CHARACTERIZATION AND VALIDATION OF FLOWERING TIME GENES IN A DIVERSITY PANEL OF RICE (*Oryza sativa* L.)

A Thesis

by

KARINA YAZMINE MORALES

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Michael Thomson
Committee Members,	Scott Finlayson
	Lee Tarpley
Head of Department,	David Baltensperger

August 2018

Major Subject: Plant Breeding

ABSTRACT

In order to meet the needs of a growing population, rice production must be doubled by 2050 in comparison to 2005 production levels. However, rice production faces many challenges including rising temperatures, drought, and salinity stress. Flowering is one of the most heat sensitive periods of the rice life cycle, causing significant decreases to yield if temperatures are too high during the flowering process. As an agronomically important trait, days to flowering determines the climate rice varieties can grow in. Days to flowering is highly variable and quantitative in nature with over 30 genes known to control days to flowering. This study aimed to characterize genes controlling days to flowering through a genome wide association study (GWAS), analysis of genetic variation in known flowering time genes, and genome editing of two known flowering time genes. Genotyping of the material used in the GWAS was performed using a rice 7K SNP chip referred to as the Cornell-IR LD Rice Array (C7AIR).

The C7AIR successfully distinguished between the five subgroups of *Oryza sativa*; however, due to all varieties genotyped being inbred, the amount of heterozygosity was low, causing problems with identifying where the heterozygous cluster should fall when creating the custom cluster file in GenomeStudio. The GWAS identified 5 candidate loci which contribute to days to flowering, including a region co-localizing with the known flowering time genes *Hd3a* and *RFT1*. The five genes analyzed for structural variation were highly conserved among the varieties observed with no nonsense, and few missense mutations being observed. Guide RNAs were designed to knockout the function of *Hd3a* and *RFT1*; however, the ribonucleoprotein (RNP) transfection was unsuccessful due to the need for optimizing the rice protoplast isolation protocol.

DEDICATION

I'd like to dedicate this to the women in my life, especially my mom, Great Grandma Polly, Dr. Louise Huang, and Dr. Jennifer Young. Thank you for your mentorship and encouragement to persevere through challenging circumstances. I am grateful for each of your pioneering spirits and the ways you have intentionally used your skills to fight against injustices.

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Thomson, and my committee members, Dr. Finlayson and Dr. Tarpley, for their advice and support throughout this research project. I'd especially like to thank Dr. Thomson for listening to my ideas and helping me to find ways to implement them.

Thanks go out to my lab mates and employees at the AgriLife Research Center in Beaumont, Texas for supporting me with work done in the fields, especially Dr. Rodante Tabien, Chersty Harper, and Pat Carre. I'd also like to thank my friends Ammani, Nolan, and Jales for their constant encouragement and support. Finally, I'd like to thank my family for continuously motivating me to pursue my dreams.

CONTRIBUTORS AND FUNDING SOURCES

This work was supported by a thesis committee consisting of Professor Michael Thomson (Advisor) and Scott Finlayson of the Department of Soil and Crop Sciences and Professor Lee Tarpley of the Department of Molecular and Environmental Plant Sciences.

Graduate study was supported by a College of Agriculture and Life Sciences Excellence Fellowship from Texas A&M University.

The data analyzed for Chapter II from Cornell University was provided by Dr. Susan McCouch. Data analyzed in Chapter II from the International Rice Research Institute was provided by Dr. Tobias Kretzschmar. The SNP genotyping data from Texas A&M in Chapters II and III was produced by Eurofins, Inc.

All other work conducted for the thesis was completed by the student independently.

NOMENCLATURE

C6AIR	Cornell_6K_Array_Infinium_Rice
C7AIR	Cornell-IR LD Rice Array
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
DTH	Days to heading
GMO	Genetically Modified Organism
GWAS	Genome Wide Association Study
IRRI	International Rice Research Institute
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
QTL	Quantitative trait locus
RNP	Ribonucleoprotein complex
sgRNA	Single guide RNA
SNP	Single nucleotide polymorphism
tracrRNA	trans CRISPR RNA

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

I.1 Introduction

Rice is one of the most important food crops in the world with over 50% of the world depending on rice for their daily caloric intake (Muthayya et al., 2014). When comparing global rice demand in 2005 to projected demands for 2050, a 100-110% increase in production is needed to ensure food security (Tilman et al., 2011). In recent years, yield growth per year has begun to stagnate with models constructed upon the current growth rate demonstrating that yields will not rise fast enough to meet future demands (Ray et al., 2013). Although rice yields need to increase in the coming decades, the production system faces many challenges associated with climate change which are projected to cause a decrease in annual crop production (Gregory et al., 2005). Abiotic stresses expected to decrease rice yield in the future include: drought, submergence, salinity, and elevated temperature (Wassmann et al., 2009, Dixit et al., 2017). Breeding for climate resilient traits is one method of improving yields in order to ensure that future yield demands can be met. One trait thought to improve climate resilience is days to flowering, as flowering is one of the most heat sensitive times of the rice life cycle (Jagadish, 2010). Exposure to temperatures of 41 ºC and higher within one hour of flowering has been demonstrated to cause complete sterility in rice (Shah et al., 2011). Strategies for avoiding heat stress include flowering earlier in the season and shorter maturity; however, these can result in decreased yield (Zafar et al., 2017). Furthermore, elevated temperatures have been shown to decrease days to flowering in some varieties of rice (Shah et al., 2011). This places great importance on understanding the genetics behind days to flowering to allow breeding for optimal

days to flowering in elite varieties of rice. Through this project I aim to further characterize the genetic control of days to flowering in rice through a genome wide association study (GWAS), development of CRISPR/Cas9 knockout mutants, and the relationship of gene haplotypes to photoperiod sensitivity.

I.1.1 Rice Genetic Resources

The *Oryza* genus contains 27 species with *Oryza sativa* L. and *Oryza glaberrima* Steud being the only cultivated species. Although these species are comprised of the AA genome, they have significantly distinct evolutionary histories with the domestication of *O. sativa* occurring in Asia approximately 10,000 years ago and that of *O. glaberrima* occurring about 3,000 years ago in Africa (Stein et al., 2018). *O. sativa* is grown all over the world with varieties adapted to almost all continents, while *O. glaberrima* is primarily grown in Africa (Muthayya et al., 2014). In each of these species there is a large amount of phenotypic variation across a multitude of traits including: yield, height, grain color, grain shape, and days to flowering. Hundreds of thousands of rice accessions are preserved in germplasm collections; however, it is unfeasible to screen all varieties for a specific trait of interest within these collections. As a result, core collections, such as the USDA Core (Yan et al., 2007) and Mini-Core (Agrama et al., 2009) collections, are often developed. These core collections are constructed to capture as much of the genetic variation in a population as possible and have no duplicated varieties (Yan et al., 2007). Core collections serve as a valuable resource for genetic studies, giving the opportunity to screen a defined population for a multitude of traits across many environments in order to best identify the genes controlling a phenotype of interest.

The genetic resources for rice are extensive with multiple reference genomes available for the primary cultivated species, *Oryza sativa*, and its wild relatives (Stein et al., 2018). Due to its small genome and inbreeding (i.e., self-pollinating) reproduction, rice is a relatively easy crop in which to perform genetic studies as pure lines are naturally produced (McCouch et al., 2016). Along with having a fully sequenced genome there are multiple gene annotation databases for rice including the Rice Annotation Project Database (Sakai et al., 2013) and Gramene (Tello-Ruiz et al., 2016). Furthermore, a number of single nucleotide polymorphism (SNP) arrays, ranging from low to high density, have been developed for genotyping rice including the High Density Rice Array (McCouch et al., 2016) and the Cornell 6K Array Infinium Rice (C6AIR) (Thomson et al., 2017). The combination of these tools has allowed for the identification of many genotype to phenotype relationships that have been utilized in trait development for improved rice varieties.

I.1.2 Association Mapping in Rice

The combination of extensive genetic resources and diversity within the *O. sativa* genome makes this species an ideal candidate for performing genome wide association studies (GWAS) (McCouch et al., 2016). Genome wide association studies improve upon quantitative trait locus (QTL) mapping as these studies evaluate a diverse panel of individuals, which ideally do not contain significant population structure, rather than only looking at a single biparental population. This allows for a larger amount of the diversity within a species to be captured and utilized in describing the genetic control of the trait of interest (Zhao et al., 2011). In rice, GWAS studies have been used to identify genetic loci controlling hull color, drought tolerance, spikelet number, leaf angle, days to flowering, and a variety of other traits (Zhao et al., 2011,

McCouch et al., 2016, Dingkuhn et al., 2017) . Despite this rich history of GWAS in rice, there is still genetic variation that has not been captured in the panels observed. Furthermore, the environmental impact upon observed phenotypes can be difficult to determine in these populations as each diversity panel is generally observed in only one region of the world.

I.1.3 Flowering Time (Days to Flowering) in Rice

Flowering time (also known as days to flowering or heading date) is a trait in rice that is both quantitatively inherited and can be greatly dependent upon the environment that a variety is grown in. Rice is a short-day plant, meaning it will initiate flowering in response to days becoming shorter. Traditionally, many landraces would exhibit strong photoperiod sensitivity, with delayed flowering under long days until the days begin to shorten; however, as rice growth spread northward, photoperiod sensitivity was suppressed in order to allow for adaptation to these regions (Brambilla and Fornara, 2013, Hori et al., 2016). Over 30 genes are known to control days to flowering with significant amounts of variation in the genotypic response. Depending on the daylength, rice has two separate pathways to signal the occurrence of flowering with significantly more genes involved in the long day response (Hori et al., 2016).

Although much is already known about the flowering pathway in rice, this trait is important to the breeding community as flowering is one of the most heat sensitive periods in the rice life cycle (Jagadish, 2010, Zafar et al., 2017). If rice can flower earlier in the season it can potentially avoid having peak anthesis coincide with hot spells, which are more likely to occur in mid- to late-summer. Furthermore, early flowering in rice is critical in multiple cropping systems, such as the ratooning system used along the U.S. Gulf Coast, as a shorter period of days to maturity

allows for the ratoon crop to mature before cool weather in fall inhibits grain maturation, decreasing the yield of the ratoon crop (Dou et al., 2016).

I.1.4 Genome Editing using CRISPR/Cas9

Cas9 is a bacterial-derived, RNA-guided, double strand nuclease that can be utilized to mutate specific genes of interest by the CRISPR/Cas9 system (Doudna and Charpentier, 2014). This system has been shown to create targeted edits in a variety of plant species including Arabidopsis, wheat, tomato, sorghum, and rice (Belhaj et al., 2015). At this time, CRISPR/Cas9 edited plants are not being classified as genetically modified, leading to an easier path for commercialization of an edited crop (Waltz, 2016). When performing edits using CRISPR/Cas9, the only requirements are the Cas9 protein combined with either a CRISPR RNA (crRNA) and trans CRISPR RNA (tracrRNA) or a single guide RNA (sgRNA) where the crRNA and tracrRNA have been combined into one RNA (Belhaj et al., 2013). As a result, there is no need for the Cas9 and gRNA components to be stably integrated into the genome for edits to occur, thus allowing an alternative ribonucleoprotein complex (RNP) approach to be utilized (Woo et al., 2015).

I.1.5 Haplotype identification in flowering time genes

Single nucleotide polymorphisms (SNPs) can be identified through a variety of sequencing strategies including: expressed sequence tag analysis, array data, and amplicon resequencing. Amplicon resequencing currently has the lowest false discovery rate of these techniques and allows for the pooling of multiple individuals to determine allele frequency in a population (Ganal et al., 2009). In amplicon sequencing experiments primers are designed for each gene of interest and PCR is performed to obtain read lengths of 500-1000 bp (Nasu et al., 2002, Ganal et

al., 2009). The SNPs identified in these experiments can then be used in marker development for the trait of interest.

In order to characterize the days to flowering in rice and develop genetic resources for better understanding this trait and others in rice, I pursued the following objectives:

1. Development of a reference cluster file for an Illumina 7K SNP chip for rice, the Cornell-IR LD Rice Array (C7AIR), which will aid in standardizing and increasing efficiency in the use of this SNP chip for diversity analysis and association mapping.
2. Completion of a genome wide association study of a new rice diversity panel for days to flowering using field data and the 7K SNP chip.
3. Testing a CRISPR/Cas9 gene editing system to knock out flowering time related genes in rice protoplasts.
4. Characterization of photoperiod sensitivity in a subset of the rice diversity panel using controlled environments and amplicon sequencing across gene targets.

I.2 Approach and Rationale

I.2.1 Development of a reference cluster file for the Cornell-IR LD Rice Array (C7AIR) which will aid in standardizing and increasing efficiency in use of this SNP chip.

The Cornell-IR LD Rice Array is a 7K SNP chip designed on the Illumina Infinium platform that was designed by Cornell University as an improvement upon the previously published Cornell_6K_Array_Infinium_Rice (C6AIR) SNP chip (Thomson et al., 2017). The C7AIR SNP chip includes 7,183 SNPs and improves upon the C6AIR design by removing approximately 2,000 poorly performing SNPs and incorporating new SNPs from the High Density Rice Array

(McCouch et al., 2016), the 384-SNP GoldenGate sets (Thomson et al., 2012), SNPs identified as descriptive for US rice germplasm from the 44K array (Zhao et al., 2011), and SNPs from unpublished work performed at the International Rice Research Institute (IRRI). This SNP array will be informative in breeding applications and for distinguishing among the different sub-groups of rice. Furthermore, this array provides a low-cost alternative to high density arrays while still maintaining enough markers for performing genetic studies.

Although the Illumina GenomeStudio software has automated clustering and allele calling functions, it requires manual curation to correct mislabeled clusters and to improve the allele calling accuracy. A custom cluster file developed in this project could then be used as a reference file to improve the allele calling using polar coordinates for every genotype across every SNP which is then used by GenomeStudio's algorithm to sort each data point as the correct allele. We developed a cluster file for the C7AIR based upon 384 varieties genotyped by Texas A&M, 96 varieties genotyped by Cornell University, and 72 varieties genotyped by IRRI. This SNP chip was primarily designed for sequencing *Oryza sativa and Oryza rufipogon* Griff.; however, a number of *O. glaberrima*, and *Oryza nivara* S.D.Sharma & Shastry were also included in this panel. All four of these rice species are diploid and, as a result, have three possible clusters that can be observed (Stein et al., 2018). These are often labeled as AA, AB, and BB, where AA and BB are the two homozygous options for that location and AB is the heterozygous cluster. Rice is primarily a self-pollinating crop (McCouch et al., 2016). As a result, when genotyping varieties obtained from gene banks that have been through many generations of reproduction, there is a high amount of homozygosity and almost no heterozygosity for all varieties across almost all markers. This causes problems in GenomeStudio's clustering process as the algorithm expects

the presence of heterozygotes. As a result, when using the GenomeStudio's default parameters in a population with zero to low heterozygosity, GenomeStudio will incorrectly cluster one of the homozygous classes as heterozygous. As part of this project, each marker was manually clustered to sort individuals into the three possible classes. These clusters were then compared against a reference genome in order to ensure correct sorting had occurred. Completion of this project resulted in the creation of a standardized cluster file that can be used in all genotyping experiments performed with the C7AIR. Furthermore, when manually reclustering a marker with no heterozygotes it is difficult to predict where to place the outline of the heterozygous cluster in order to ensure heterozygotes will be called correctly in future panels. In order to resolve this problem, IRRI formed a panel of 24 F1 hybrid crosses that are expected to have a significantly higher number of heterozygous markers than the inbred varieties genotyped and will provide even more accurate allele calling for subsequent studies once this data is incorporated into an updated version of the cluster file.

I.2.2 Genome wide association study of a rice diversity panel for days to flowering.

A diversity panel made up of 384 varieties was grown at the Texas A&M AgriLife Research Center in Beaumont, Texas during the summer 2017 field season. This panel was constructed from a subset of the USDA core and mini-core collections and geographically diverse global accessions from the USDA National Small Grains Collection, along with additional Southern US rice varieties added in. All varieties were grown in at least one single row plot with 208 varieties grown in an additional two three-row replicates and 54 varieties coming from the replicated field trial performed as part of Dr. Rodante Tabien's experiments. Flowering notes were taken approximately once every week and were collected in the Field Book app (Rife and Poland,

2014). Days to flowering was defined as the number of days it took for the majority of plants in a replicate to reach 50% flowering, where half of the panicle was flowering for most panicles on the plant. All materials were genotyped using the 7K SNP chip (Cornell-IR LD Rice Array) described above to define the genetic relationships of the diversity panel and provide genotype data for association mapping.

Upon obtaining the genotypic and phenotypic data an initial analysis was performed in Tassel (Bradbury et al., 2007) to identify potential SNPs of interest. Population structure was then analyzed using fastSTRUCTURE (Raj et al., 2014) in order to best determine the number of groups the population can be subset in. GAPIT (Lipka et al., 2012) was used to analyze the genotype to phenotype relationship with population structure accounted for. After performing these analyses, significant loci were compared to gene annotation libraries, such as RapDB and Gramene, in order to identify genes that may be associated with these SNPs of interest.

I.2.3 Knock out of flowering time related genes in protoplasts using the CRISPR/Cas9 system.

In this experiment, I planned to use a ribonucleoprotein (RNP) approach to edit *Hd3a* and *RFT1* in protoplast cells isolated from the popular Texas rice variety 'Presidio'. *Hd3a* and *RFT1* are both known to promote flowering in rice in a photoperiod dependent manner (Hori et al., 2016). Primers were designed based on *indica* and *japonica* references to amplify these genes in Presidio background and perform Sanger sequencing on the product. CRISPRdirect (Naito et al., 2015) and CRISPR R-GEN tools (Bae et al., 2014, Park et al., 2015) were used to design gRNAs and ensure that there were no mismatches or potential off-targets. Two gRNAs were designed for each gene with both targeting the first exon. I chose to use two gRNAs as this would potentially create larger mutations if both gRNAs are functional. This also gives assurance that if one gRNA

didn't work the second may be able to create the desired mutation. The first exon was chosen as the target for a frameshift mutation to cause a premature stop codon; this would have the greatest impact on the functionality of the protein produced by leading to a truncated protein (a "knockout" mutation). Before using the Cas9 and gRNA in protoplasts I attempted to confirm their ability to efficiently target the gene of interest by combining them with our PCR product described above and running the product on a gel to confirm that the cut had been made. Protoplast isolation was performed using leaf tissue of 10 day old seedlings (Zhang et al., 2011). The Cas9-gRNA RNP complex was inserted into the cells using PEG-mediated transfection. After incubation with the RNP complex the target genes and potential off targets could then be sequenced in order to ensure that the desired mutation was created without any off-target mutations. RNP methods were used instead of traditional transformation methods as there would be no transgene integration and RNP methods have been shown to decrease the amount of off-target effects as the cell will degrade the Cas9 protein and gRNA so it is not continuously expressed (Woo et al., 2015).

I.2.4 Characterization of photoperiod sensitivity in a subset of the rice diversity panel using a controlled environment and amplicon sequencing.

The amount of days to flowering in rice is highly dependent upon environment, especially the photoperiod, or day length, the plant is exposed to. Traditionally, rice is classified as a short-day plant, meaning it will flower as the nights become longer. However, through various breeding efforts as rice has been adapted to new regions of the world, the photoperiod response of rice has evolved to create day-neutral varieties. These day-neutral varieties will flower at a specific number of degree-days after planting, rather than changing the number of days to

flowering based on daylength (Hori et al., 2016). In this experiment, 10 different rice varieties were grown in growth chambers with one of three conditions: field, short day (10 hours light plus 14 hours dark), and short day plus elevated carbon dioxide (700 ppm CO_2). I chose to include the elevated CO_2 environment as this has been shown to slightly decrease the number of days from planting to flowering (Hasegawa et al., 2016). After determining the days to flowering under each of these conditions, I planned to sequence key flowering time genes from each variety that are known to control the days to flowering. Upon obtaining this sequence data, the structure of each gene across varieties was compared to identify SNPs and haplotypes. Phenotypic data was also to be used to identify any potential SNPs or haplotypes that could be associated with photoperiod sensitivity.

