et al., 1986; Govina Raj and Virmani 1988; Bharaj et al., 1991, 1995; and Yao et al., 1997). The other two CMS types-BT and HL were found to follow a single gene inheritance (Shinjyo 1969; Huang et al.; 2000). Finally, segregation analysis of PNST showed a satisfactory chi-square fit of 15F:1S

indicating that pollen sterility was therefore controlled by two recessive genes. Genetic control of pollen sterility in previous studies in rice has been found to be controlled primarily by monogenic recessive genes (Razzaque, 1974; Pavithran and Mohandas, 1976; Trees and Rutger, 1978; Singh and Ikehashi, 1981), although Pavithran and Mohandas (1976) also observed trigenic segregation of an induced male sterile mutant. Male sterility in rice mutants from three different studies was reported recently to be controlled by a single recessive gene (Zuo et al., 2008; Zhang et al., 2008). Although results in the current study deviate from the previous reports, additional confirmation and analysis are warranted.

Table 5.5 Segregation analyses of pubescence, spikelet fertility, and pollen sterility among 478 $F_2$ individuals derived from the Red Rice-CL161 x Cocodrie cross, Rice Research Station, Crowley, LA, 2008.

| Trait | Observed | | Expected | | Chi-Square | P-value |
|---|---|---|---|---|---|---|
| | No. of plants | | | | | |
| PBGL | Pubescent | Glabrous | Pubescent | Glabrous | (3:1) | |
| | 354 | 124 | 358 | 120 | 0.2259 | 0.6345 |
| | Fertile | Sterile | Fertile | Sterile | (15:1) | |
| SPFT | 454 | 24 | 448 | 30 | 0.2324 | 0.2669 |
| PNST | 411 | 37 | 448 | 30 | 1.8126 | 0.1782 |

PBGL-pubescent/glabrous leaves, SPFT-spikelet fertility, PNST-pollen sterility

**5.4 Discussion**

Segregation analysis is a classical but powerful method to determine inheritance of a trait in $F_2$ populations. In hybrid rice for example, most genetic and inheritance studies rely on developing hybrids (AxR) and growing large $F_2$ populations to determine mode of inheritance. Previous studies reported that CMS-WA (Satarri et al., 2008, Virmani et al., 1986; Govina Raj and Virmani, 1988; Bharaj et al., 1991, 1995; and Yao et al., 1997), and recently CMS Dissi, and Gambiaca (Sattari et al., 2008) followed a 15F:1S segregation ratio indicating that fertility restoration was controlled by two dominant genes while the CMS-BT and CMS-HL segregated

into 3F:1S, indicating that fertility/sterility was governed by one major gene (Shinjyo, 1969; Huang et al., 2000). Our result is consistent with the CMS-WA and CMS Dissi and Gambiaca in a 15F:1S ratio and is believed to be controlled by two dominant genes.

The use of pollen-sterile lines has been successfully exploited in China over the last 30 years to develop commercial hybrid cultivars that exceed the yield of pure-line cultivars by ~ 15% to 20% (Virmani, 1994). However, the number of pollen-sterile lines is limited, so extensive research has been conducted in China to discover or develop new pollen-sterile lines. The focus of the current study was to initiate a genetic analysis of a single hybrid plant derived from a cross between a weedy red rice and an elite commercial cultivar. The red rice hybrid plant exhibited unique and interesting characteristics in terms of fertility and agronomics traits that warranted further investigation of CMS-based sterility.

The results from this study indicate that seed fertility was governed by two dominant genes similar to CMS-WA while pollen sterility was controlled by two recessive genes. These results are characteristics of genetic/cytoplasmic male sterile lines used to produce hybrid cultivars in China and elsewhere. Overall, the results suggest that the selected red rice-CL161 hybrids could be developed as possible new sources of cytoplasmic genetic male sterile lines for hybrid rice breeding in Louisiana. However, additional research at the whole-plant and molecular level is needed to confirm if lines developed from this study represent a new source of CMS similar to CMS-WA systems developed in Asia. With recent advances in molecular biology, tools for comparison at the molecular level are possible through sequence data of different types/sources of CMS and the one reported here. Finally, from an agronomic standpoint, it is interesting to note the high correlations of most of the quantitative traits measured. This opens an avenue for indirect selection of traits that are difficult to measure.

