January 2009

Evolution of Endosperm Starch Synthesis Pathway genes in the Context of Rice: Oryza sativa) Domestication

Guoqin Yu
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Program of Evolution, Ecology, and Population Biology

Dissertation Examination Committee:
Barbara A. Schaal, Chair
Kenneth Olsen, Co-chair
Jim Cheverud
Justin Fay
Tiffany Knight
Mick Richardson

EVOLUTION OF ENDOSPERM STARCH SYNTHESIS PATHWAY GENES IN THE CONTEXT OF RICE (*Oryza sativa*) DOMESTICATION

By

Guoqin Yu

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
Or Doctor of Philosophy

December 2009

Saint Louis, Missouri

Abstract of the Dissertation

The evolution of metabolic pathways is a fundamental but poorly understood aspect of evolutionary change. The rice endosperm starch biosynthetic pathway is one of the most thoroughly characterized biosynthesis pathways in plants, and starch is a trait that has evolved in response to strong selection during rice domestication and subsequent crop improvement. In this study, I have examined six key genes in the rice endosperm starch biosynthesis pathway to investigate the evolution of this pathway before rice domestication and during rice domestication. *Oryza rufipogon* is the wild ancestor of cultivated rice (*Oryza sativa*). *Oryza sativa* has five variety groups: *aus, indica, tropical japonica, temperate japonica and aromatic*. I have sequenced five genes (*shrunken2, Sh2; brittle2, Bt2; waxy, Wx; starch synthase IIa, SsIIa; starch branching enzyme IIb, SbeIIb;* and *isoamylase1, Iso1*) in 70 *O. rufipogon* accessions, 99 cultivated rice accessions (*aus*, 10; *indica*, 34; *tropical japonica,* 26; *temperate japonica*, 21; *aromatic* rice, 8) and two accessions of two closely related species, *O. barthii*, *O. meridionalis*. The published sequence data for *Wx* in rice are included in the analysis as well.

The difficulty of detecting selection is often caused by the complex demographic history of a species. Genome-wide sequence data in a species would mainly reflect its demographic history. I have compared the pattern of nucleotide variation at each starch gene with published genome-wide sequence data and with a standard neutral model for detecting selection. Results show no evidence of deviations from neutrality at these six starch genes in *O. rufipogon* and no evidence of deviations from neutrality at four starch genes in *O. sativa*. Evidence of selection is observed at *Wx* in *tropical japonica* and *temperate japonica,* and at *Wx* and *SbeIIb* in *aromatic* rice.

Starch quality is one of the most important agronomic traits in rice. Starch synthase IIa (*SsIIa*) has been mapped as a gene which contributes to the starch quality variation in cultivated rice, *O. sativa*. Within the gene, three nonsysnonymous mutations in the exon 8 region were shown to affect its enzyme activity in *Escherichia coli*. In order to identify the mutation in *SsIIa* exon 8 region that is responsible for starch quality variation in rice, I have sequenced *SSIIa* exon 8 region and recorded the alkali spreading score in 57 *O. rufipogon* accessions and 151 cultivated rice accessions (*aus*, 8; *indica*, 51; *tropical japonica,* 55; *temperate japonica*, 29; *aromatic*, 8). Starch alkali spreading score is used to quantify rice endosperm starch quality and has been shown to be significantly associated with *SsIIa* enzyme activity in rice. Both a general linear model and nested clade analysis were used to detect an association between the three nonsynonymous mutations in *SSIIa* exon 8 and the alkali spreading score. In order to avoid the effect of population structure on the association analysis, both association analyses are conducted within each rice variety group. Among the previously identified nonsynonymous mutations, my results show strong evidence of association at one nonsynonymous mutation (SNP3, see Fig 2 of Chapter 2), and evidence of no association at another nonsynonymous mutation. Tests of association for the other nonsynonymous mutation are inconclusive with current samples and will require further investigation.

This dissertation reveals the relative role of evolutionary forces in shaping the variation pattern of six starch genes in *O. sativa* and its wild ancestor, *O. rufipogon*. It also reveals an association between a nonsynonymous mutation in *SSIIa* exon 8 and rice endosperm starch quality.

Acknowledgments

I would like to acknowledge the financial support from Washington University at St. Louis, the National Science Foundation plant genomics program, which made the lab work for this project possible.

Numerous people have helped me in completing this thesis. I want to take this opportunity to express my sincere thanks to every one of them.