I.3 Outcome

Upon completing each of these objectives, the following outcomes were expected: [1] creation of a cluster file for the 7K Cornell-IR LD Rice Array that can be used across institutions to optimize future genotyping studies, [2] identification of significant genetic loci that are correlated with days to flowering, [3] validation of protoplasts with non-functional *Hd3a* and *RFT1* genes, and [4] obtainment of sequence data for known flowering genes that explain the photoperiod response of the rice varieties screened. Each of these experiments developed new resources for rice genetics studies, specifically for describing the variation in days to flowering. Furthermore, data developed in these experiments can lead to future experiments such as validating the impact of our target genes on the days to flowering in regenerated CRISPR edited

plants, editing significant loci identified in the GWAS in order to validate their impact on days to flowering, and optimizing RNP approaches in rice to edit other genes of interest.

CHAPTER II

IMPROVEMENT AND UTILIZATION OF A RICE 7K SNP ARRAY

II.1 Synopsis

Single nucleotide polymorphisms (SNPs) are ubiquitously found in all organisms and are inexpensive to screen for. SNP arrays have been created in a variety of densities with low density arrays being cheaper to run than high density; however, when decreasing the number of SNPs the ability to distinguish between population subgroups becomes increasingly more difficult. The Cornell-IR LD Rice Array (C7AIR) is a mid-density SNP array containing 7,098 markers. This SNP chip has been used for genotyping hundreds of rice varieties and was able to successfully differentiate between the five subgroups of *Oryza sativa* while also providing valuable information for wild relatives including: *Oryza rufipogon* and *Oryza nivara*.

II.2 Introduction

Rice production needs to increase by 100-100% when comparing 2005 production to demand in 2050 in order to meet the needs of a growing population (Tilman et al., 2011). At the same time rice breeders are needing to increase resistance to a variety of abiotic and biotic stresses including: heat, drought, salinity, insect pests, and disease. In order to advance breeding at a faster rate it is beneficial to identify genetic loci linked with a trait of interest which can be incorporated into elite breeding material. High-throughput genotyping methods with low marker density offer an affordable method to generate data for the multitude of samples produced in large breeding programs around the world. Genotyping data produced from these tools can then

be applied in the selection process with key loci identified through genome wide association studies (GWAS), DNA fingerprinting, QTL mapping, and genomic selection.

Single nucleotide polymorphisms (SNPs) are often chosen as markers as they are the most common mutation in all species. SNPs are easier to utilize than insertions and deletions (INDELS) as there are only two options for each allele and there are a number of rapid, high-throughput SNP genotyping methods available. Assays based on SNPs are relatively inexpensive and can be developed at a variety of densities with examples including the RICE6K (Huihui et al., 2014) and C6AIR (Thomson et al., 2017) each having 6,000 markers, the 44 K SNP chip (Zhao et al., 2011), and the High Density Rice Array (McCouch et al., 2016) with approximately 700,000 markers. Data developed using SNP arrays offer informative markers that can differentiate groups commonly used in breeding for improved crop varieties. In the past, low-density SNP arrays, including 384 SNP bead sets, have been used for trait integration and confirmation of cultivar identity (Thomson et al., 2012). Unfortunately, identifying differences within groups often requires multiple sets of low density arrays or the use of higher density arrays, which may quickly become financially unsustainable due to their high cost. SNP arrays with six to seven thousand markers offer an affordable alternative with a high enough density to distinguish between and within breeding groups while also offering a user-friendly pipeline that can be used without being an expert in bioinformatics.

The Cornell-IR LD Rice Array (C7AIR) improves upon the Cornell_6K_Array_Infinium_Rice (C6AIR (Thomson et al., 2017)) by removing poorly performing SNPs and incorporating new SNPs from sources such as the HDRA and 384 bead sets. This SNP array offers a vast output of information and applications while its lower density

allows for inexpensive genotyping and straightforward data analysis with a low threshold of required computational resources and skills.

II.3 Materials and Methods

II.3.1 Plant Materials

The rice varieties utilized in this section are listed in Appendix A. The Texas A&M material was collected approximately 55 days after planting. Samples were then lyophilized before being sent to Eurofins, Inc. for DNA extraction and running of the samples on the 7K SNP chips. Cornell University contributed data from an additional 96 varieties and the International Rice Research Institute contributed data from 48 varieties of diverse backgrounds.

II.3.2 Design of the C7AIR

The Cornell-IR LD Rice Array (C7AIR) design was primarily based off of the Cornell_6K_Array_Infinium_Rice (C6AIR) (Thomson et al., 2017). The C7AIR improves upon the C6AIR by removing poorly performing SNPs and adding SNPs which are informative for elite tropical japonica breeding material. Both chips were designed with the following metrics in mind: no SNPs within 10 bp of each locus, no SNPs within 35 bp of the chosen locus with a minor homozygote count greater than 4, removal of all INDELS, and removal of SNPs with repetitive sequences and low minor allele frequencies. The C7AIR included 4007 SNPs from the C6AIR, 2056 SNPs from the High Density Rice Array (HDRA) (McCouch et al., 2016), 910 SNPs from the 384-SNP GoldenGate sets (Thomson et al., 2012), 189 SNPs from the 44K array which added higher information content for U.S. rice varieties (Zhao et al., 2011), and 21 SNPs based on genes from IRRI (unpublished data).

II.3.3 Genotyping and SNP Allele Calling

The manufacturer's protocol was followed for amplification of DNA before hybridizing to the Infinium II BeadChip, staining with fluorescent dye and scanning to measure the fluorescence intensity of the beadchip (processed at Eurofins BioDiagnostics, River Falls, WI). Raw intensity values were then converted to SNP data using Illumina's GenomeStudio software. The 7098 SNPs were filtered down to 6565 SNPs based on a call rate above 0.8, where at least 80% of samples were called for each SNP (less than 20% missing data per locus). The 528 varieties genotyped were then filtered down to 448 based on a call rate above 0.939 (less than 6.1% missing data per sample) and a P10 GC score above 0.45. P10 GC is a score developed by Illumina to identify samples which may have failed genotyping as described in the user manual for GenomeStudio. Upon compiling all data in GenomeStudio, SNPs were manually re-clustered in order to correctly sort the clusters as the correct genotype.

II.3.4 Tree Construction and Data Analysis

Reclustered data was exported from GenomeStudio and imported into TASSEL GUI 5.2.43 where CenteredIBS kinship using a maximum of 6 alleles was calculated (Bradbury et al., 2007). Kinship values were then imported into MEGA7 to create a phylogenetic tree (Kumar et al., 2016). TASSEL was also used to calculate linkage disequilibrium using a sliding window of 2000 alleles (Bradbury et al., 2007). Genotyping information was also input to GAPIT where marker density, VanRaeden kinship, and other linkage disequilibrium statistics were determined (Lipka et al., 2012).

II.4 Results

II.4.1 Design of the C7AIR

The C7AIR contains 7,098 SNPs which passed Illumina's production quality standards; however, after filtering this was decreased to 6,565 SNPs used in clustering. Over 50% of SNPs are less than 50 kb away from each other (Figure 1) with an average distance between neighboring SNPs of 52 kb. This subset of SNPs was further decreased to 6,132 after removing SNPs with a GenTrain score, GenomeStudio's score for determining how well a cluster pattern fits the alleles called, of less than 0.7 after manual re-clustering.

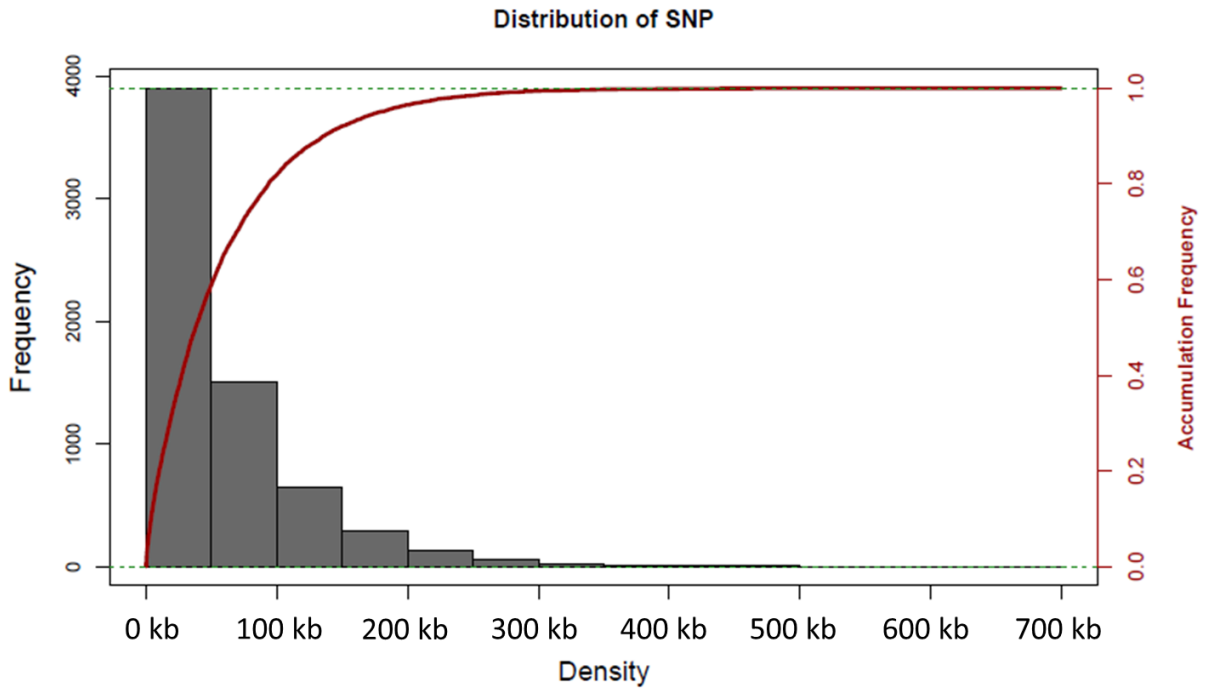


Figure 1 Distribution of gap size between a SNP and its neighbor. The majority of SNPs are within 100 kb of their neighbor.

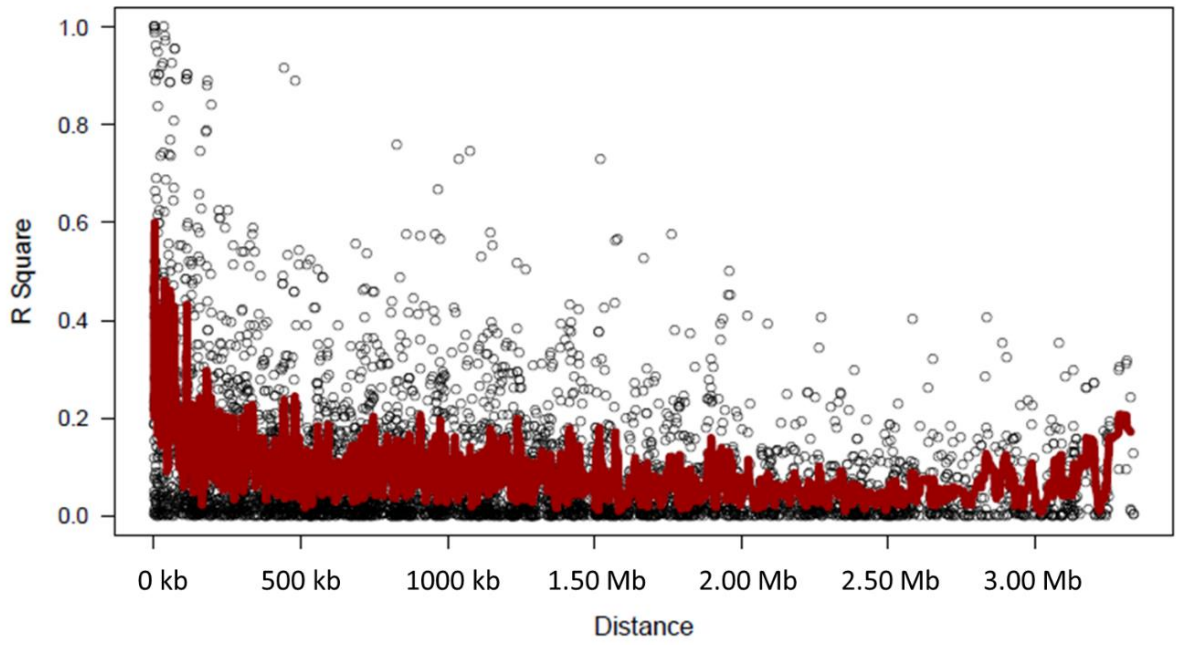


Figure 2 Linkage disequilibrium decay across all SNPs calculated using a sliding window of 10 where each dot is distance between two markers and red line is moving average of 10 adjacent markers

II.4.2 Linkage Disequilibrium

Linkage disequilibrium, the measurement of how often recombination will occur between two markers with 0 meaning the markers are in equilibrium with each other (not linked at all), and 1 meaning the two markers are completely linked to each other. The linkage disequilibrium was fairly low between all markers; however, there were many non-zero values (Figure 2). This could be caused by the co-selection of markers during the breeding process as many of the varieties genotyped are elite breeding varieties. As expected, markers physically closer to each other were more correlated with each other and had higher significance to their correlation with each other (Figure 3).

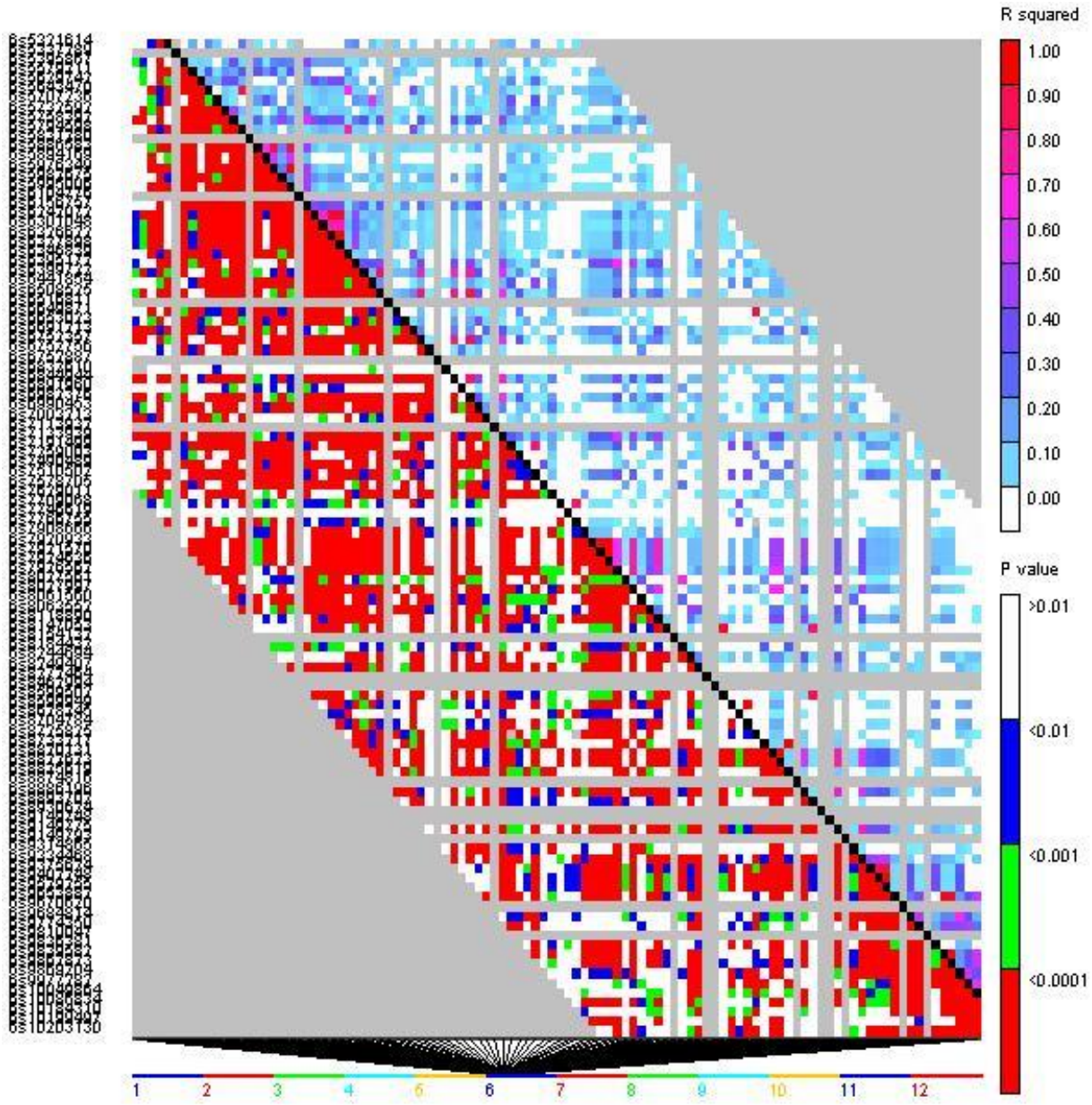


Figure 3 Linkage disequilibrium plot for a subset of SNPs on chromosome 6 calculated using a sliding window of 50 where the y axis represents a SNP and the x axis is the 50 neighboring SNPs

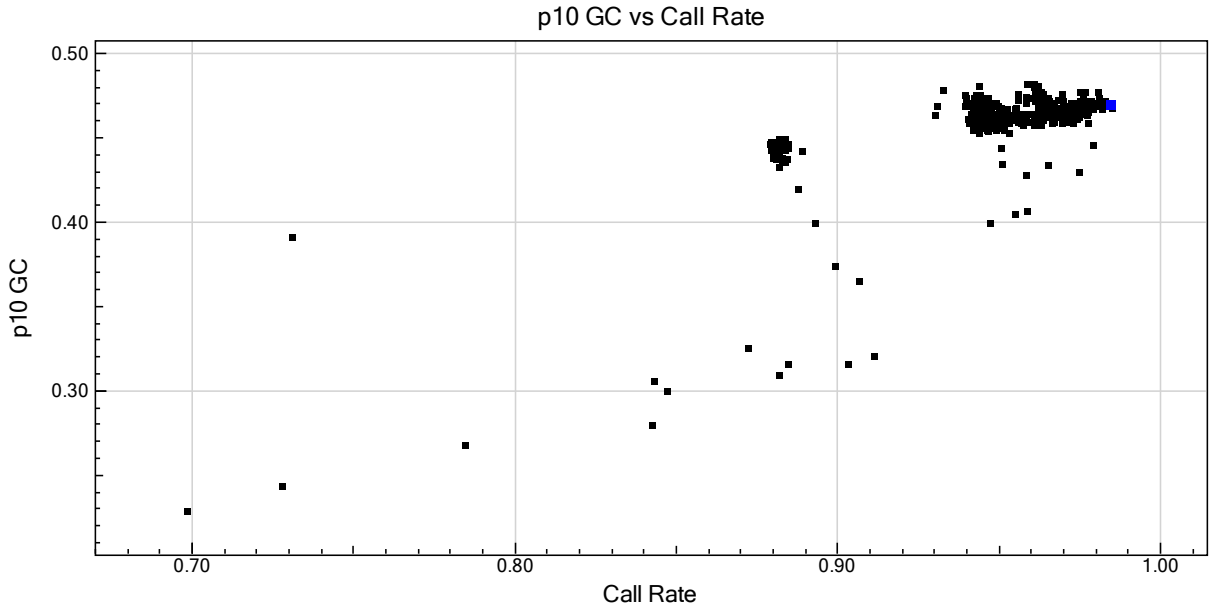


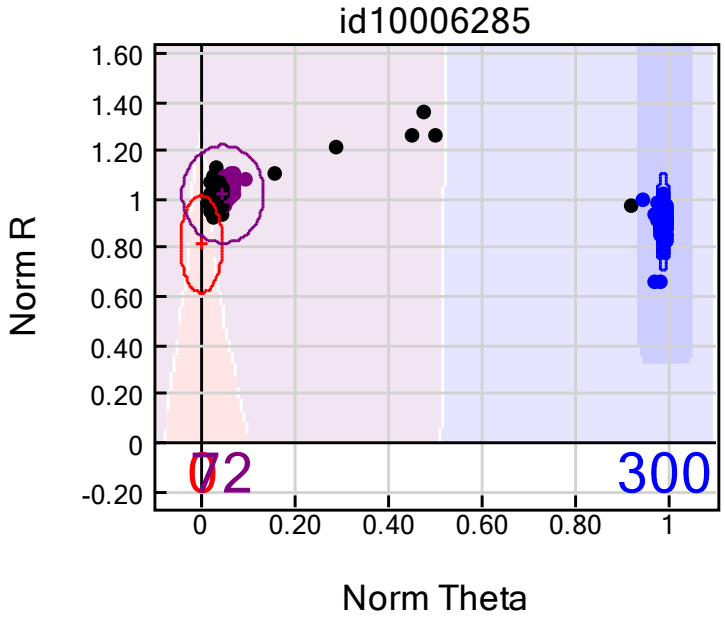
Figure 4 p10 GC versus Call Rate for all samples