## 5.5 References


Bharaj TS, Bains SS, Sidhu GS, Gagneja MR (1991) Genetics of fertility restoration of 'wild abortive' cytoplasmic male sterility in rice *Oryza sativa* L. Euphytica 56:199-203.

Bharaj TS, Virmani SS, Khush GS (1995) Chromosomal location of fertility restoring genes for 'wild abortive' cytoplasmic male sterility using primary trisomics in rice. Euphytica 83:169-173.

Bruskiewich RM, Cosico AB, Eusebio W, Portugal AM, Ramos LM, Reyes MT, et al., 2003. Linking genotype to phenotype: The International Rice Information System (IRIS). Bioinformatics 19 (Suppl. 1): 63-65.

Cheng SH, Zhang JY, Fan YY, Du JH, Cao LY (2007) Progress in research and development on hybrid rice: A super-domesticate in China. Annals of Botany 100(5):959-966.

Fisher AJ, Ramirez A (1993) Red rice (*Oryza sativa* L.): Competition studies for management decisions. Int J Pest Manage 39:133-138.

Forde BG. Leaver CJ (1980) Nuclear and cytoplasmic genes controlling synthesis of variant mitochondrial polypeptides in male sterile maize. Proc Natl Acad Sci USA 77:418-422.

Gealy DR, Yan W, Rutger JW (2006) Red Rice (*Oryza sativa* L.) Plant Types Affect Growth, Coloration, and Flowering Characteristics of First- and Second-Generation Crosses with Rice. Weed Technology 20:839-852.

Govinda Raj K and Virmani SS (1988) Genetics of fertility restoration of 'WA' type cytoplasmic male sterility in rice. Crop Sci 28:787-792.

Huang QY, He QY, Jing RC, Zhu RS, Zhu YG (2000) Mapping of the nuclear fertility restorer gene for HL cytoplasmic male sterility in rice using microsatellite markers. Chin Sci Bull 45(5):430-432.

Huang W, Wang L, Yi P, Tan XL, Zhang XM, Zhang ZJ, LI YS, Zhu YG (2006) RFLP analysis of mitochondrial genome of CMS rice. Acta Genetica Sinica 33 (4) 330-338.

Kadowaki K, Harada K (1989) Differential organization of mitochondrial genes in rice with normal male sterile cytoplasms. Jpn J Breed 39: 179-186.

Kadowaki K, Oh-Fuchi T (1989) Molecular analysis of cytoplasmic male sterility of rice ( *Oryza sativa* L.). (Eng.). In: Iyama S, Takeda G, (eds) Breeding research the key to the survival of the earth. Tsukuba, Japan. pp 527-530.

Kadowaki K, Ishige T, Suzuki S, Harada K, Shinjyo C (1986) Differences in the characteristics of mitochondrial DNA below a normal and male sterile cytoplasms of japonica rice. Jpn J Breed 36: 333-339.

Levings CS III, Pring DR (1976) Restriction endonucleases of mitochondrial DNA from normal and Texas cytoplasmic male-sterile maize. Science 193:158-160.

Li S, Yang D, and Zhu Y (2007) Characterization and use of male sterility in hybrid rice breeding. Journal of Integrative Plant Biology 49 (6): 791-804.

Liu Z, Zha S, Zhan Q, Chen Y (1989) Mitochodrial genome translation products and cytoplasmic male sterility in rice. Acta Genetica Sinica 16(1) 19 (Abstract).

Mignouna H, Virmani SS, Briquet M (1987) Mitochondrial DNA modifications associated with cyptoplasmic male sterility in rice Theor Appl Genet (74) 666-669.

Oard JH, Cohn MA, Linscombe S, Gealy DR, Kenneth G (2000) Field evaluation of seed production , shattering, and dormancy in hybrid population of transgenic rice (*Oryza sativa* L) and the weed, red rice (*Oryza sativa* L.). Plant Sci 157:13-22.

Oard J (2005) Update on outcrossing between Clearfield rice and red rice. Rice Research Station News. La. State Univ. Agric. Cent. 2(4):1-2.

Pavithran K, Mohandas C (1976) Genetics of male sterility in rice. J Hered 67: 252.

Razzaque CA (1974) Genetics of male sterility in rice. Indian J Genet 34:303-308.

Sattari M, Kathiresan A, Gregorio GB and Virmani SS (2008) Comparative genetic analysis and molecular mapping of fertility restoration genes for WA, Dissi, and Gambiaca cytoplasmic male sterility systems in rice. Euphytica 160:305-315.