I would like to thank all the members of my thesis committee for their support and valuable suggestions on this project. I would like to express my greatest appreciation to Dr. Barbara Schaal for her unwavering support of me. I could not have reached this point without her generous support in both science and life. I am lucky to have such a wonderful advisor who always listens, encourages, and gives the right suggestions. I am also deeply indebt to Dr. Ken Olsen for his recommendation of this fascinating subject of starch synthesis pathway and for all the valuable assistance in all these years. I thank Dr. Justin Fay for the valuable suggestions on data analysis.

I also would like to thank Dr. Peter Raven for his generous support and for helping me obtain a lot of samples of *Ambrosia*. Although I switched my thesis project from *Ambrosia* to rice, his support still means a lot to me.

This project would not have been completed in such a short period of time without the great rice collections in Schaal lab, which were made by previous Schaal lab members, Dr. Jason Londo and Dr. Yu-Chung Chiang. I thank the assistance of the Washington University Greenhouse staffs. Mike Dyer, Darlene Branson and Barb Bauwens helped me with planting, collecting seeds and fighting with rice mites. I also thank the Biology department sequencing facility staffs for regularly maintaining the sequencer, which I used a lot to collect most of my thesis data.

Numerous students in the Evolution, Ecology and Population Biology program have provided advice, emotional support and great companionship during these years. I thank past and present members of the Schaal lab, who have provided great help for me to establish in St. Louis as an international student. I particularly thank James Beck, Kuo-fang Chung, Yu-chang Chiang, Ben Torke, and Brad Oberle for their intellectual exchange and social commentary. I thank the great friendship of my cohorts, Nicole Miller and Taina Price. I thank the high school student, Charles Washington, for help at the Greenhouse and for providing me an opportunity as a mentor. I thank Ana Caicedo for providing the genome-wide sequence data of rice, which I used in my thesis.

Finally, I would like to thank my family for their strong support. My parents do not really understand what I was doing. But they always pray for me and wish the best of my thesis project. Special thanks are needed to my husband, Jun Xu, for his love and unwavering support. He always tried to help, to give some suggestions and to company me through a phone no matter how late it is. I also need to thank numerous Chinese friends here for their great companion. I specially thank my good friends, Jing-yu Zhang, Lena Wang, Hua-wen Lin and Wei Wang.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AGP, ADP-glucose pyrophosphorylase

ANOVA, analyses of variance

CNI, close-neighbor-interchange

DP, degree of polymerization

GWA, genome-wide association

GT, gelatinisation temperature

HKA test, Hudson-Kreitman-Aguadé test

MLHKA test, maximum likelihood HKA test

MITE, miniature inverted-repeat transposable element

MK test, McDonald-Kreitman test

MP tree, maximum parsimony

NJ tree, Neighbor joining

NI, neutral index

PCR, polymerase chain reaction

QTL study, quantitative trait locus study

RCSTS, randomly combined STS loci

SASS, starch alkali spreading score

SINE, short interspersed element

SNM, standard neutral model

STS, sequence tagged site

Introduction of the Dissertation

One of the fundamental questions of the study of evolution is why there is such great life diversity on Earth. Different species exhibit great morphological and functional divergence between species, much of which is thought to be adaptive. How do species adapt to different and changing environments? What is the genetic basis of adaptive divergence between species? The adaptive divergence between species is caused by selection. These questions can be addressed by examining the role of selection in shaping the pattern of genetic diversity within a species. Within species, great phenotypic variation commonly exists. Population genetic techniques can also be applied to elucidate the contribution of genetic variation to phenotypic variation within a species. The research of this dissertation applies population genetics techniques to understand the genetic basis of phenotypic divergence between species and the genetic basis of phenotypic variation within a species.

In order to determine the relative role of the two major evolutionary forces, selection and genetic drift, in shaping the genetic diversity within species or populations, a number of methods have been developed (FAY and WU 2000; FU and LI 1993; KIM and NIELSEN 2004; LI and STEPHAN 2005; SABETI *et al.* 2002; TAJIMA 1989; WONG and NIELSEN 2004). Most of these methods use a neutral equilibrium model (NE) as a null hypothesis (FAY and WU 2000; FU and LI 1993; KIM and NIELSEN 2004; LI and STEPHAN 2005; SABETI *et al.* 2002; TAJIMA 1989; WONG and NIELSEN 2004). The NE model developed from the neutral theory of evolution (KIMURA 1968), which hypothesized that vast majority of genetic variation is selectively neutral and its fate is determined by genetic drift. All of these tests for selection compare the expectation of NE model to the extant pattern of genetic diversity. The rejection of NE model suggests that selection might play major role in shaping the observed pattern genetic diversity. Failure to reject

NE model, suggests that genetic drift might play a major role in shaping the extant pattern of genetic diversity.