II.4.3 Filtering and Manual Re-clustering of SNPs

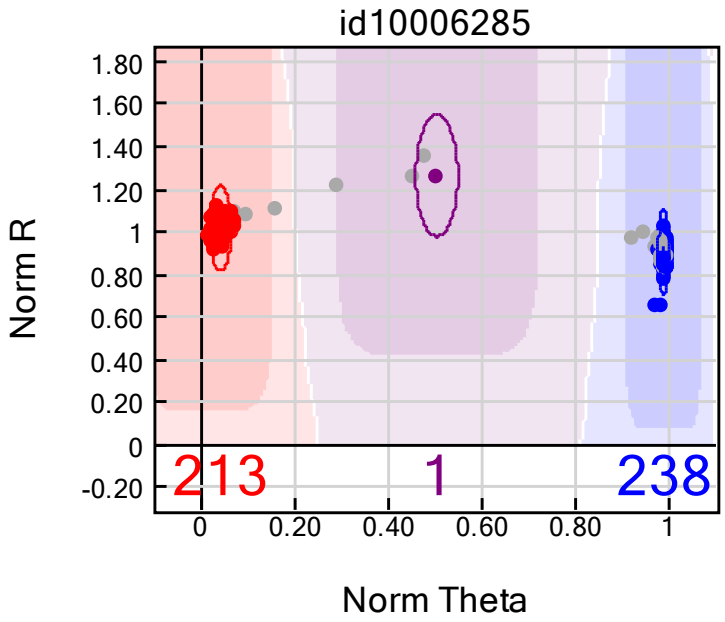
When filtering samples by p10 GC and call rate, 76 varieties were removed for having values which fell below quality standards. The cluster close to 0.88 call rate was primarily made up of *O. glaberrima* (Figure 4). The genetic architecture of *O. glaberrima* is significantly different from *O. sativa* and may require development of a separate cluster file. Although *O. glaberrima* did not meet quality standards, *O. rufipogon* and *O. nivara* both clustered with most *O. sativa* above 0.94 call rate and 0.45 p10 GC.

The majority of varieties genotyped were inbred, leading to high levels of homozygosity across all markers; however, when clustering was performed based on GenomeStudio's clustering algorithm the percentage of heterozygous calls was much higher than expected from the inbred populations genotyped (Figure 5a). After filtering and re-clustering, the number of

heterozygous calls significantly decreased. This created difficulties in placing the heterozygous clusters as there were oftentimes only two or three samples to base the cluster on.



(a)



(b)

Figure 5 Uncorrected SNP cluster (a) compared to corrected SNP cluster (b)

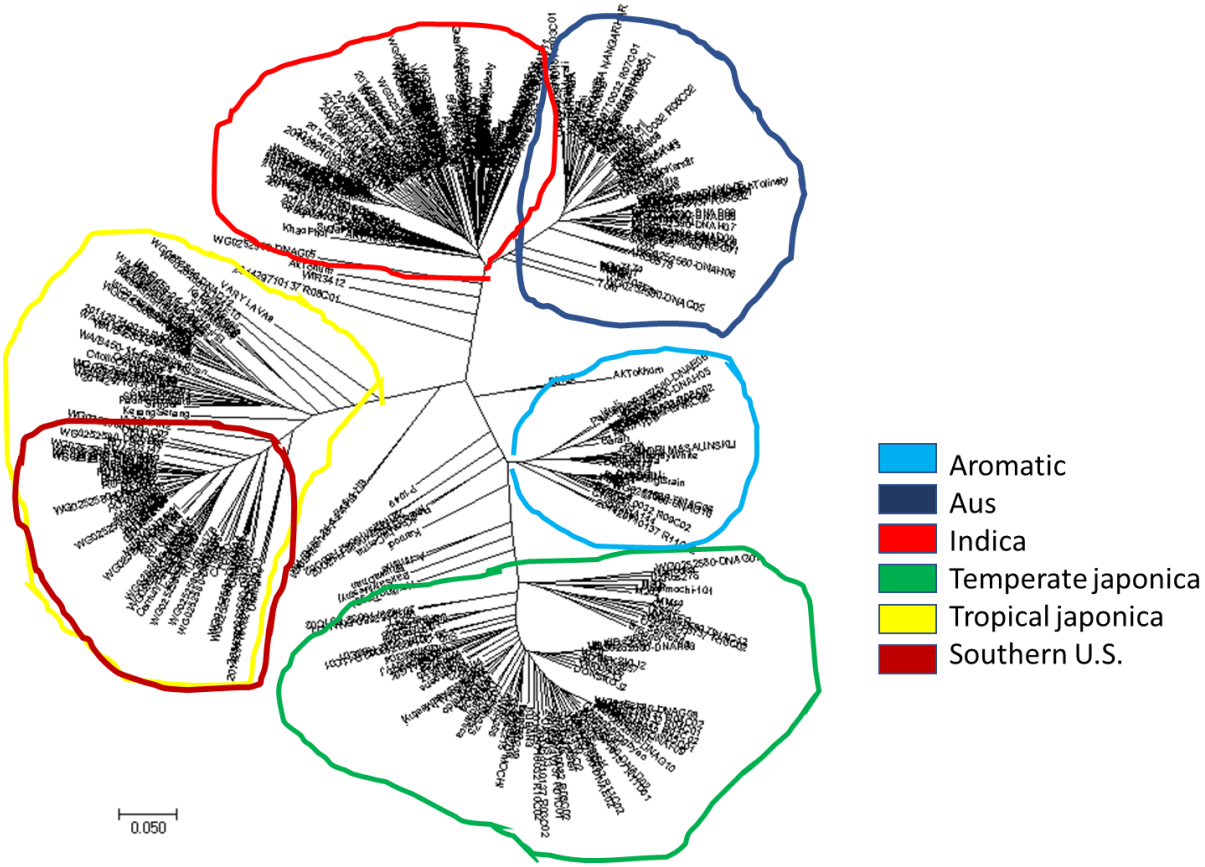


Figure 6 Phylogenetic tree displaying subgroups of *O. sativa*

II.4.4 Diversity Analysis

The C7AIR successfully differentiated between the 5 major subgroups of *O. sativa* using Centered IBS (Figure 6) and VanRaeden kinship measurements (Figure 7). Distinguishing between varieties was much more difficult in the *indica* and *aus* subgroups compared to the *tropical japonica, temperate japonica,* and *aromatic* subgroups. The Southern U.S. varieties appropriately clustered with the *tropical japonica* as expected with their breeding history.

II.4.5 Implementation of the C7AIR

The C7AIR is currently being utilized at the International Rice Research Institute (IRRI), Cornell University, Texas A&M University, and Louisiana State University. IRRI's primary use of this tool is for genotyping *indica* varieties used in their breeding programs. Cornell utilizes the C7AIR for identification of quantitative trait loci (QTL) in crosses made between *O. sativa* and its wild relatives *O. rufipogon* and *Oryza meridonalis* Ng. Texas A&M's use of the C7AIR is to distinguish between the 5 subgroups of *O. sativa* and *O. glaberrima* in a diversity panel comprised of varieties in the USDA core and mini-core collections. Cornell, Texas A&M, and Louisiana State University all utilize genotyping data obtained from the C7AIR in characterizing elite U.S. breeding material.

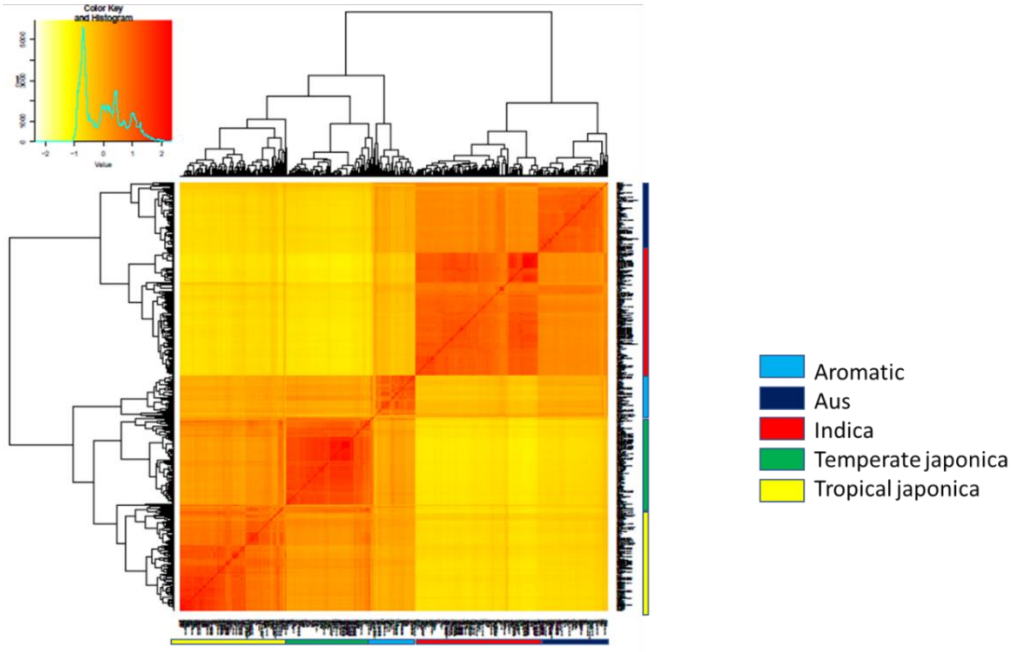


Figure 7 VanRaeden kinship heat map, displaying relatedness of individuals to each other with the legend identifying subgroups of *O. sativa*

II.5 Discussion

The C7AIR successfully differentiates between the 5 subgroups of *O. sativa* and is a useful tool for DNA fingerprinting. This SNP array has been successfully used to genotype hundreds of samples between the four institutions currently using it. However, there is currently no standardized pipeline among users of the C7AIR. Development of a cluster file allows for quick, automated genotyping while also creating reproducible results across multiple research programs. When analyzing raw genotypic data, filtering, sorting through, and re-clustering problematic SNPs can take hundreds of hours. If this same process is performed utilizing an optimized cluster file the processing time can significantly decrease to approximately 20 minutes. Availability of a cluster file will therefore allow processing of larger numbers of samples while also increasing certainty in the genotypic calls produced.

Although many of the SNPs have passed initial quality control, the number of heterozygous calls for each locus were fairly low. This causes challenges when developing an automated cluster file because placing the cluster in the wrong location can result in the incorrect sorting of samples for that specific locus. The number of true heterozygous calls can be increased by including more F1 samples from crosses of diverse accessions. By adding more heterozygous calls there will be more certainty on what allele combinations are present within a population as it will be less likely to exclude good quality markers while also increasing the ability to remove low quality markers within each sample.

Low density SNP chips, such as the C7AIR, allow quick, efficient fingerprinting of hundreds of varieties at low cost. In combination with the ability to process data with minimal resources, skills, and time this tool allows breeders to genotype hundreds of varieties.

Furthermore, the C7AIR has been shown to be effective at differentiating between subgroups of rice and clearly displaying relatedness between varieties. These characteristics allow the C7AIR to serve as a valuable resource for gene banks, giving assurance and confidence in the identity of the many varieties stored and distributed.

CHAPTER III

GENOME WIDE ASSOCIATION STUDY OF DAYS TO FLOWERING IN RICE

III.1 Synopsis

Days to heading is a quantitative trait in rice that is an important factor in determining the adaptation of varieties for specific environments. This experiment performed a genome wide association study on a subset of the USDA rice core and mini-core collections, along with geographically-diverse global rice accessions from the USDA National Small Grains Collection, to identify quantitative trait loci associated with days to heading. All materials were genotyped with a 7K SNP chip (the Cornell-IR LD Rice Array). GWAS models using mixed linear models and enhanced compression mixed linear models were used with K values of 2 and 5 (representing the number of subpopulations). The enhanced compression mixed linear model identified the most significant loci while a K value of 2 was most informative. All models were similar in their ability to attribute variation to genotypic differences. Five significant loci were identified through all models (two on chromosome 1, and one each on chromosomes 2, 5, and 6). The SNP on chromosome 6 co-localized with known flowering time genes *RFT1* and *Hd3a* while other SNPs co-localized with annotated genes that had predicted functions of binding, nuclear transport, and leaf senescence.

III.2 Introduction

Rice is one of the most important cereal crops in the world. As the global population increases, there is a need to double the current rice production to meet worldwide rice demands (Ray et al., 2013). However, as demand for rice production increases, there are also many

challenges to boosting yields including: rising temperatures, increased frequency of catastrophic storms, and salinity stress caused by rising sea levels (Wassmann et al., 2009). Days to heading (DTH) is important in determining what environments a variety may be suited for, as influenced by many factors such as photoperiod, water availability, and temperature (Hori et al., 2016). The timing of flowering can be vital in allowing avoidance of stressful conditions while also improving seed and biomass yields in crops (Jung and Müller, 2009).

Traditionally rice is described as a short-day crop, meaning it will flower sooner as days become shorter (Vergara and Chang, 1985). However, the degree of photoperiod sensitivity present in a variety can differ widely with some being completely photoperiod dependent and others being photoperiod insensitive (Izawa, 2007). Flowering occurs through one of two independent pathways. The first, which is triggered in short day conditions, is evolutionarily similar to the flowering pathway in *Arabidopsis* and contains *OsGI, Hd1,* and *Hd3a*. The second pathway occurs under long day conditions and is promoted by *DTH2* and *Ehd1,* while *Hd1* and *OsPRR37* repress the flowering pathway. Over 20 genes are known to contribute to one or both of these pathways, causing DTH to be a highly quantitative trait (Hori et al., 2016).

Genome wide association studies (GWAS) aim to take advantage of linkage disequilibrium in a panel of individuals that capture the diversity of a population in order to correlate regions of the genome to a specific phenotype (Myles et al., 2009). The average linkage disequilibrium, or correlation between two loci close to each other on the same chromosome, in cultivated rice generally ranges from approximately 100 kb to 500 kb (Garris et al., 2005, Tung et al., 2010). GWAS studies utilize diversity panels in order to identify alleles controlling a trait of interest that are in disequilibrium with a SNP or marker identified in the genotypic data (Myles et al., 2009). The USDA has created the core and mini-core collections of rice germplasm with the aim

of grouping the most diverse rice varieties in one collection for use in association mapping studies (Yan et al., 2007, Agrama et al., 2009). The Cornell-IR LD Rice Array (C7AIR) contains 6,565 high quality markers with an average coverage of one marker per 52 kb. This genotyping tool was used in tandem with a subset of the USDA core and mini core collections in order to identify quantitative trait loci associated with days to heading in rice.

III.3 Materials and Methods

III.3.1 Plant Materials

The varieties chosen for this study were primarily selected from the USDA core (Yan et al., 2007) and mini core collections (Agrama et al., 2009) with other diverse global rice varieties added from the USDA National Small Grains Collection. Elite Texas breeding varieties were also added in to this study's collection as well. Varieties used in this experiment are listed in Appendix A. All material was planted in the field at the Texas A&M AgriLife Research Center in Beaumont, Texas on April 10th, 2017 in two three-row replicates. Standard agronomic practices for Texas were used in growing these varieties.

III.3.2 Phenotyping

Flowering notes were taken approximately once a week beginning about 2 months after planting. All notes were collected using the Field book app (Rife and Poland, 2014). Days to flowering was defined as the number of days it took for the majority of plants in a replicate to reach 50% flowering, where half of the panicle was flowering for most panicles on the plant.

III.3.3 Genotyping

Genotyping data used in this experiment was obtained using the C7AIR protocols described in Chapter 2. Varieties with p10 GC below 0.45 and call rate below 0.939 were removed from downstream analyses.

III.3.4 Data Analysis

Population structure was determined using fastStructure (Raj et al., 2014) which returned an optimal K (number of groups) of 2; however, due to our observation of 5 subgroups in the phylogenetic tree (Figure 6) we ran GWAS models based on both k values. K=2 is preferred as it represents the ancestral split of the population into the two major *Oryza sativa* subgroups *japonica* and *indica.* GAPIT was used to identify significant loci that correlated with days to heading using mixed linear models (MLM) and enhanced compression MLM (CMLM). Markers and individuals with more than 10% heterozygosity were removed prior to performing the GWAS analysis as the inbred nature of our population led us to expect a fairly low amount of heterozygosity. RAP-DB was used to search 100 kb up and downstream of significant SNPs to identify genes which may correlate with the trait of interest (Kawahara et al., 2013, Sakai et al., 2013). Annotation of these regions was further confirmed using the Rice Genome Annotation Project and the Oryza Genome Evolution set of annotations from Gramene (Ouyang et al., 2006, Tello-Ruiz et al., 2016).

III.4 Results

III.4.1 Flowering Time Distribution

Days to flowering ranged from 60 days after planting to 190 days after planting (DAP). The majority of varieties flowered between 90 and 120 DAP. Almost all varieties flowered before 150 DAP with a few extremely late varieties flowering after this (Figure 8).