Shi C, Shen Z (1995) Genetic correlation analysis of agronomic traits in rice. Rice Genetics Newsletter 12:34.

Shinjyo C (1969) Cytoplasmic-genetic male sterility in cultivated rice (*Oryza sativa* L.) II. The inheritance of male sterility. Jpn J Genet 44:149-156.

Shinjyo C (1975) Genetical studies of cytoplasmic male sterility and fertility restoration in rice (*Oryza sativa* L.) Sci Bull Coll Agric Univ Ryukus 22:1-57.

Singh RJ, Ikehashi H (1981) Monogenic male sterility in rice: induction, identification and inheritance. Crop Sci 21:286-288.

Trees SC, Rutger JN (1978) Inheritance of four genetic male steriles in rice. J Hered 69:270-272.

Virmani SS (1994) Monographs on Theoretical and Applied Genetics 22. Springer-Verlag.

Virmani SS, Govinda Raj K, Casal C, Dalmacio RD, Aurin PA (1986) Current knowledge and future outlook on cytoplasmic genetic male sterility and fertility restoration in rice. In IRRI (ed) Rice Genetics. International Rice Research Institute, Manila, Philippines. pp 633-647.

Xiong LZ, Liu KD, Dai KX, Xu DC, Zhang Q (1999) Identification of genetic factors controlling domestication-related traits of rice using an $F_2$ population of a cross between *Oryza sativa* and *O. rufipogon*. Theor Appl. Genet. 98:243-51.

Yao FY, Xu CG, Yu SB, Li JX, Gao YJ, Li XH, Zhang Q (1997) Mapping and genetic analysis of two fertility restorer loci in the wild abortive cytoplasmic male sterility system of rice (*Oryza sativa* L) Euphytica 98:183-187.

Yuan LP (1977) The execution and theory of developing hybrid rice. Chin Agric Sci 1:27-31.

Yuan LP and Peng JM (2005) Hybrid Rice and World Food Security. China Science and Technology Press, Beijing, China.

Wang B, Li YN, Chen W, Li DN (1987) Plasmid-like mitochondrial DNA associated with cytoplasmic male sterility. Rice Genet Newsl (4) 118-120.

Zhang W, Linscombe SD, Webster E, Tan S, Oard J ( 2006) Risk assessment of the transfer of imazethapyr herbicide tolerance from Clearfield rice to red rice (*Oryza sativa*). Euphytica 152:75-86.

Zhang Y, Mao JX, Yang K, Li YF, Zhang J, Huang YX, Shen FC, and Zhang CD (2008) Characterization and mapping of a male-sterility mutant, *tapetum desquamation* (t), in rice. Genome 51:368-374.

Zhang Y, Li YF, Zhang J, Shen FC, Huang YX, and Wu Z (2008) Characterization and mapping of anther advanced dehiscence (t) in rice. J. Genet. Genomics 35:177-182.

Zuo L, Li SC, Chu M, Wang S, Deng Q, Ding L, Zhang J, Wen Y, Zheng A, and Li P (2008) Phenotypic characterization, genetic analysis, and molecular mapping of a new mutant gene for male sterility in rice. Genome 51:303-308.

# CHAPTER 6 SUMMARY AND CONCLUSIONS

## 6.1 Mixed Model (TASSEL) and GLMSelect Procedures for Association Genetics

The SAS GLMSelect and TASSEL mixed model approaches were evaluated for ability to identify candidate markers associated with heading date, head rice, and amylose content among 192 lines of inbred rice lines grown in replicated trials at 5 locations in AR, LA, MO, MS, and TX. The TASSEL approach was able to identify individual candidate markers detected across locations, but the overall ability to model complex phenotypic variation was poor. This low performance level was attributed to the identification of candidate markers one at a time by TASSEL that ignored multiple loci effects and two-way interactions (epistasis). Therefore, selected marker-effects from TASSEL were subsequently evaluated by GLMSelect that allowed consideration of epistasis and allowed selection of effects by criteria other than F statistics. The combined TASSEL-GLMSelect procedure proved optimal in terms of relatively high adjusted $R^2$ values, minimal selected effects, and accounting for kinship effects to reduce Type I errors within a narrow germplasm base.