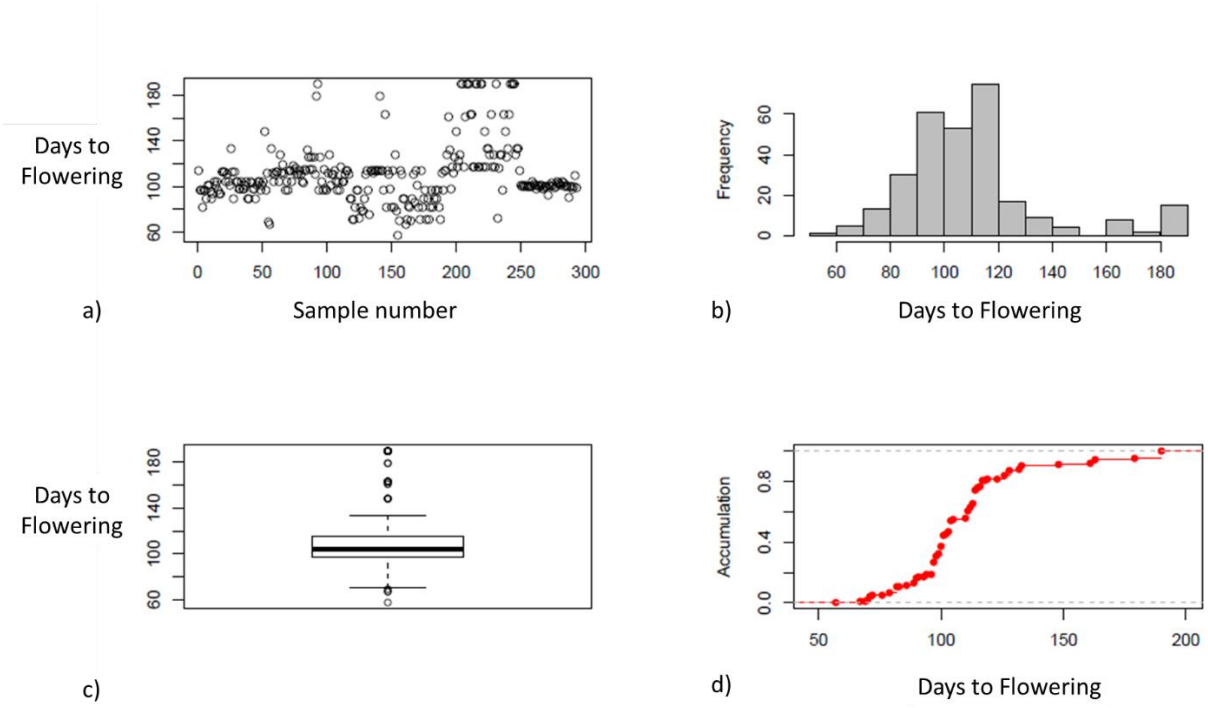


Figure 8 Distribution of days to flowering: (a) scatter plot of days to flowering, (b) histogram of days to flowering, (c) box plot of days to flowering, (d) accumulative distribution days to flowering

III.4.2 Genome Wide Association Prediction Models

Filtering markers and genotypes to remove potential outliers with high percentages of heterozygosity slightly improved the model as demonstrated through the fit of the QQ-plots. The QQ-plots show the fit of the observed p-value for a SNP compared against the expected p-value

with the expectation that there will be a 1:1 fit as shown through the red line with the grey window giving the deviation a sample can show before being considered significant (Figure 9b and 10b). In all models at least 80% of the variance was explained by genetics. The optimal compression for all model predictions clustered based on mean and grouped based on average. SNPs 4824 and 4806 were consistently among the top 3 highest impact SNPs with a significance of FDR-adjusted p-value < 0.05 for both default setting models, and the MLM and enhanced compression MLM based on k=2 (Figure 9-14, Table 1-6). The enhanced compression MLM with k=2 returned the highest number of significant SNPs with five SNPs having an FDR-adjusted p-value less than 0.05 (Figure 13, Table 5). In most cases accounting for kinship did not change the identity of the highest impact SNPs; however, it did change the level of significance for specific SNPs. The K value of 2 subpopulations resulted in higher significance of SNPs than a K value of 5 subpopulations. The enhanced compression MLM differed from the MLM by reducing the count of samples that are the same to increase variance in order to improve ability to detect true quantitative trait loci. This was demonstrated through the enhanced compression MLM with K=2 outputting 5 significant loci while the MLM with K=2 only showed 2 significant loci. Increasing the variance allows the significance to increase as this makes it easier to differentiate between potential controls for a highly quantitative trait.

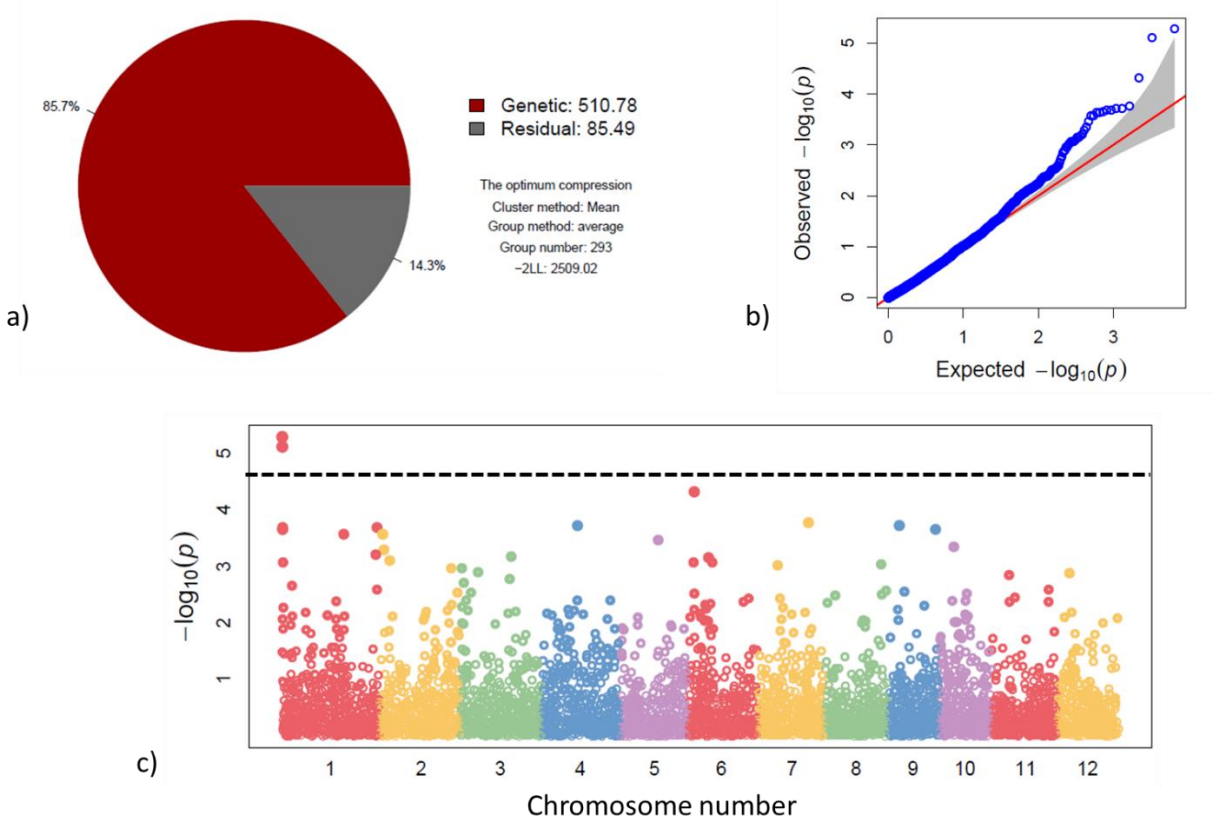


Figure 9 Model prediction based on default GAPIT settings prior to filtering. (a) Optimal model, (b) QQ-plot where red line shows 1:1 correlation and shaded area shows confidence of fit, (c) Manhattan plot displaying significance of SNPs

Table 1 Five highest impact SNPs using default GAPIT settings prior to filtering where SNP is SNP name, Chr is chromosome, Pos is position, and maf is minor allele frequency

SNP	Chr	Pos	P.value	maf	FDR_Adjusted_P-values	Allelic Effect
4824	1	194844	5.21E-06	0.399317	0.025651	-11.1014
4806	1	192836	7.81E-06	0.404437	0.025651	11.08765
id6002535	6	3154730	4.82E-05	0.479522	0.105441	11.31496817
7787812	7	22609202	0.000171	0.104096	0.136828	12.09869202
4226233	4	15845898	0.00019	0.307167	0.136828	11.67450173

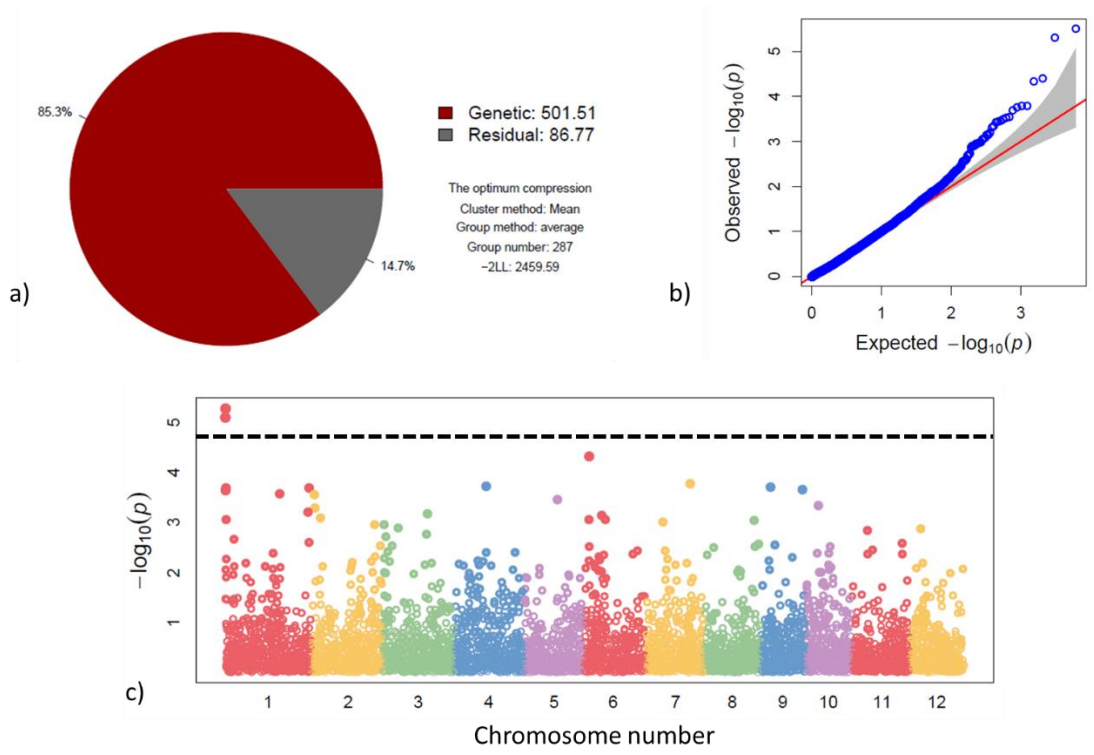


Figure 10 Model prediction based on default GAPIT settings after filtering. (a) Optimal model, (b) QQ-plot where red line shows 1:1 correlation and shaded area shows confidence of fit, (c) Manhattan plot displaying significance of SNPs

Table 2 Five highest impact SNPs using default GAPIT settings after filtering where SNP is SNP name, Chr is chromosome, Pos is position, and maf is minor allele frequency

SNP	Chr	Pos	P.value	maf	FDR_Adjusted P-values	Allelic Effect
4824	1	194844	3.11E-06	0.39547	0.015116	-11.5764
4806	1	192836	4.91E-06	0.400697	0.015116	11.54673
7787812	7	22609202	4.01E-05	0.102787	0.072564	14.27916
id6002535	6	3154730	4.72E-05	0.472125	0.072564	11.44802
SNP-1.303375.	1	304376	0.000163	0.418118	0.152546	-8.98933

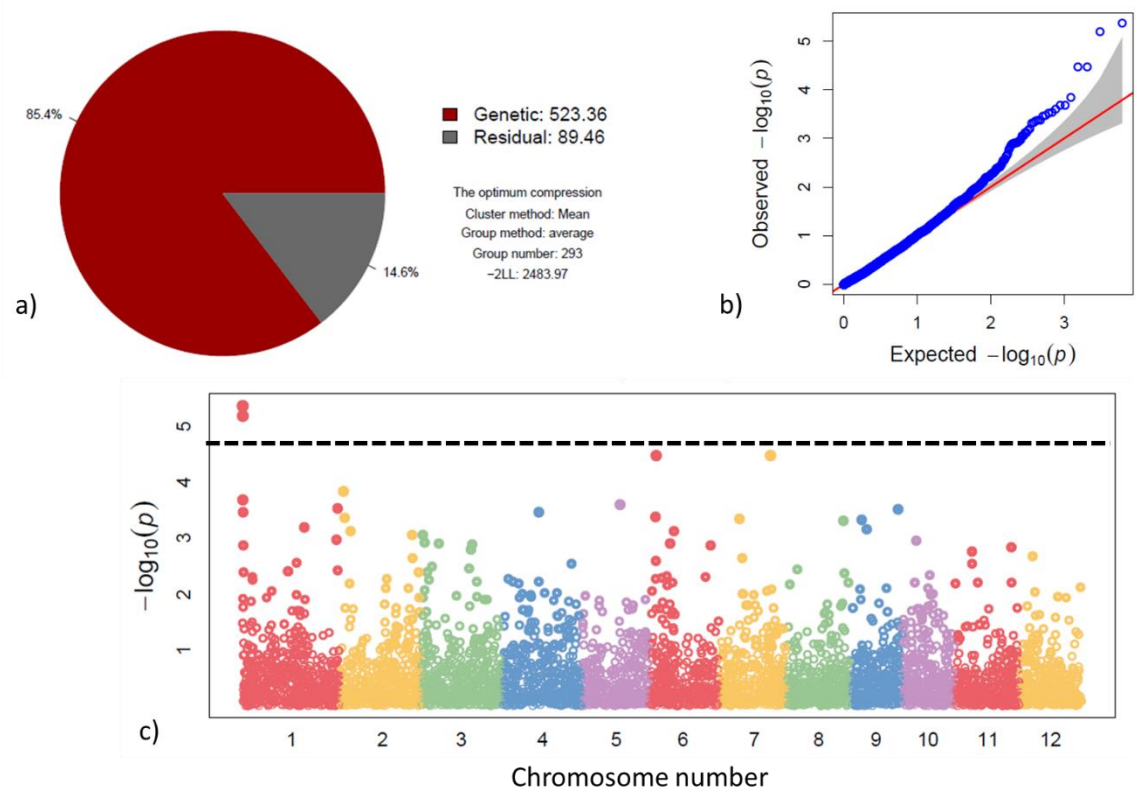


Figure 11 Model prediction using MLM with k=2. (a) Optimal model, (b) QQ-plot where red line shows 1:1 correlation and shaded area shows confidence of fit, (c) Manhattan plot displaying significance of SNPs

Table 3 Five highest impact SNPs using MLM with k=2 where SNP is SNP name, Chr is chromosome, Pos is position, and maf is minor allele frequency

SNP	Chr	Pos	P.value	maf	FDR_Adjusted_ P-values	Allelic Effect
4824	1	194844	7.23E-06	0.399317	0.036809	-11.004
4806	1	192836	1.12E-05	0.404437	0.036809	10.95741
id6002535	6	3154730	4.83E-05	0.479522	0.105721	11.47523
7787812	7	22609202	0.000168	0.104096	0.156476	11.99424
4226233	4	15845898	0.000182	0.307167	0.156476	11.69138

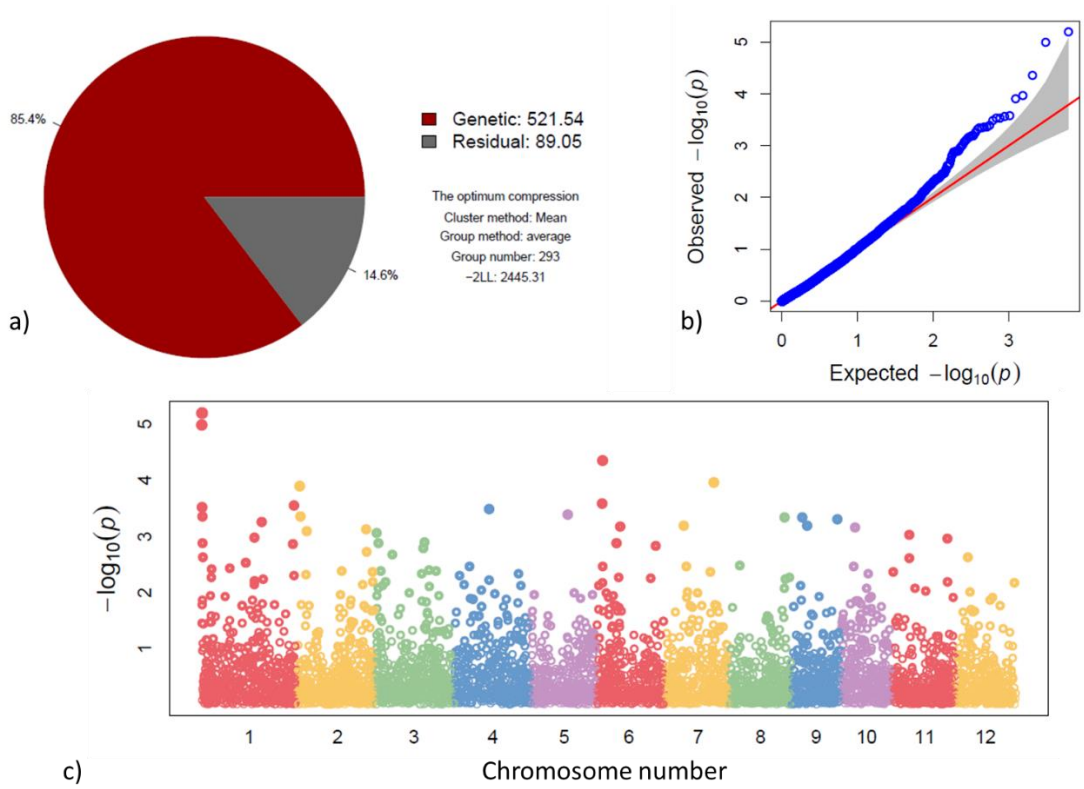


Figure 12 Model prediction using MLM with k=5. (a) Optimal model, (b) QQ-plot where red line shows 1:1 correlation and shaded area shows confidence of fit, (c) Manhattan plot displaying significance of SNPs

Table 4 Five highest impact SNPs using MLM and k=5 where SNP is SNP name, Chr is chromosome, Pos is position, and maf is minor allele frequency

SNP	Chr	Pos	P.value	maf	FDR_Adjusted_ P-values	Allelic Effect
4824	1	194844	1.05E-05	0.399317	0.056798	-10.8404
4806	1	192836	1.73E-05	0.404437	0.056798	10.76538
id6002535	6	3154730	6.03E-05	0.479522	0.131999	11.28087
4226233	4	15845898	0.000169	0.307167	0.191645	11.84797
9250909	9	5008706	0.0002	0.075085	0.191645	11.8298

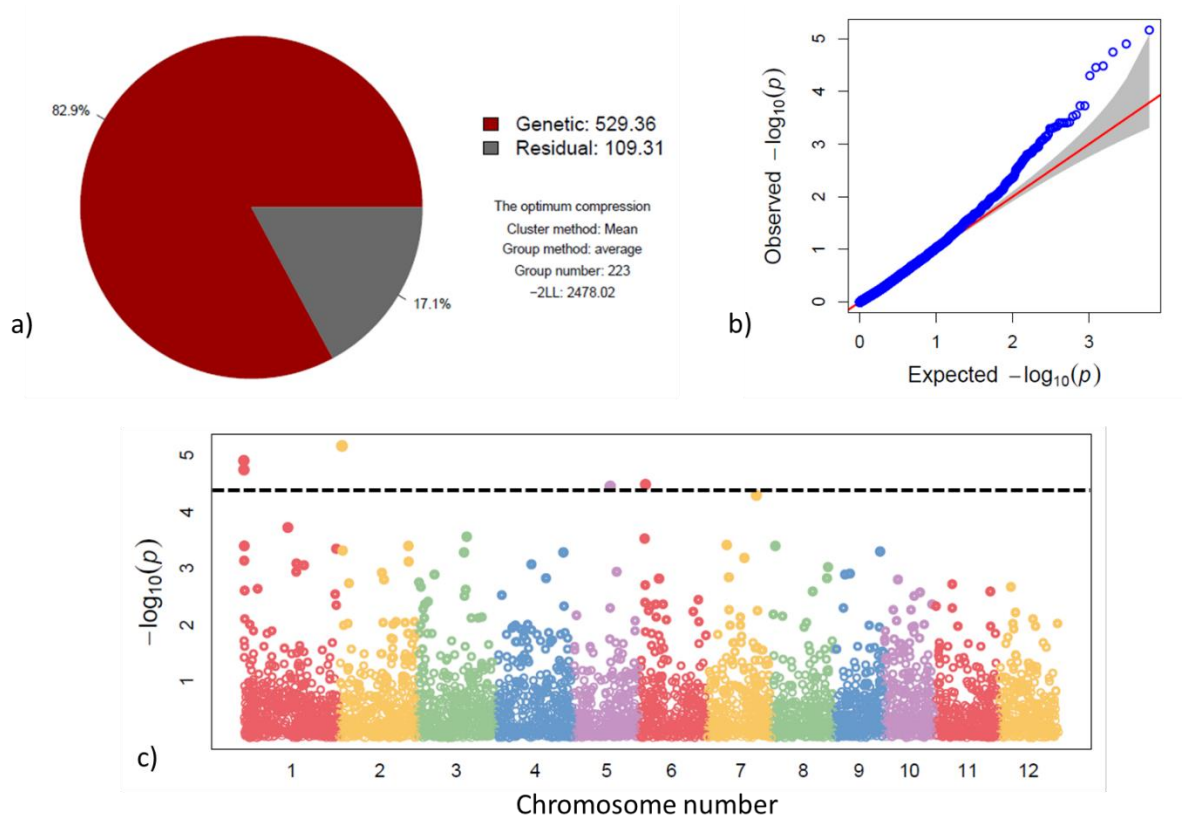


Figure 13 Model prediction using enhanced compression MLM with k=2. (a) Optimal model, (b) QQ-plot where red line shows 1:1 correlation and shaded area shows confidence of fit, (c) Manhattan plot displaying significance of SNPs

Table 5 Five highest impact SNPs using enhanced compression MLM with k=2 where SNP is SNP name, Chr is chromosome, Pos is position, and maf is minor allele frequency

SNP	Chr	Pos	P.value	maf	FDR_Adjusted_ P-values	Allelic Effect
1407860	2	1660713	6.81E-06	0.188153	0.036815	-13.2955
4824	1	194844	1.25E-05	0.39547	0.036815	-10.4012
4806	1	192836	1.79E-05	0.400697	0.036815	10.5516
id6002535	6	3154730	3.32E-05	0.472125	0.043284	11.77453
5423023	5	16808642	3.52E-05	0.339721	0.043284	-10.8914

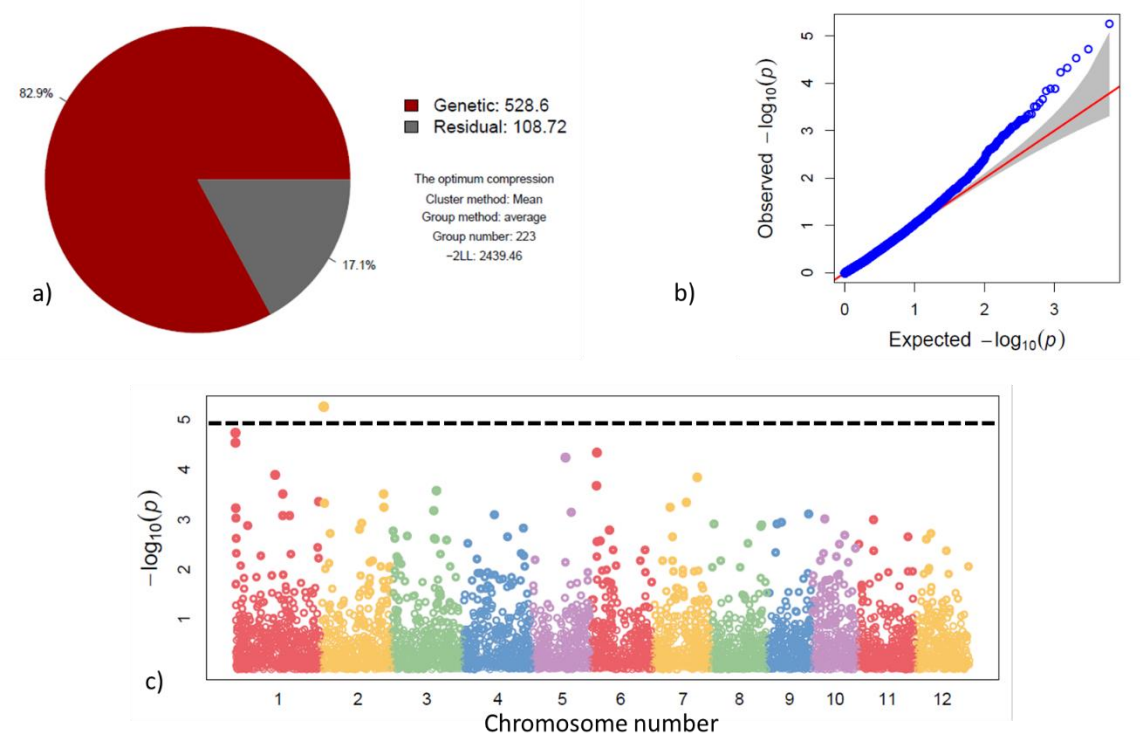


Figure 14 Model prediction using enhanced compression MLM with k=5. (a) Optimal model, (b) QQ-plot where red line shows 1:1 correlation and shaded area shows confidence of fit, (c) Manhattan plot displaying significance of SNPs

Table 6 Five highest impact SNPs using enhanced compression MLM with k=5 where SNP is SNP name, Chr is chromosome, Pos is position, and maf is minor allele frequency

SNP	Chr	Pos	P.value	maf	FDR_Adjusted_ P-values	Allelic Effect
1407860	2	1660713	5.54E-06	0.188153	0.034082	-13.5194
4824	1	194844	1.89E-05	0.39547	0.058179	-10.2222
4806	1	192836	2.92E-05	0.400697	0.059819	10.32737
id6002535	6	3154730	4.72E-05	0.472125	0.071283	11.49664
5423023	5	16808642	5.79E-05	0.339721	0.071283	-10.601

Table 7 Genes associated with SNPs identified as significant and their function

SNP	Chr	Pos	Gene	Function
4824	1	194844	Os01t0103700	Binding, reproduction
4806	1	192836	Os01t0103700	Binding, reproduction
id6002535	6	3154730	RFT1, Hd3a	Flowering Time Genes
1407860	2	1660713	OSJAP02G02560	Nuclear Transport Factor
5423023	5	16808642	OSJAP05G14860	Leaf Senescence Protein

III.5 Discussion

During the analysis of the genotype by phenotype correlation, the models utilizing K=2 outperformed those using K=5. This suggests the major subgroups of rice, *japonica* and *indica*, better describe the population structure than the minor subgroups: *tropical japonica, temperate japonica, aus, indica,* and *aromatic*. In all models, genetics was the major descriptor of the variance with at least 80% of the phenotypic variation described by how related different varieties are to each other. The enhanced compression model using K=2 found the highest number of significant SNPs, as expected, as this clustered individuals that were considered the same in order to increase the variation. As all models were fairly similar when comparing the genetic variance explained, the enhanced compression MLM with K=2 was identified as the optimal model due to its ability to identify true SNPs of significance and the fit of the QQ-plot observed.

Moving forward, all SNPs identified as significant in any of the six models were compared against 100 kb upstream and downstream of the locus in the *japonica* and *indica* reference genomes. The majority of genes found within 100 kb of significant SNPs of interest are not traditionally described as flowering time related genes. However, SNP id6002535, which was found to be significant in the enhanced compression MLM with k=2, co-localized with two

genes previously described as related to DTH. *RFT1* and *Hd3a* are both mobile signaling molecules which promote flowering with double RNAi of these genes preventing flowering even after 300 days from planting (Komiya et al., 2008). The two SNPs on chromosome 1 are close enough to suggest they are the same locus as they are only separated by 2 kb. For the significant locus on chromosome 2, the nearest gene *OSJAP02G02560* is simply described as a coding protein on RAP-DB and Gramene. However, the Rice Genome Annotation Project describes this as a nuclear transport factor and Gramene shows this is an orthologue to the Arabidopsis version of *Nuclear Transport Factor 2 (NTF2)* (Zhao et al., 2006). Finally, for the significant locus on chromosome 5, the nearest relevant gene *OSJAP05G14860* is described as a promoter of leaf senescence. Leaf senescence proteins have previously been described as regulators of and/or correlated with the occurrence of heading (El Mannai et al., 2017). This provides some support of the relevance of this candidate gene, as flowering begins the reallocation of nutrients from leaves into the grain, associated with enhanced leaf senescence under some conditions. However, the ultimate function of all underlying candidate genes would need further confirmation by subsequent studies.

Oftentimes GWAS experiments are performed using genotyping by sequencing or microarrays with tens of thousands of markers with high marker density. The C7AIR in comparison has a marker density of one marker per 52 kb which is much lower than arrays generally used. Although the C7AIR had relatively low density in comparison to SNP arrays normally used in GWAS, it was able to successfully identify loci that correlate with known flowering time genes. Through this study novel regions of interest were also identified as correlating with days to heading. In order to validate the effect of significant loci, future plans are to use the CRISPR/Cas9 system of genome editing to knockout genes of interest in the Texas

variety Presidio. In addition, a second-year field trial is currently being grown in Beaumont, Texas in order to ensure the effects described arose from genotypic differences rather than environmental effects.

CHAPTER IV

CHARACTERIZATION OF SEQUENCE VARIANTS ACROSS DIVERSE ACCESSIONS OF EARLY FLOWERING RICE

IV.1 Synopsis

Days to heading is a complicated trait with dozens of genes controlling the response of a plant to environmental cues. Phenotypes can range from completely photoperiod dependent, where a plant will flower after exposure to a specific daylength, to fully independent, where the plant will flower after a certain number of days no matter the daylength. This study observed 8 early flowering varieties under field, short-day, and short-day plus elevated carbon dioxide environments to characterize differences in response to environmental conditions. Five known flowering time genes were then characterized using Sanger sequencing. These genes were highly conserved among these varieties with only five missense mutations being observed. Days to flowering did not differ between field and short-day conditions; however, the response to elevated carbon dioxide was highly variable with some varieties showing decreased days to flowering and others increased days to flowering.

IV.2 Introduction

Days to flowering is a complex trait in rice with over 30 known genes contributing to the observed phenotype for this trait (Hori et al., 2016). Days to flowering can often control the areas where a variety is environmentally suited. Traditionally rice has been described as a short-day plant, meaning it will flower as the days become shorter (Vergara and Chang, 1985). However, there are varieties which are photoperiod insensitive, meaning they will flower after a certain

number of days, not dependent upon the daylength. Furthermore, the number of days to flowering can be highly variable with days to flowering ranging from extremely early (50-60 days after planting (DAP)) to extremely late or never flowering varieties under specific environmental conditions. Despite the quantitative nature of this trait, days to flowering is an important agronomic trait as flowering is one of the most heat sensitive periods in the rice life cycle (Jagadish, 2010) and time to flowering can control the latitude and climate that different varieties can grow in (Koo et al., 2013).

Depending on the daylength a rice plant is grown in, genes involved in the flowering pathway can serve different functions. For example, *Hd1* promotes days to flowering under short day conditions, causing flowering to occur sooner when this gene is expressed; however, under long day conditions this promoter becomes a repressor, preventing expression of *Hd3a* and *RFT1* (Hori et al., 2016). *Hd1* is evolutionarily similar to the Arabidopsis gene *CONSTANS* with both having nuclear localization signals and zinc finger binding sites and regulating the photoperiod response of their respective species (Yano et al., 2000). Both *CONSTANS* and *Hd1* expression is regulated through *GIGANTEA* (*OsGI* in rice), a circadian clock gene (Park et al., 1999, Hori et al., 2016). *Hd3a* is one of the final regulators of flowering, with its expression promoting flowering under both short and long day conditions. Under short day conditions, *Hd3a* expression is promoted by *Hd1* while under long days *Hd1* represses the expression of *Hd3a* (Kojima et al., 2002, Hori et al., 2016). Beyond *Hd1* expression coming under the control of *OsGI*, it is also controlled by *Hd6* which acts independently from the circadian clock (Ogiso et al., 2010). Genes related to days to flowering can also have pleiotropic effects, as demonstrated by *Ghd7* which simultaneously contributes to yield-related traits and days to flowering (Xue et al., 2008). Natural variation exists within each of these, leading to diversity in function and

expression. This experiment aims to characterize genetic differences in these five genes and how they may impact the days to flowering in rice in short day, short day plus elevated carbon dioxide, and field conditions.

IV.3 Materials and methods

IV.3.1 Plant Materials

Eight early flowering rice varieties were planted in the field (Texas A&M AgriLife Center in Beaumont, TX) and two controlled environment conditions (growth chambers) to determine variation in days to flowering. Varieties tested included: Antonio, Arpa Shali, Cocodrie, Colorado, Kubanets, N22, Nahodka, and Presidio. Planting in the field occurred on April 10th, 2017; crop management used standard agronomic practices. Daylength during the growth of these plants ranged from 12 hours and 46 minutes to 14 hours and 5 minutes with a mean daylength of 13 hours and 35 minutes. Both controlled environments had a short-day cycle with 10 hours of light and 14 hours dark. The day temperature was set at 30 ºC while the night temperature was set at 21 ºC. Plants grown in the first chamber were grown with ambient atmospheric conditions while plants grown in the second chamber were grown with an elevated carbon dioxide (eCO_2) condition of 700 ppm. Spectral analysis comparing the two chambers is included in Appendix B. Days to flowering was defined as the number of days it took for the majority of plants in a replicate to reach 50% flowering, where half of the panicle was flowering for most panicles on the plant.

IV.3.2 Primer Design and Amplicon Sequencing

Five known flowering time genes were chosen to sequence through Sanger sequencing. These genes were *OsGI, Hd1, Ghd7, Hd6,* and *Hd3a*. Primers were designed to capture all

exons and the first 500 bp before the first exon in 1 kb segments using Primer3 (Koressaar and Remm, 2007, Untergasser et al., 2012). Due to the length of the sequences we aimed to create, intron regions were included for the majority of sequences; however, not all intronic regions were covered completely. Primer options output by Primer3 were then compared to reference sequences of *japonica* and *indica* varieties using the Oryza Genome Evolution (OGE) subsection of Gramene to ensure that primers would work across multiple varieties (Tello-Ruiz et al., 2016). DNA was extracted using a modified CTAB method where tissue was grinded using the FASTPRE 24 and PVP was not used (Healey et al., 2014). PCR was performed using Q5 polymerase and product quality was observed by running each sample on a 1% agarose gel (New England BioLabs, Ipswich, MA). Upon confirming the PCR product, samples were sequenced using Sanger sequencing (Laboratory for Genome Technology, Borlaug Center, Texas A&M University), producing sequences in the forward and reverse directions. Sequence alignments were performed using the MegAlign program included in the DNASTAR Lasergene suite. Single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELS) were only considered significant if they were found in both the forward and reverse sequences. If a SNP or INDEL was identified in an exon, the open reading frame (ORF) was identified for each region and was aligned to the ORF of the cDNA for the gene of interest. MegAlign was then used to convert the DNA sequences to protein alignments to identify changes made by the mutation.

IV.4 Results

IV.4.1 Varietal and Environmental Effects on Days to Heading

When considering effects on the days to flowering, variety had the largest and most significant impact on when a plant would flower (Table 8). Environment did not have a

significant impact on days to flowering; however, this was likely caused by the variability of response to the eCO_2 conditions while the short day and field were consistently similar with each other (Figure 15).

Table 8 Analysis of Variance for days to flowering by variety and environment

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Variety	5589.958	7	798.5655	8.782913	0.000329	2.764199
Environment	529.75	2	264.875	2.913191	0.087537	3.738892
Error	1272.917	14	90.92262			
Total	7392.625	23				

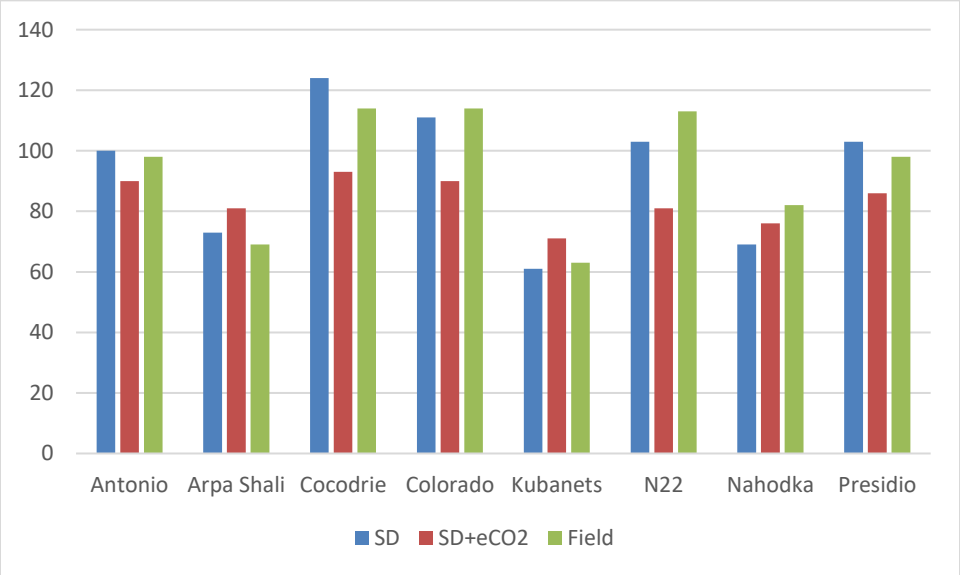


Figure 15 Distribution of days to flowering among environments and varieties

IV.4.2 Variant Effects in Target Genes

All INDELS observed occurred in non-coding regions; however, the INDELS observed in *Hd3a* occurred in the promoter region upstream of the first exon (Table 9). While this may not have an effect on the structure of the protein produced it could impact the transcriptional activation or repression of this gene. The majority of amplicons did not display INDELS and with 2 of 4 INDELS observed being 1 base pair, 1 being 2 base pairs in length, and the final being 14 base pairs (Figure 16); however, due to the location of many of these INDELS it is expected they would not have much of an impact on gene expression or function as they were primarily found in introns which are spliced out prior to translation.