The overall results suggest that a combined mixed model-multiple regression procedure that considers epistasis with a validation step should be explored in future studies for association studies in rice and other crops. Although the number of selected markers from this study was too large for an effective and practical breeding program, increasing the number of markers for analysis in this population may increase power and precision. Our study showed that while allelic diversity of microsatellite was relatively low in this narrow germplasm, the frequency of rare alleles was very high. The use of SNP markers would help alleviate the problem associated with rare alleles because they are more prevalent than microsatellite markers and are amenable to high throughput analysis. We conclude that high density SNP markers coupled with the methods outlined in this study should be further explored for association genetics in rice.

### 6.2 Support Vector Regression (SVR)

SVR was implemented in R-software to evaluate prediction accuracy and power of DNA markers associated with heading date, head rice, and amylose content among 192 lines of inbred rice lines in the U.S. grown in 5 states (AR, LA, MO, MS, and TX). The results showed that the modified SVR procedure produced high levels of accuracy using radial basis kernel which is consistent with previous studies of maize inbred lines (De Baets et al., 2008). High levels of power were detected with the SVR procedure for all three complex traits compared to a multiple regression approach carried out by GLMSelect. The SVR approach for marker selection was supported by previous QTL mapping studies that identified the same genetic regions for the three traits evaluated. The outcome and procedures developed in this study could provide insights and guidance for the development of model simulations and design of future validation experiments.

SSR markers were utilized in this study, but dense SNP "chips" and maps with > 40,000 markers will soon become available. One advantage of SVR under those circumstances may be the ability to obtain sparse solutions with relatively few variables versus other methods involving large data sets (Vapnik, 1995). Another advantage of SVR maybe an internal validation step to estimate parameters that give rise to high power and precision. All results obtained from this study suggested that SVR exhibited desirable features for association genetics in rice and other inbred species should be further explored and developed for optimum power and prediction accuracy of marker-trait association.

### 6.3 SNP Markers for Marker-Assisted Selection

SNP markers were evaluated for aroma, amylose content (AC), and gelatinization temperature (GT) in marker-assisted selection of LSU breeding lines aimed at developing new elite aromatic varieties. Results from the molecular and agronomic analyses clearly showed that the SNP marker approach enriched the frequency of desired alleles in lines with good plant type

in only two generations. The strong implication is that marker-assisted methods for certain traits such as aroma, AC and GT, could speed up and increase efficiency in development of new Louisiana aromatic breeding lines. Results from this study showcased the ability of molecular markers to screen and select individuals that possess the quality traits of interest in rice. Without the use of markers such as those used in this study, identification of individual plants or lines that carry all favorable alleles for aroma, low GT, and low AC would be labor-intensive, costly and time-consuming.

### 6.4 Genetic Analysis of Pollen Sterility from Natural Outcross of Weedy and Commercial Rice

A genetic analysis was carried out for pollen sterility/male fertility in a single $F_2$ population derived from a natural outcross of a red rice biotype with the commercial Louisiana variety Clearfield161. Based on segregation analysis using the Chi-Square test, seed fertility restoration was governed by two dominant genes, while pollen sterility in contrast was controlled by two recessive genes. These results are typical of genetic/cytoplasmic male sterile lines used to produce hybrid cultivars in China and elsewhere. Overall, the results suggest that the selected red rice-CL161 hybrid and the resulting progeny may be developed further as possible new sources of cytoplasmic genetic male sterile lines for hybrid rice breeding in Louisiana through a series of backcrosses. However, additional research at the whole-plant and molecular level is needed to confirm if lines developed from this study represent a new source of CMS similar to CMS-WA systems in Asia. With recent advances in molecular biology, tools for comparison at the molecular level are possible through sequence data of different types/sources of CMS and the one reported in this study.

## APPENDIX R SOURCE CODE FOR THE SVR PROCEDURES

**Step 1**: **Type the following commands in the R console window to assess the SVR procedure**

```
install.packages('e1071') + Enter
library("e1071") + Enter
svm
```

**Step 2: Read the raw amylose content data by the "read.Table" command:**d<-

```
read.Table("ac.dat", header = TRUE)
    y<-d[,1]
    x0<-d[,-1]
```

**Step 3: Fit a SVR on each individual marker.**

```
for (i in 1:194){
x<-x0[,i]
sm<- svm(x, y, kernel ="radial", scale=FALSE, cost = 200, nu = .001, cachesize = 100,
tolerance = 0.01, epsilon = .01, fitted = TRUE, cross=10)
py<-predict(sm, x)
R2<-1-sum((y-py)**2)/2601.3    }
```