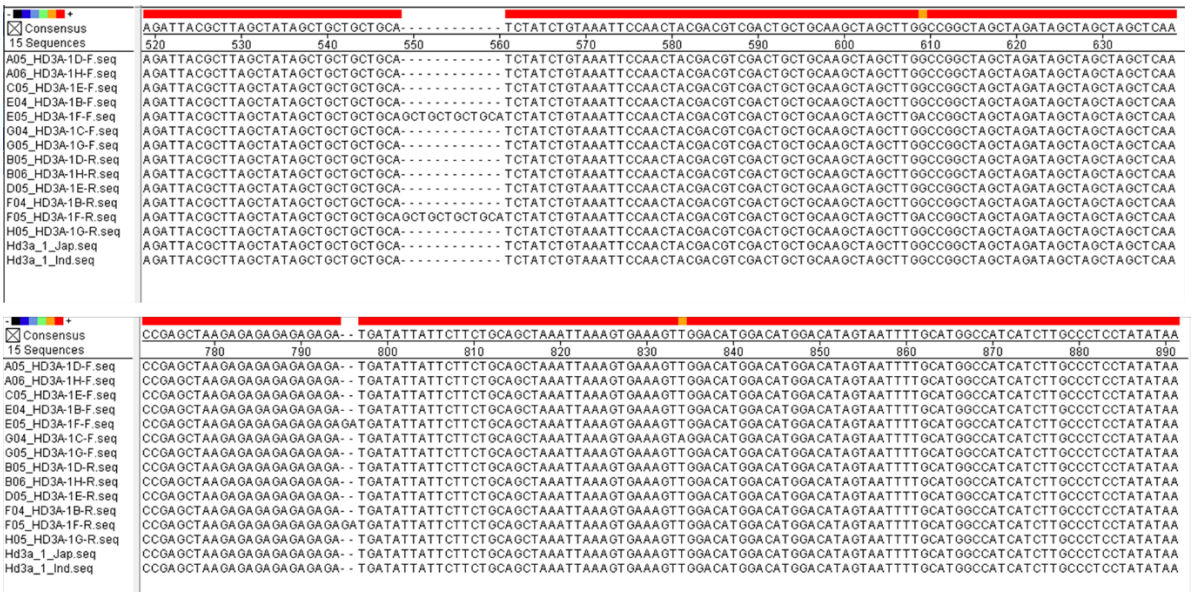


Figure 16 Multiple sequence alignment displaying observed INDELS in promoter region of *Hd3a*

Table 9 Count and effect of INDELS in each gene for all varieties

	Antonio	Arpa Shali	Cocodrie	Colorado	Kubanets	N22	Nahodka	Presidio	INDEL effects
Ghd7_1	0	NA	0	0	NA	0	0	0	
Hd1_3	0	0	0	0	0	0	0	NA	
Hd3a_1	NA	1	NA	0	0	2	0	0	Promoter region
Hd6_3	NA	NA	NA	0	0	1	0	1	Intron
Hd6_5	NA	0	0	0	0	0	0	0	
Hd6_6	0	0	0	0	0	0	0	0	
OsGI_1	NA	NA	NA	0	0	0	0	0	
OsGI_2	0	0	0	0	0	1	0	0	Intron
OsGI_3	0	0	0	0	0	0	0	0	
OsGI_4	NA	0	0	0	0	0	0	0	
OsGI_5	0	0	0	0	0	0	0	0	
OsGI_6	0	0	0	0	0	0	0	0	
OsGI_7	0	0	0	0	0	0	0	0	
OsGI_8	0	0	0	0	0	0	1	0	Intron
OsGI_9	0	0	0	0	0	0	0	0	

Observed SNPs were much higher than INDELs with 18 SNPs observed compared to 4 INDELs. Almost all SNPs were found in introns; however, 4 were found in the promoter region of a gene and 5 were found in exons. All mutations observed in exons were nonsynonymous substitutions leading to missense mutations (Figure 17). It is expected that these could have an impact on protein function but is more likely that the protein will still be functional. The mutations observed in the promoter region could potentially impact transcription of this target gene as they were all found in the same gene and may have changes in the binding sites of transcriptional regulators.

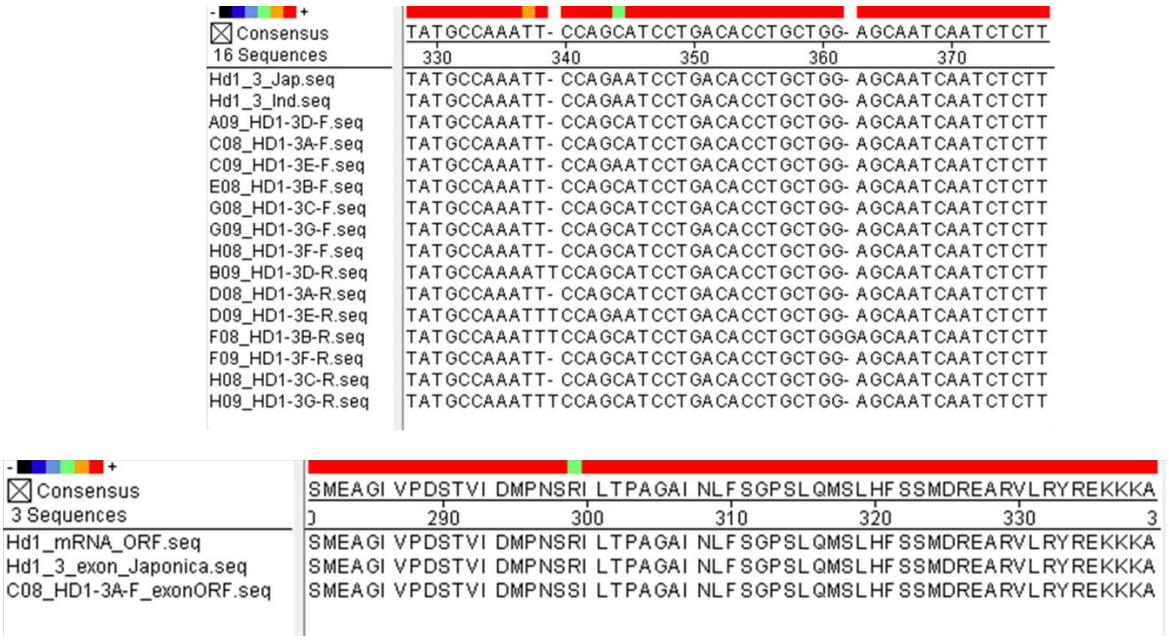


Figure 17 DNA and protein multiple sequence alignment for *Hd1* showing a SNP mutation which causes arginine in the reference genomes to become serine.

Table 10 Count and effect of SNPs in each gene for all varieties

	Antonio	Arpa Shali	Cocodrie	Colorado	Kubanets	N22	Nahodka	Presidio	SNP Effects
Ghd7_1	0	0	0	0	0	1	0	1	Intron
Hd1_3	2	2	2	2	0	2	2	2	2 in exon: arginine to serine, glycine to serine
Hd3a_1	NA	0	NA	0	0	4	0	0	All in upstream region
Hd6_3	NA	NA	NA	0	0	0	0	0	
Hd6_5	NA	0	0	0	0	2	0	2	Intron
Hd6_6	0	0	0	0	0	0	0	0	
OsGI_1	NA	NA	NA	0	0	0	0	0	
OsGI_2	0	0	0	0	0	1	0	0	Intron
OsGI_3	1	0	1	1	0	0	1	1	Intron
OsGI_4	0	0	0	0	0	0	0	0	
OsGI_5	0	0	0	0	0	1	0	0	Intron

Table 110 Continued

	Antonio	Arpa Shali	Cocodrie	Colorado	Kubanets	N22	Nahodka	Presidio	SNP Effects
OsGI_6	0	0	0	0	0	2	0	0	Both in exon: serine to leucine, threonine to asparagine
OsGI_7	0	0	0	0	0	0	0	0	
OsGI_8	1	0	1	1	0	2	0	1	1 in exon, matches indica. Causes valine in japonica to become isoleucine. Indica matches mRNA given by NCBI. 2nd in intron
OsGI_9	2	0	2	2	0	2	0	2	Both intron

IV.5 Discussion

The varieties grown in this experiment showed variable responses to growth in elevated carbon dioxide, including five instances where elevated carbon dioxide led to earlier flowering and three cases where it led to slightly later flowering (Figure 15). These results differfrom those of previous reports that eCO_2 consistently decreases days to flowering (Hasegawa et al., 2016). Although days to flowering in a controlled environment long day condition was not observed, it is likely that these varieties are not photoperiod sensitive as the days to flowering under short day and field conditions, which would have been closer to long day at that time of year, are very similar to each other. Furthermore, when comparing the short day and short day plus elevated carbon dioxide chambers, the two had different photon flux measurements (Appendix B). This

has the potential for impacting the days to flowering, as light is a primary signal for the initiation of days to flowering. Moreover, the carbon dioxide concentration was not measured in the ambient chamber, which could show larger fluctuations in concentration than would be observed in field conditions.

The structure of the genes observed is highly conserved through all genotypes with few genotypes showing missense mutations. This suggests the flowering pathway previously described is conserved among varieties from diverse origins (Hori et al., 2016). Complementation studies using base editing can be used to confirm the impact of these missense SNPs. In the future it may be beneficial to observe the transcriptional expression of target genes displaying mutations in the upstream promoter region to see if these SNPs and INDELs change the level of expression observed. However, although the genes currently observed are conserved amongst these varieties, there are at least 2 dozen genes known to contribute to days to flowering. As a result, there are other genes which could impact the days to flowering in these varieties, leading to the highly significant variation in genotypes. In the future other previously-described flowering time genes will be sequenced in order to identify sources of variation which may contribute to early flowering in rice.

CHAPTER V

CRISPR/CAS9 GENOME EDITING OF FLOWERING TIME REGULATORS IN TEXAS RICE VARIETY PRESIDIO UTILIZING DNA-FREE METHODS

V.1 Synopsis

The CRISPR/Cas9 system offers opportunities to speed up breeding processes for crops through targeted editing of genes controlling traits of interest. USDA has recently ruled that CRISPR edited crops will not be regulated as transgenic crops. However, in order to avoid regulation, all transgene materials must be removed from the target genome. We propose using a ribonucleoprotein (RNP) complex approach as RNPs will naturally be degraded and this DNA-free approach does not involve integration into the genome. Beyond avoiding integration of the Cas9 DNA, RNP approaches further decrease the occurrence of off-target events due to their transient nature. This experiment aims to optimize RNP approaches in the Texas rice variety Presidio, using *Hd3a* and *RFT1* as target genes of interest.

V.2 Introduction

Cas9 is a bacterial derived, RNA-guided, double strand nuclease that can be utilized to mutate specific genes of interest by the CRISPR/Cas9 system (Doudna and Charpentier, 2014). This system has been shown to create targeted edits in a variety of plant species including Arabidopsis, wheat, tomato, sorghum, and rice (Belhaj et al., 2015). At this time, CRISPR/Cas9 edited plants are not being classified as genetically modified organisms (GMOs), leading to an easier path for commercialization of an edited crop (Waltz, 2016). When performing edits using CRISPR/Cas9, the only requirements are the Cas9 protein combined with either a CRISPR RNA

(crRNA) and trans CRISPR RNA (tracrRNA) or a single guide RNA (sgRNA) where the crRNA and tracrRNA have been combined into one RNA (Belhaj et al., 2013). As a result, there is no need for the Cas9 and gRNA components to be stably integrated into the genome for edits to occur and an alternative ribonucleoprotein complex (RNP) approach can be utilized (Woo et al., 2015).

RNP genome editing has proven effective in crop species such as rice, wheat, Arabidopsis, tobacco, and lettuce (Woo et al., 2015, Liang et al., 2017). When compared to integration of Cas9 and gRNA material into the genome, the occurrence of off -target effects is expected to decrease as the Cas9 nuclease and gRNA are not continuously expressed (Liang et al., 2017). Furthermore, current USDA regulations require that all Cas9 transgene material is removed before a plant can be commercialized. Using an RNP approach removes the need to rid the Cas9 gene from edited plants as the protein will be naturally degraded through cellular processes. This project aims to optimize RNP delivery methods in rice for knocking out the gene function of *Hd3a* and *RFT1*, two genes known to contribute to the days to flowering.

V.3 Materials and Methods

V.3.1 Guide RNA Design

DNA was extracted from Presidio leaves using the modified CTAB method (Healey et al., 2014). Primers were designed using Primer3 to amplify the first exon of *Hd3a* and *RFT1* based on both *japonica* and *indica* reference genomes (Koressaar and Remm, 2007, Untergasser et al., 2012). PCR was performed using DreamTaq polymerase (Thermo Fisher, Waltham, MA) and product quality was confirmed by running on a 1% agarose gel. PCR products were then analyzed using Sanger sequencing. Cas-Designer and CRISPRdirect were then used to design

guide RNAs with two gRNA chosen for each targeted gene to ensure with the aim of ensuring a complete knockout of the gene of interest (Naito et al., 2015, Park et al., 2015). Synthetic gRNA and Cas9 protein were ordered from Synthego (Menlo Park, CA).

V.3.2 Protoplast Isolation and RNP Transfection

Dehulled seeds sterilized using 70% bleach were grown on Murashige and Skoog media for 12 days in dark conditions at 25 ºC (Li and Murai, 1990). Etiolated shoots were chopped into fine pieces and digested with Macroenzyme and Cellulase (Yakult, Tokyo, Japan) with vacuum infiltration at -20 kPa for 30 minutes. This was followed by a 4-hour incubation with gentle shaking in the dark. Protoplasts were then filtered through a 38-µm cell filter and Miracloth. The protoplasts were centrifuged and washed before they were quantified using a hemocytometer under 100x magnification (Jabnoune et al., 1993). RNP complexes were formed following the manufacturer's protocol (Synthego) and were inserted into protoplasts using a modified polyethylene glycol (PEG) transformation method as displayed in Figure 18 (Hayashimoto et al., 1990). Protoplasts were incubated with the RNP complexes for 16 hours before harvesting using centrifugation. DNA was extracted from harvested protoplasts using the Qiagen DNeasy Plant Mini Kit. PCR was performed using the primers described above and was run on a 1% agarose gel to assess the quality of the product.

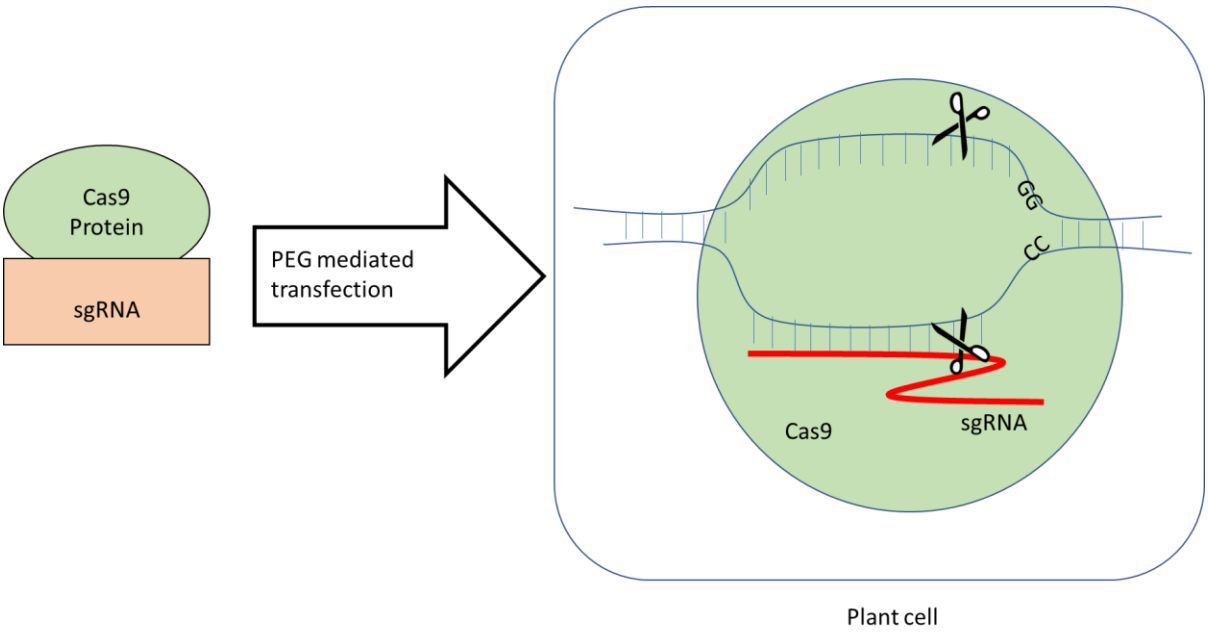


Figure 18 PEG mediated transfection of protoplasts with RNP complexes

V.4 Results

V.4.1 Hd3a gRNA design

The first exon of *Hd3a* in Presidio completely matched the *indica* reference genome which has 2 INDELS and 38 SNPs in comparison to the *japonica* reference genome. This was not expected as Presidio comes from a *japonica* background. Both gRNAs chosen were unique when compared to the rest of the genome. Furthermore, there were no sequences in the genome that matched the 12mer region for gRNA1 while gRNA2 had 4 matching regions and the number of regions matching the 8mer region was minimized with gRNA1 having 453 matches and gRNA2 having 2071 matches. The gRNAs chosen had no overlapping restriction enzyme sites. Therefore, the only way to validate the occurrence of an edit is through the use of a T7E1 assay.

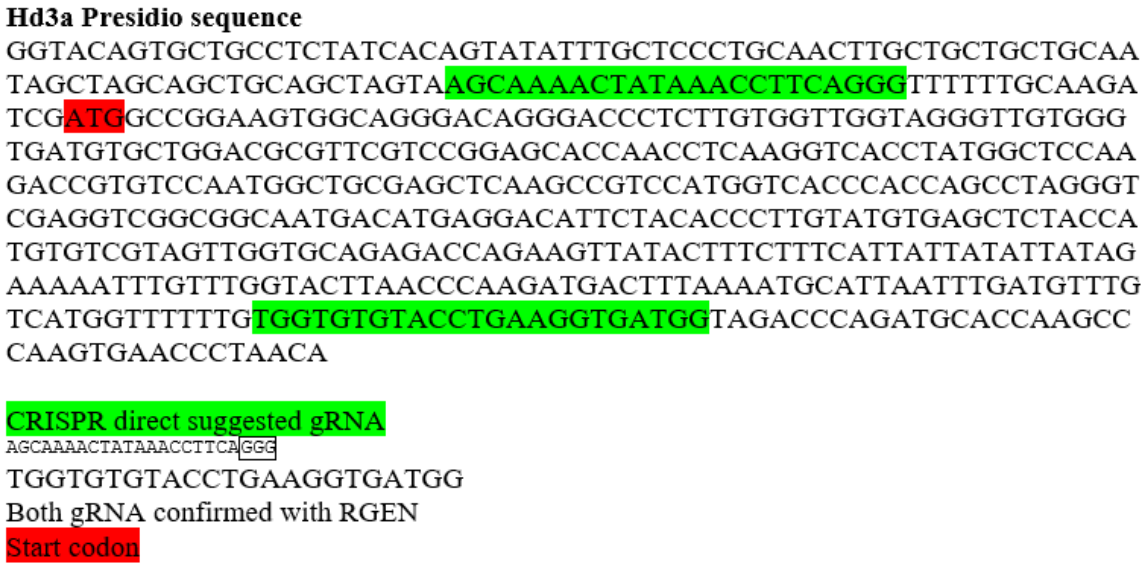


Figure 19 Sequence of Hd3a in Presidio with highlighted gRNA design and start codon

V.4.2 RFT1 gRNA design

The first exon of *RFT1* had 1 SNP in relation to the *indica* reference genome but completely matched the *japonica* reference genome. Both gRNAs are completely unique in relation to the *japonica* reference genome for both the 20mer and 12mer sequences. The 8mer sequence for gRNA1 matches 650 sequences while this same region in gRNA2 matches 813 sequences. Neither of these gRNAs had a restriction enzyme site associated with them. Therefore, the T7E1 assay is the best way to validate edits using these gRNAs (Figure 19).