**Step 4: The $R^2$s obtained from step 3 are ordered in descending manner. Thirty-five of the ordered $R^2$s are found to be greater than 0.00001. The markers corresponding to these 35 $R^2$s are then kept and interacted pairwisely, resulting in 630 terms, along with the original 35 terms, in total. The forward selection procedure is to be conducted on these 630 terms in the SVR model setting.**

```
d<-read.Table(file="actotal.dat", header =TRUE)
y<-d[,1]
x0<-d[,-1]
for (i in 1:630){
x<-x0[,i]
sm<- svm(x, y, kernel ="radial", scale=TRUE, degree = 3, cost = 200, nu = .00001, cachesize
= 100, tolerance = 0.01, epsilon = .01, shrinking = TRUE, fitted = TRUE, cross=10)
py<-predict(sm, x)
R2<-1-sum((y-py)**2)/2601.3
}
```

**Step 5: Forward selection is conducted**

```
d<-read.Table(file="ac630.dat")
y<-d[,1]
x0<-d[,-1]
x1<-x0[,1]
for (i in 1:630){
x1<-cbind(x1,x0[,i+1])
x<-x1
```

```
sm<- svm(x, y, kernel =“radial”, scale=TRUE, degree = 3, cost = 200, nu = .00001, cachesize
= 100, tolerance = 0.001, epsilon = .001, shrinking = TRUE, fitted = TRUE, cross=10
py<-predict(sm, x)
R^2[i+1]<-1-sum((py-y)**2)/2601.3
mc<-max(R^2)
n<- ncol(s2)
if (R[i+1]-mc<0.02){x1=x[, -n]}
if (R[i+1]-mc>0.02) {
R^2<-cbind(R^2,R[i+1]) }}
```

**Step 6: The MSE, squared correlation coefficient, and $R^2$ are computed**

```
d<-read.Table(file=“all.dat”, header=TRUE)
y<-d[, 1]
x0<-d[,-1]
x<-x0
sm<- svm(x, y, kernel =“sigmoid”, scale=TRUE, degree = 3, cost = .1, nu = 1, cachesize = 40,
tolerance = 0.1, epsilon = .1, shrinking = TRUE, fitted = TRUE, cross=10)
py<-predict(sm, x)
R^2<-1-sum((py-y)**2)/2601.3
mse<-sum((py-y)**2)/193
```

**Step 7**: **Compute power**

```
require(pwr)
pw<-read.Table (file=“r^2svm.dat”)
pw1<-sqrt(pw)
pwr.r.test(r=pw1, n=194, sig.level=0.05, alternative=“two.sided”)
```

## VITA

Samuel A. Ordonez Jr. was born in Jones, Isabela, Philippines, in 1975. He is the third child among four siblings. He started his formal education at Santiago Adventist Elementary School (SAES) graduating as class valedictorian and again as class valedictorian from Philippine Union College-Northeast Luzon Campus. He entered Central Luzon State University (CLSU) for his undergraduate and took a bachelor of science degree in biology and graduated *cum laude* in 1997.

Upon graduation he worked as Research Assistant at the Bureau of Postharvest Research and Extension (BPRE). He transferred to Philippine Rice Research Institute (PhilRice) a few months later and worked as project research assistant in the hybrid rice breeding program under Dr. Edilberto D. Redona. He was awarded a full time scholarship by the Department of Agriculture-Bureau of Agricultural Research (DA-BAR) to pursue a master's degree in plant breeding at the University of the Philippines, Los Banos (UPLB), and graduated in 2003. Upon graduation he resumed work at PhilRice as Plant Breeder - Senior Science Research Specialist (SrSRS) and was primarily engaged in hybrid rice breeding and genetics. In Spring 2006 he was accepted in the doctoral program at Louisiana State University and was a awarded an assistantship by the Department of Agronomy through Dr. James H. Oard, his major professor. After 3.5 years of hard work and dedication, he is finally graduating in Summer of 2008. Sam, as is he is fondly called, has accepted a Senior Biologist position from DowAgroSciences, where he will continue his career.

He is married to Carol and they have two wonderful sons-Samuel III (9 yrs old) and Ferdinand Amadeo (6 yrs old).