RFT1 Presidio sequence

GTCGTCATTGTTTTTCCACGACTTCTCTGGATTGACGGCAGGAGATACCTAAGCTAG
CTAGCAATCTCTATCGATCTGTTTGTTTACATGTTCAGTTAAAGGTTACTGAGAAATG
CCTAGAGTTTTTCCGGCTAGCTTCATAAGTTAGTGGGTTAGCTGACCTAGATTCAAA
GTCTAATCCTTTTATTTATTTGATATTAGATATCCTAACGTTTTTAGTTAGAGGTTATT
AATTTGACATGGCCGGCAGCGGCAGGGACGATCCTCTTGTGGTTGGCAGGATTGTGG
GTGATGTGCTGGATCCATTCGTCCGGATCACTAACCTCAGTGTCAGCTATGGTGCAA
GGATCGTCTCCAATGGCTGCGAGCTCAAGCCGTCCATGGTGACCCAACAGCCCAGG
GTCGTGGTCGGTGGCAATGACATGAGGACGTTCTACACACTCGTACGGATCATATCT
TGGATGCAGAGACCCACCAGAAGTTAGA

gRNA1:
CCTAGATTCAAAGTCTAATCCTT

gRNA 2: CAACAGCCCAGGGTCGTGGTCGG

Figure 20 Sequence of RFT1 in Presidio with highlighted gRNA and start codon

V.4.2 Protoplast Isolation and RNP Transfection

The protoplast isolation yielded approximately 100 times fewer protoplasts than are required for PEG-mediated transformation. These protoplasts were then split into four in order to perform RNP transfection with each gRNA. After 16 hours incubation the protoplasts were harvested and DNA was extracted. However, due to the low concentration of protoplasts there was not enough DNA extracted to be properly sensed by the Nanodrop spectrophotometer. We proceeded to attempt performing PCR on these samples; however, this failed, likely due to the low concentration of DNA obtained (Figure 20).

V.5 Discussion

RNP mediated genome editing offers opportunities to create targeted changes within the rice genome while decreasing off-target effects and preventing integration of foreign genes into the rice genome. These methods have previously proven effective in rice; however, in order to

perform these methods, the number of protoplasts harvested needs to be higher. Furthermore, PEG-mediated transformation of protoplasts has not been tested in Presidio. Therefore, it will also be beneficial to use Cas9 fused with green fluorescent protein in order to ensure the Cas9 protein is successfully entering the cell to create edits. The protoplast isolation protocol is currently being optimized to increase the number of protoplasts harvested from Presidio plantlets, and Cas9 culture methods are also being investigated in order to minimize the cost of using commercial Cas9 protein in high concentrations. This would allow performance of *in vitro* edit confirmation to ensure gRNA efficiency before applying RNP methods in the cellular context.

CHAPTER VI

SUMMARY AND CONCLUSIONS

Genetic resources in rice are extensive with multiple tools for genotyping, wide variation among accessions stored within gene banks, and many signaling pathways which have been thoroughly researched. This project aimed to better characterize days to flowering in rice by advancing a low density genotyping tool, performing a genome wide association study, identifying genetic variants in a small panel of early flowering varieties, and designing DNA-free genome editing tools for use in protoplasts. The C7AIR successfully differentiated between the five *Oryza sativa* subgroups and proved effective as a low-cost tool for DNA fingerprinting. Upon improving the automated cluster file, data from the C7AIR was used in a GWAS which identified 5 significant loci. These loci co-located with 2 known flowering time genes, a nuclear transport factor, a leaf senescence protein, and a hypothetical protein. In the smaller panel of varieties, the 5 genes that were sequenced did not show any mutations of apparent effects; however, there are other genes which may contribute to the varietal differences in days to flowering that were observed. Finally, gRNAs have been designed for *RFT1* and *Hd3a* and experiments have been initiated to perform genome editing on these target genes with the aim of knocking these genes out to confirm their function in the genetic background of the Presidio variety.

REFERENCES

Agrama, H.A., W. Yan, F. Lee, R. Fjellstrom, M.-H. Chen, M. Jia, et al. 2009. Genetic assessment of a mini-core subset developed from the USDA rice genebank. Crop Science 49: 1336-1346.

Bae, S., J. Park and J.-S. Kim. 2014. Cas-offinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics 30: 1473-1475.

Belhaj, K., A. Chaparro-Garcia, S. Kamoun and V. Nekrasov. 2013. Plant genome editing made easy: Targeted mutagenesis in model and crop plants using the CRISPR/cas system. Plant methods 9: 1.

Belhaj, K., A. Chaparro-Garcia, S. Kamoun, N.J. Patron and V. Nekrasov. 2015. Editing plant genomes with CRISPR/Cas9. Current Opinion in Biotechnology 32: 76-84.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss and E.S. Buckler. 2007. Tassel: Software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.

Brambilla, V. and F. Fornara. 2013. Molecular control of flowering in response to day length in rice. Journal of Integrative Plant Biology 55: 410-418.

Dingkuhn, M., R. Pasco, J.M. Pasuquin, J. Damo, J.-C. Soulié, L.-M. Raboin, et al. 2017. Crop-model assisted phenomics and genome-wide association study for climate adaptation of indica rice. 1. Phenology. Journal of Experimental Botany 68: 4369-4388.

Dixit, S., A. Singh, N. Sandhu, A. Bhandari, P. Vikram and A. Kumar. 2017. Combining drought and submergence tolerance in rice: Marker-assisted breeding and QTL combination effects. Molecular Breeding 37: 143.

Dou, F., L. Tarpley, K. Chen, A.L. Wright and A.R. Mohammed. 2016. Planting date and variety effects on rice main and ratoon crop production in South Texas. Communications in Soil Science and Plant Analysis 47: 2414-2420.

Doudna, J.A. and E. Charpentier. 2014. The new frontier of genome engineering with CRISPR-Cas9. Science 346: 1258096:1-9.

El Mannai, Y., K. Akabane, K. Hiratsu, N. Satoh-Nagasawa and H. Wabiko. 2017. The NAC transcription factor gene *OsY37* (*ONAC011*) promotes leaf senescence and accelerates heading time in rice. International Journal of Molecular Sciences 18: 2165.

Ganal, M.W., T. Altmann and M.S. Röder. 2009. SNP identification in crop plants. Current Opinion in Plant Biology 12: 211-217.

Garris, A.J., T.H. Tai, J. Coburn, S. Kresovich and S. McCouch. 2005. Genetic structure and diversity in *Oryza sativa* L. Genetics 169: 1631-1638.

Gregory, P.J., J.S.I. Ingram and M. Brklacich. 2005. Climate change and food security. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 2139-2148.

Hasegawa, T., H. Sakai, T. Tokida, Y. Usui, M. Yoshimoto, M. Fukuoka, et al. 2016. Rice free-air carbon dioxide enrichment studies to improve assessment of climate change effects on rice agriculture. In: J. L. Hatfield and D. Fleisher, editors, Improving modeling tools to assess climate change effects on crop response. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, Inc., Madison, WI. p. 45-68.

Hayashimoto, A., Z. Li and N. Murai. 1990. A polyethylene glycol-mediated protoplast transformation system for production of fertile transgenic rice plants. Plant Physiology 93: 857-863.

Healey, A., A. Furtado, T. Cooper and R.J. Henry. 2014. Protocol: A simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. Plant Methods 10: 21-21.

Hori, K., K. Matsubara and M. Yano. 2016. Genetic control of flowering time in rice: Integration of mendelian genetics and genomics. Theoretical and Applied Genetics 129: 1-12.

Huihui, Y., X. Weibo, L. Jing, Z. Fasong and Z. Qifa. 2014. A whole-genome SNP array (rice6k) for genomic breeding in rice. Plant Biotechnology Journal 12: 28-37.

Izawa, T. 2007. Adaptation of flowering-time by natural and artificial selection in arabidopsis and rice. Journal of Experimental Botany 58: 3091-3097.

Jabnoune, M., D. Secco, C. Lecampion, C. Robaglia, Q. Shu and Y. Poirier. 1993. An efficient procedure for protoplast isolation from mesophyll cells and nuclear fractionation in rice. The Plant Cell

Jagadish, S.V.K. 2010. Genetic analysis of heat tolerance at anthesis in rice. Crop Science 50: 1633-1641.

Jung, C. and A.E. Müller. 2009. Flowering time control and applications in plant breeding. Trends in Plant Science 14: 563-573.

Kawahara, Y., M. de la Bastide, J.P. Hamilton, H. Kanamori, W.R. McCombie, S. Ouyang, et al. 2013. Improvement of the *Oryza sativa* nipponbare reference genome using next generation sequence and optical map data. Rice 6: 4.

Kojima, S., Y. Takahashi, Y. Kobayashi, L. Monna, T. Sasaki, T. Araki, et al. 2002. *Hd3a*, a rice ortholog of the arabidopsis *FT* gene, promotes transition to flowering downstream of *Hd1* under short-day conditions. Plant and Cell Physiology 43: 1096-1105.

Komiya, R., A. Ikegami, S. Tamaki, S. Yokoi and K. Shimamoto. 2008. Hd3a and rft1 are essential for flowering in rice. Development 135: 767-774.

Koo, B.-H., S.-C. Yoo, J.-W. Park, C.-T. Kwon, B.-D. Lee, G. An, et al. 2013. Natural variation in *osprr37* regulates heading date and contributes to rice cultivation at a wide range of latitudes. Molecular Plant 6: 1877-1888.

Koressaar, T. and M. Remm. 2007. Enhancements and modifications of primer design program primer3. Bioinformatics 23: 1289-1291.

Kumar, S., G. Stecher and K. Tamura. 2016. Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33: 1870-1874.

Li, Z. and N. Murai. 1990. Efficient plant regeneration from rice protoplasts in general medium. Plant Cell Reports 9: 216-220.

Liang, Z., K. Chen, T. Li, Y. Zhang, Y. Wang, Q. Zhao, et al. 2017. Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. Nature Communications 8: 14261.

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, et al. 2012. Gapit: Genome association and prediction integrated tool. Bioinformatics 28: 2397-2399.

McCouch, S.R., M.H. Wright, C.-W. Tung, L.G. Maron, K.L. McNally, M. Fitzgerald, et al. 2016. Open access resources for genome-wide association mapping in rice. Nature Communications 7: 10532.

Muthayya, S., J.D. Sugimoto, S. Montgomery and G.F. Maberly. 2014. An overview of global rice production, supply, trade, and consumption. Annals of the New York Academy of Sciences 1324: 7-14.

Myles, S., J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich, et al. 2009. Association mapping: Critical considerations shift from genotyping to experimental design. The Plant Cell 21: 2194-2202.

Naito, Y., K. Hino, H. Bono and K. Ui-Tei. 2015. Crisprdirect: Software for designing CRISPR/cas guide RNA with reduced off-target sites. Bioinformatics 31: 1120-1123.

Nasu, S., J. Suzuki, R. Ohta, K. Hasegawa, R. Yui, N. Kitazawa, et al. 2002. Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa* , *Oryza rufipogon*) and establishment of SNP markers. DNA Research 9: 163-171.

Ogiso, E., Y. Takahashi, T. Sasaki, M. Yano and T. Izawa. 2010. The role of Casein Kinase II in flowering time regulation has diversified during evolution. Plant Physiology 152: 808-820.

Ouyang, S., W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, et al. 2006. The tigr rice genome annotation resource: Improvements and new features. Nucleic Acids Research 35: D883-D887.

Park, D.H., D.E. Somers, Y.S. Kim, Y.H. Choy, H.K. Lim, M.S. Soh, et al. 1999. Control of circadian rhythms and photoperiodic flowering by the arabidopsis *GIGANTEA* gene. Science 285: 1579-1582.

Park, J., S. Bae and J.-S. Kim. 2015. Cas-designer: A web-based tool for choice of CRISPR-Cas9 target sites. Bioinformatics 31: 4014-4016.

Raj, A., M. Stephens and J.K. Pritchard. 2014. Faststructure: Variational inference of population structure in large SNP data sets. Genetics 197: 573-589.

Ray, D.K., N.D. Mueller, P.C. West and J.A. Foley. 2013. Yield trends are insufficient to double global crop production by 2050. PLOS One 8: e66428.

Rife, T.W. and J.A. Poland. 2014. Field book: An open-source application for field data collection on android. Crop Science 54: 1624-1627.

Sakai, H., S.S. Lee, T. Tanaka, H. Numa, J. Kim, Y. Kawahara, et al. 2013. Rice annotation project database (RAP-DB): An integrative and interactive database for rice genomics. Plant and Cell Physiology 54: e6-e6.

Shah, F., J. Huang, K. Cui, L. Nie, T. Shah, C. Chen, et al. 2011. Impact of high-temperature stress on rice plant and its traits related to tolerance. The Journal of Agricultural Science 149: 545-556.

Stein, J.C., Y. Yu, D. Copetti, D.J. Zwickl, L. Zhang, C. Zhang, et al. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nature Genetics 50: 285-296.

Tello-Ruiz, M.K., J. Stein, S. Wei, J. Preece, A. Olson, S. Naithani, et al. 2016. Gramene 2016: Comparative plant genomics and pathway resources. Nucleic Acids Research 44: D1133-D1140.

Thomson, M.J., N. Singh, M.S. Dwiyanti, D.R. Wang, M.H. Wright, F.A. Perez, et al. 2017. Large-scale deployment of a rice 6 k SNP array for genetics and breeding applications. Rice 10: 40.

Thomson, M.J., K. Zhao, M. Wright, K.L. McNally, J. Rey, C.-W. Tung, et al. 2012. High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the beadxpress platform. Molecular Breeding 29: 875-886.

Tilman, D., C. Balzer, J. Hill and B.L. Befort. 2011. Global food demand and the sustainable intensification of agriculture. Proceedings of the National Academy of Sciences 108: 20260-20264.

Tung, C.-W., K. Zhao, M.H. Wright, M.L. Ali, J. Jung, J. Kimball, et al. 2010. Development of a research platform for dissecting phenotype–genotype associations in rice (*Oryza* spp.). Rice 3: 205-217.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B.C. Faircloth, M. Remm, et al. 2012. Primer3—new capabilities and interfaces. Nucleic Acids Research 40: e115-e115.

Vergara, B.S. and T.-T. Chang. 1985. The flowering response of the rice plant to photoperiod: A review of the literature. 4th ed. International Rice Research Institute, Los Banos, Laguna, Philippines.

Waltz, E. 2016. Gene-edited CRISPR mushroom escapes US regulation. Nature 532: 293.

Wassmann, R., S.V.K. Jagadish, S. Heuer, A. Ismail, E. Redona, R. Serraj, et al. 2009. Chapter 2 Climate change affecting rice production: The physiological and agronomic basis for possible adaptation strategies. Advances in Agronomy 101: 59-122.

Woo, J.W., J. Kim, S.I. Kwon, C. Corvalan, S.W. Cho, H. Kim, et al. 2015. DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. Nature Biotechnology 33: 1162-1164.

Xue, W., Y. Xing, X. Weng, Y. Zhao, W. Tang, L. Wang, et al. 2008. Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. Nature Genetics 40: 761-767.

Yan, W., J.N. Rutger, R.J. Bryant, H.E. Bockelman, R.G. Fjellstrom, M.-H. Chen, et al. 2007. Development and evaluation of a core subset of the USDA rice germplasm collection. Crop Science 47: 869-876.

Yano, M., Y. Katayose, M. Ashikari, U. Yamanouchi, L. Monna, T. Fuse, et al. 2000. Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the arabidopsis flowering time gene Constans. The Plant Cell 12: 2473-2483.

Zafar, S.A., A. Hameed, M.A. Nawaz, M. Wei and M.A. Noor. 2017. Mechanisms and molecular approaches for heat tolerance in rice (*Oryza sativa* L.) under climate change scenario. Journal of Integrative Agriculture 16: 60345-60347.

Zhang, Y., J. Su, S. Duan, Y. Ao, J. Dai, J. Liu, et al. 2011. A highly efficient rice green tissue protoplast system for transient gene expression and studying light/chloroplast-related processes. Plant methods 7: 30.

Zhao, K., C.-W. Tung, G.C. Eizenga, M.H. Wright, M.L. Ali, A.H. Price, et al. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nature communications 2: 467.

Zhao, Q., S. Leung, A.H. Corbett and I. Meier. 2006. Identification and characterization of the arabidopsis orthologs of nuclear transport factor 2, the nuclear import factor of ran. Plant Physiology 140: 869-878.

APPENDIX A

VARIETES GENOTYPED FOR USE IN CLUSTERING AND GWAS

Original Samples	Source	Clustering	GWAS
WG0252580-DNAA01	Cornell	x	
WG0252580-DNAA02	Cornell	x	
WG0252580-DNAA03	Cornell	x	
WG0252580-DNAA04	Cornell	x	
WG0252580-DNAA05	Cornell	x	
WG0252580-DNAA06	Cornell	x	
WG0252580-DNAA07	Cornell		
WG0252580-DNAA08	Cornell	x	
WG0252580-DNAA09	Cornell	x	
WG0252580-DNAA10	Cornell	x	
WG0252580-DNAA11	Cornell	x	
WG0252580-DNAA12	Cornell	x	
WG0252580-DNAB01	Cornell	x	
WG0252580-DNAB02	Cornell	x	
WG0252580-DNAB03	Cornell	x	
WG0252580-DNAB04	Cornell	x	
WG0252580-DNAB05	Cornell	x	
WG0252580-DNAB06	Cornell	x	
WG0252580-DNAB07	Cornell		
WG0252580-DNAB08	Cornell	x	
WG0252580-DNAB09	Cornell		
WG0252580-DNAB10	Cornell		
WG0252580-DNAB11	Cornell		
WG0252580-DNAB12	Cornell	x	
WG0252580-DNAC01	Cornell	x	
WG0252580-DNAC02	Cornell	x	
WG0252580-DNAC03	Cornell	x	
WG0252580-DNAC04	Cornell		
WG0252580-DNAC05	Cornell	x	
WG0252580-DNAC06	Cornell	x	
WG0252580-DNAC07	Cornell		
WG0252580-DNAC08	Cornell		
WG0252580-DNAC09	Cornell		
WG0252580-DNAC10	Cornell	x	
WG0252580-DNAC11	Cornell		
WG0252580-DNAC12	Cornell	x	
WG0252580-DNAD01	Cornell	x	

WG0252580-DNAD02	Cornell	x
WG0252580-DNAD03	Cornell	x
WG0252580-DNAD04	Cornell	
WG0252580-DNAD05	Cornell	x
WG0252580-DNAD06	Cornell	x
WG0252580-DNAD07	Cornell	
WG0252580-DNAD08	Cornell	x
WG0252580-DNAD09	Cornell	x
WG0252580-DNAD10	Cornell	x
WG0252580-DNAD11	Cornell	x
WG0252580-DNAD12	Cornell	x
WG0252580-DNAE01	Cornell	x
WG0252580-DNAE02	Cornell	x
WG0252580-DNAE03	Cornell	
WG0252580-DNAE04	Cornell	
WG0252580-DNAE05	Cornell	x
WG0252580-DNAE06	Cornell	x
WG0252580-DNAE07	Cornell	
WG0252580-DNAE08	Cornell	
WG0252580-DNAE09	Cornell	x
WG0252580-DNAE10	Cornell	x
WG0252580-DNAE11	Cornell	x
WG0252580-DNAE12	Cornell	x
WG0252580-DNAF01	Cornell	x
WG0252580-DNAF02	Cornell	x
WG0252580-DNAF03	Cornell	x
WG0252580-DNAF04	Cornell	
WG0252580-DNAF05	Cornell	
WG0252580-DNAF06	Cornell	x
WG0252580-DNAF07	Cornell	
WG0252580-DNAF08	Cornell	
WG0252580-DNAF09	Cornell	x
WG0252580-DNAF10	Cornell	x
WG0252580-DNAF11	Cornell	x
WG0252580-DNAF12	Cornell	x
WG0252580-DNAG01	Cornell	x
WG0252580-DNAG02	Cornell	x
WG0252580-DNAG03	Cornell	x
WG0252580-DNAG04	Cornell	x
WG0252580-DNAG05	Cornell	x
WG0252580-DNAG06	Cornell	x

WG0252580-DNAG07	Cornell	
WG0252580-DNAG08	Cornell	x
WG0252580-DNAG09	Cornell	x
WG0252580-DNAG10	Cornell	x
WG0252580-DNAG11	Cornell	x
WG0252580-DNAG12	Cornell	x
WG0252580-DNAH01	Cornell	x
WG0252580-DNAH02	Cornell	x
WG0252580-DNAH03	Cornell	x
WG0252580-DNAH04	Cornell	x
WG0252580-DNAH05	Cornell	x
WG0252580-DNAH06	Cornell	x
WG0252580-DNAH07	Cornell	x
WG0252580-DNAH08	Cornell	x
WG0252580-DNAH09	Cornell	x
WG0252580-DNAH10	Cornell	x
WG0252580-DNAH11	Cornell	x
WG0252580-DNAH12	Cornell	x
201429710032_R01C01	IRRI	x
201429710032_R01C02	IRRI	x
201429710032_R02C01	IRRI	x
201429710032_R02C02	IRRI	x
201429710032_R03C01	IRRI	x
201429710032_R03C02	IRRI	x
201429710032_R04C01	IRRI	x
201429710032_R04C02	IRRI	x
201429710032_R05C01	IRRI	x
201429710032_R05C02	IRRI	x
201429710032_R06C01	IRRI	x
201429710032_R06C02	IRRI	x
201429710032_R07C01	IRRI	x
201429710032_R07C02	IRRI	x
201429710032_R08C01	IRRI	x
201429710032_R08C02	IRRI	x
201429710032_R09C01	IRRI	x
201429710032_R09C02	IRRI	x
201429710032_R10C01	IRRI	x
201429710032_R10C02	IRRI	x
201429710032_R11C01	IRRI	x
201429710032_R11C02	IRRI	x
201429710032_R12C01	IRRI	x

201429710032_R12C02	IRRI	x	
201429710137_R01C01	IRRI	x	
201429710137_R01C02	IRRI	x	
201429710137_R02C01	IRRI	x	
201429710137_R02C02	IRRI	x	
201429710137_R03C01	IRRI	x	
201429710137_R03C02	IRRI	x	
201429710137_R04C01	IRRI	x	
201429710137_R04C02	IRRI	x	
201429710137_R05C01	IRRI	x	
201429710137_R05C02	IRRI	x	
201429710137_R06C01	IRRI	x	
201429710137_R06C02	IRRI	x	
201429710137_R07C01	IRRI	x	
201429710137_R07C02	IRRI	x	
201429710137_R08C01	IRRI	x	
201429710137_R08C02	IRRI	x	
201429710137_R09C01	IRRI	x	
201429710137_R09C02	IRRI	x	
201429710137_R10C01	IRRI	x	
201429710137_R10C02	IRRI	x	
201429710137_R11C01	IRRI	x	
201429710137_R11C02	IRRI	x	
201429710137_R12C01	IRRI	x	
201429710137_R12C02	IRRI	x	
PURPLE	TAMU	x	x
Kin Shan Zim	TAMU	x	x
N 32	TAMU	x	x
*Dular	TAMU	x	x
Hashikalmi Aus	TAMU	x	x
Kataktara Aus	TAMU	x	x
Bala	TAMU	x	x
Angkrang	TAMU	x	x
Srav Prapay	TAMU	x	x
WAB450-I-B-P-62-HB	TAMU	x	x
WAB450-I-B-P-160-HB	TAMU	x	
Pan Ju	TAMU	x	x
Dhala Shaitta	TAMU	x	x
Early No. 1	TAMU	x	x
Sel. No. 388	TAMU	x	x
SHIMIZU MOCHI	TAMU	x	x

Charmarumi	TAMU	x	x
Dharial	TAMU	x	x
N 22	TAMU	x	
ARC 6578	TAMU	x	x
Tia Heret	TAMU	x	x
Pelu	TAMU	x	x
Sufaida	TAMU	x	x
Jhona	TAMU	x	x
Mahlar	TAMU	x	x
P 22	TAMU	x	x
Ziri	TAMU	x	x
Sathra	TAMU	x	x
Ai Chueh Ta Pai Ku	TAMU	x	x
N 22	TAMU		
Sigoendaba	TAMU	x	x
Sabharaj	TAMU	x	x
*ASWINA 330	TAMU	x	x
DNJ 151	TAMU	x	x
SUNG LIAO 2	TAMU	x	
Trandeup Kandir	TAMU	x	x
Banajira	TAMU	x	x
Brondol	TAMU	x	x
DJ 90	TAMU	x	x
DJ 102	TAMU	x	x
Nam Dawk Mai	TAMU	x	
DV 132	TAMU	x	x
Gubuh Balai	TAMU	x	
Bengawan	TAMU	x	x
Sug	TAMU	x	x
Calmochi-101	TAMU	x	
Pakhe dhan	TAMU	x	
Ghorbhai	TAMU	x	x
Ak Tokhum	TAMU	x	x
EMBRAPA 1200	TAMU	x	x
Jefferson	TAMU	x	x
Bhuwa Dhan	TAMU	x	
Caucasica	TAMU	x	x
Vulgaris	TAMU	x	
ARPA SHALI	TAMU	x	x
UZ ROSZ 5	TAMU	x	
Kyzyl Shala	TAMU	x	

GHATI KAMMA NANGARHAR	TAMU	x		
Shato Lua	TAMU	x		
Kabre	TAMU			x
Atemo	TAMU	x		
Dulugu	TAMU	x		
Dulugu	TAMU			
Boma	TAMU	x		x
Kabre	TAMU	x		x
TOg 5603	TAMU			
TOg 5882	TAMU	x		
TOg 6231	TAMU			
TOg 6238	TAMU			
IRGC-103571	TAMU	x		x
WAB450-11-1-3-P40-HB	TAMU	x		x
WAB450-24-2-3-P33-HB	TAMU	x		x
WAB450-24-3-P38-1-HB	TAMU	x		x
WAB450-24-3-2-P18-HB	TAMU	x		
WAB450-I-B-P-24-HB	TAMU	x		
WAB450-I-B-P-38-HB	TAMU	x		x
Kharsu	TAMU	x		x
Long Gnar Jim	TAMU	x		x
Mayang Khang	TAMU	x		
Century Patna Original	TAMU	x		x
Spin Mere	TAMU	x		x
Secano do Brazil	TAMU	x		x
Sigadis	TAMU	x		x
Red Khosha Cerma	TAMU	x		x
GHRAIBA	TAMU	x		x
Bala	TAMU			x
Khao Phoi	TAMU	x		x
Karayal	TAMU	x		x
PURPLE Rep	TAMU	x		x
C 8447	TAMU	x		x
Padi Pohon Batu	TAMU	x		x
Ratna	TAMU	x		x
Jhona	TAMU			x
Heo Trang	TAMU	x		x

Sipirasikkam	TAMU	x	x
Lantjang	TAMU	x	x
Ali Combo	TAMU	x	x
Torh	TAMU	x	x
Sugdasi	TAMU	x	x
Jira Shahi	TAMU	x	x
Kerr Sail	TAMU	x	x
AS 46	TAMU	x	x
Janeri	TAMU	x	
Kuning Tinggi	TAMU	x	x
Koi Murali	TAMU	x	x
Angana	TAMU	x	
DJ 53	TAMU	x	x
DM 55	TAMU	x	x
DV 85	TAMU	x	x
79	TAMU	x	
Daudzai Field Mix	TAMU	x	x
*M202	TAMU	x	x
*NIPPONBARE	TAMU	x	
Achar Dhog	TAMU		
Jhinga Sail	TAMU	x	
Bak Tushi	TAMU	x	x
Gambir	TAMU	x	x
Shoni	TAMU	x	x
*FIROOZ	TAMU	x	x
Kalamkati	TAMU		
Dom Zard	TAMU	x	
Sirkat	TAMU	x	
Keriting Tinggi	TAMU	x	
Shali-i-Luk	TAMU	x	
Berenj	TAMU	x	
2	TAMU	x	
Safut Khosha	TAMU	x	
Shevkati Kundry	TAMU	x	x
Uz Rosz 17	TAMU	x	x
UZ ROSZ 269	TAMU	x	x
Uz Rosz 2741	TAMU	x	
Uz Rosz 2832	TAMU	x	x
Vrosz 213	TAMU	x	x
UZ ROSZ M8	TAMU	x	x
Uz Rosz M9	TAMU	x	x

Uz Rosz 215	TAMU	x	
Nahodka	TAMU	x	x
VILKID ZIRE	TAMU	x	x
Primorsk 6	TAMU	x	
KUBAN 3	TAMU	x	x
Uz Ros 275	TAMU	x	x
AZ ROS 637	TAMU	x	
SADRI MASALINSKIJ	TAMU	x	x
Ambarby White	TAMU	x	x
Shirkati	TAMU	x	
Besudi Short-Grain	TAMU	x	x
Maien Garm	TAMU	x	
Jubilejnyj	TAMU		
Dera Wadi 1/43	TAMU	x	x
Cat 1747	TAMU	x	x
P 817	TAMU	x	x
Y 2272	TAMU	x	
Vulgaris Ko Ch Azpasaly	TAMU	x	x
Dieebeca Bat Vros 3716	TAMU		
UZ ROS 59	TAMU	x	x
Subdi Chroa Kora Muazah Tolinsty	TAMU	x	
Italica Alef Ambeste Royj	TAMU	x	x
P 1042	TAMU	x	x
Caloro	TAMU	x	
Koshihikari	TAMU	x	
Colorado	TAMU	x	
IR64 Sub1	TAMU	x	
Quilla 66304	TAMU	x	
IR42	TAMU	x	
Fujisaka 5	TAMU	x	
Chuncheongbyeo	TAMU	x	
Cocodrie	TAMU	x	
Diamante	TAMU	x	
Antonio	TAMU	x	
P 1048	TAMU	x	x
HZ ROS 637	TAMU	x	x
Affinis	TAMU	x	
Kasakstanica	TAMU	x	
Vavilovi	TAMU	x	x

Nigrescens	TAMU	x	x
Melanotrix	TAMU	x	x
Ak Tohum	TAMU	x	x
Dicolorata	TAMU	x	x
Az Ros 1646	TAMU	x	x
Bak Saly Mestnyj	TAMU	x	x
DONSKOJ 2	TAMU	x	x
Dv Ros 0219	TAMU	x	x
Dv Ros 2568	TAMU	x	x
Hokkajdo	TAMU	x	
Ali Combo Rep	TAMU	x	
Kasaki Shala Mestnyj	TAMU	x	x
Kesa	TAMU		x
KUBANETS 508	TAMU		x
Mallai	TAMU		
Severnyj	TAMU		x
UZBEKSKIJ 2	TAMU	x	x
Uz Ros 421	TAMU	x	x
Zolotyje Vshody	TAMU	x	
Pioner 320	TAMU	x	x
Krasnodarski	TAMU	x	x
LUK TAKHAR	TAMU	x	x
WIR 623	TAMU	x	x
Chernyj	TAMU	x	
WIR 1528	TAMU	x	x
Hi Muke	TAMU	x	x
WIR 2623	TAMU	x	x
Celiaj	TAMU	x	x
UZ ROS 2759	TAMU		x
WIR 3764	TAMU	x	
Kendzo	TAMU	x	
KROS 358	TAMU	x	x
NF-1	TAMU	x	x
NF-9	TAMU	x	x
SPALCIK	TAMU	x	x
M-667	TAMU	x	
INTENSIVNYJ	TAMU	x	x
ZEMCYZNYJ	TAMU	x	x
WIR 3419	TAMU	x	x
DAL'RIS 13	TAMU	x	
WIR 2462	TAMU	x	

Shala	TAMU	x	x
Styk	TAMU	x	x
Gidej	TAMU	x	
Bankoram	TAMU		x
Saka	TAMU	x	
Mala	TAMU	x	
Sanganyan	TAMU	x	x
Sakatiana	TAMU	x	
TOg 5548	TAMU	x	x
TOg 6248	TAMU		
TOg 6264	TAMU		
TOg 7201	TAMU		
WAB450-I-B-P-20-HB	TAMU	x	
Presidio	TAMU	x	x
Quinimpol	TAMU	x	x
Stormproof	TAMU	x	
Creole Bred	TAMU	x	x
Kerang Serang	TAMU	x	x
Sadri Type	TAMU	x	
Karang Serang	TAMU	x	x
Kerang Serang Selection	TAMU		
Criollo Chivacoa 2	TAMU	x	x
Juppa	TAMU	x	x
Peta	TAMU	x	x
Kamodi	TAMU	x	
PATNAI 23	TAMU	x	x
WC 4443	TAMU		x
Hassawi	TAMU	x	x
Samanis	TAMU	x	x
Ash Kata Aus	TAMU	x	x
ARC 10638	TAMU	x	x
ARC 11524	TAMU	x	
ARC 11611	TAMU	x	x
*Basmati	TAMU	x	x
Achhame	TAMU	x	x
Gendjah Banten	TAMU	x	x
WW 8/2290	TAMU	x	x
VARY LAVA 9	TAMU	x	x
Vary Vato 275	TAMU	x	x
Bengaly Morino 120	TAMU	x	x
Jayanthi	TAMU	x	x

Mushkan	TAMU	x	x
Akabona	TAMU	x	x
Aus 8	TAMU	x	x
Ngoba	TAMU	x	x
*IR64	TAMU	x	x
Karngi	TAMU	x	
Dudhel	TAMU	x	x
Mushkan	TAMU		x
Toga	TAMU	x	
Besudi Long-Grain	TAMU	x	x
Barah	TAMU	x	
Dehraduni	TAMU	x	
Qumanani	TAMU	x	x
Shinali	TAMU	x	
P 807	TAMU	x	x
Uz Begohef 2	TAMU	x	
Guanicagust Soclri Masayensnif	TAMU	x	
Known Kros 358	TAMU	x	
P 1041	TAMU	x	x
P 1049	TAMU	x	x
Azerbaidjanica	TAMU	x	x
WIR 3412	TAMU	x	
TOg 6281	TAMU		
TOg 6288	TAMU		
Zolotyje Vshody Rep	TAMU	x	
TOg 6302	TAMU		
TOg 6303	TAMU		
TOg 6307	TAMU		
TOg 6314	TAMU		
TOg 6328	TAMU		
TOg 6328B	TAMU		x
TOg 6335	TAMU		x
TOg 6342	TAMU		
TOg 6367	TAMU		x
TOg 6392	TAMU		
TOg 6405	TAMU		x
TOg 6422	TAMU		
TOg 6464	TAMU		
TOg 6465	TAMU		
TOg 6468	TAMU		
TOg 6474	TAMU		

TOg 6511	TAMU		x
TOg 6512	TAMU		
TOg 6916	TAMU		x
TOg 6943	TAMU		
TOg 6946	TAMU		
TOg 6951	TAMU		
TOg 7194	TAMU	x	
TOg 7199	TAMU	x	
CG 14	TAMU		
WAB450-I-B-P-23-HB	TAMU	x	x
Pokkali	TAMU	x	x
Simpor	TAMU	x	x
Putih Montor	TAMU	x	x
Bulu Pote	TAMU	x	
Daudzai	TAMU	x	
Pakkali	TAMU	x	x
Kirak	TAMU	x	x
I-363	TAMU	x	
Suga Paukhi Dhan	TAMU	x	
Latisail	TAMU	x	x
Padi Kasalle	TAMU	x	x
CHAHORA 144	TAMU	x	x
Kalo Marsi	TAMU		x
TOg 6244	TAMU		
TOg 6249	TAMU		
TOg 6250	TAMU	x	
TOg 6251	TAMU		
TOg 6253	TAMU		
TOg 6259	TAMU		
TOg 6266	TAMU		
TOg 6271	TAMU		
TOg 7174	TAMU	x	
RU1303138	TAMU	x	
RU0803147	TAMU	x	
RU1303153	TAMU	x	
RU0803153	TAMU	x	
RU1003123	TAMU	x	
RU1503175	TAMU	x	
RU1503147	TAMU	x	
RU1603138	TAMU	x	
RU1403141	TAMU	x	

RU1603144	TAMU	x
RU1603178	TAMU	x
RU1603113	TAMU	x
RU1003098	TAMU	x
RU1303181	TAMU	x
RU1403089	TAMU	x
RU1403138	TAMU	x
RU1403153	TAMU	x
RU1503169	TAMU	x
RU1603086	TAMU	x
RU1603089	TAMU	x
RU1603116	TAMU	
RU1603123	TAMU	x
CL 111	TAMU	x
CL 153	TAMU	x
PRESIDIO	TAMU	
MERMENTAU	TAMU	x
JUPITER	TAMU	x
WELLS	TAMU	x
LAKAST	TAMU	x
DIAMOND	TAMU	x
MM-14	TAMU	x
REX	TAMU	x
CHENIERE	TAMU	x
COCODRIE	TAMU	x
CL272	TAMU	x
ROY J	TAMU	x
TITAN	TAMU	x
JAZZMAN 2	TAMU	x
CL 172	TAMU	x
M206	TAMU	x
CL 163	TAMU	x
DELLA 2	TAMU	x
ANTONIO	TAMU	x
THAD	TAMU	x
RONDO	TAMU	x
CL 151	TAMU	x
IR64_Sub1	TAMU	x
TOg 6342 rep	TAMU	
IR64 (IRRI)	TAMU	x
Ciherang_Sub1_AG1	TAMU	x

Ciherang_Sub1	TAMU	x
Ciherang	TAMU	x
IR64_AG1	TAMU	x
Darij 8	TAMU	x
IR64 (Beaumont)	TAMU	x

If a variety was not used in clustering it was not used in the final data analysis of the GWAS as the genotyping data did not have good quality.

APPENDIX B

SPECTRAL ANALYSIS OF GROWTH CHAMBERS UTILIZED IN CHAPTER IV

Short-day chamber

Quantum ratio: 15.82

Photosynthetically active radiation: 464.5

Short-day plus elevated carbon dioxide chamber

Quantum ratio: 21.57

Photosynthetically active radiation: 494.5

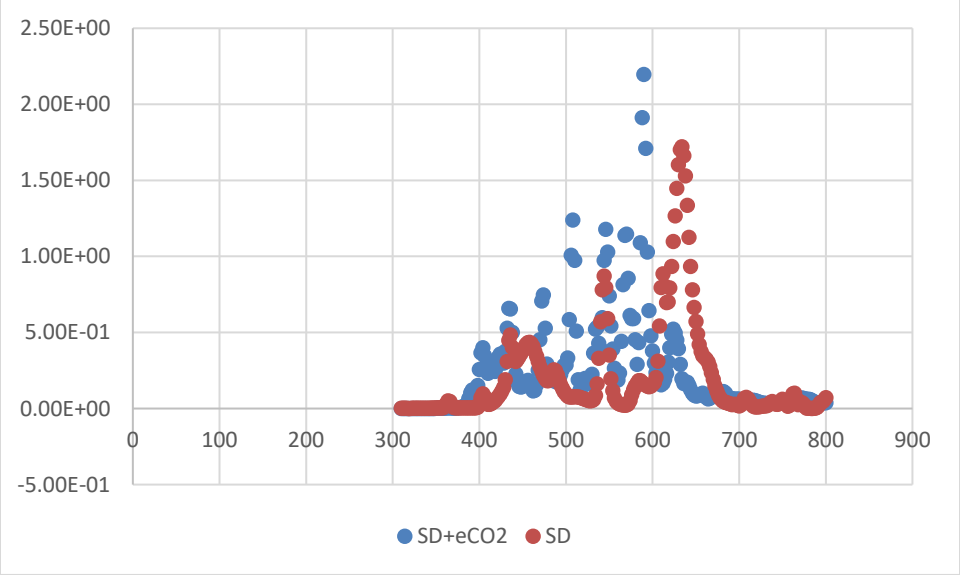


Figure 21 Photon flux for each wavelength of light in both chambers