The adaptive divergence between species is mainly caused by directional selection. Much attention has been given to directional selection (WRIGHT *et al.* 2005). Directional selection has been identified in female reproductive proteins in mammals (SWANSON *et al.* 2001), and many genomic region across the *human* genome (SABETI *et al.* 2007) and *Zea mays* genome (WRIGHT *et al.* 2005). Detecting positive selection is a challenge. First, the ability to reject NE models depends on a number of factors, the strength of selection, the time since fixation of the beneficial mutation, and the amount of recombination between the selected and neutral sites (BRAVERMAN *et al.* 1995; PRZEWORSKI 2002; PRZEWORSKI 2003). The reasons are straightforward: mutations arise after positive selection, and recombination breaks down association between variants. The signature of a selective sweep will disappear over time due to mutations and recombination, and the signature of a selective sweep will disappear quickly if the strength of selection is weak (PRZEWORSKI 2002). Thus, many positive selection events during the history of a species can not to be detected by population genetic approaches. Only recent and strong selection events are likely to be detected (KIM and NIELSEN 2004). This challenge is the reality, and can be overcome by selecting a treatable study system. Domesticated species are ideal model systems for the study of positive selection because domesticated species have undergone recent and strong selection at numerous loci (INNAN and KIM 2004). Positive selection across many loci has been identified in domesticated species (SATO *et al.* 2001; WRIGHT *et al.* 2005; YAMASAKI *et al.* 2005; YAMASAKI *et al.* 2007).

The second challenge for detecting selection is a false negative result for neutrality tests due to the complex demographic history of many species. The null hypothesis used by most neutrality tests is NE models, which assumes random mating and constant population size. Therefore, rejection of NE models does not necessarily mean selection; it might be the violation of NE model assumptions (WRIGHT and GAUT 2005). Most species violate the assumptions of random mating and have complex demographic histories (RAMOS-ONSINS *et al.* 2008). For example, almost all domesticated species experienced a history of population size change, which includes at least a bottleneck event and population expansion event during domestication (ZEDER 2006). Therefore, the best way to overcome this challenge is to use the most likely demographic model for a species as the null hypothesis for neutrality tests (RAMOS-ONSINS *et al.* 2008). However, constructing the most likely demographic model is itself a challenge. A proposed alternative is to use genome-wide variation to reflect the complex demographic history of a species (BISWAS and AKEY 2006). This method of detecting selection compares the extant pattern of variation of a gene with genome-wide variation. In general, the better way of studying selection so far is to study the domesticated species and to use the genome wide variation pattern as the null expectation (BISWAS and AKEY 2006).

Ever since Darwin, domesticated species have been used as model species for the study of evolution. Domestication of plants or animals from their wild ancestors has typically involved rapid phenotypic evolution in response to strong directional selection (HARLAN 1992). These dramatic, human-mediated transformations provide an excellent model for studying directional selection. The advantages of a domestication model system also include a well documented and short timescale for domestication, a suite of known traits that were under intense selection during domestication and the accumulation of genetic information.

Asian rice, *Oryza sativa,* is one of the oldest domesticated species, which was domesticated in Asia at least 10000 years ago (DIAMOND 2002). The population structure of

Asian rice has been well surveyed. Asian rice is highly variable in phenotype with an estimated 120,000 varieties (KHUSH 1997). Most varieties of rice can be placed into two subspecies or races, *Oryza sativa ssp.indica* and *Oryza sativa ssp. japonica*, based on their morphological, physiological and ecological differences, such as the length of hull-hairs, seed dormancy, cold tolerance and the disintegration of endosperm starch in alkali solution (KHUSH 1997; OKA and MORISHIMA 1997). Recent studies by 169 microsatellite and 2 chloroplast sequence markers identified five cultivated rice variety groups including *aus, indica, tropical japonica, temperate japonica* and *aromatic* rice (GARRIS *et al.* 2005). Among these five variety groups, *aus* and *indica* are closely related; *tropical japonica*, *temperate japonica* and *aromatic* are also closely related. *Indica*, *tropical japonica* and *temperate japonica* are the major variety groups in rice, which are widely grown in Asia (MATHER *et al.* 2007). Due to their ecological differences, *indica* and *tropical japonica* varieties are mainly grown in the tropical or subtropical regions such as India, Southeast Asian and Southern USA; *Temperate japonica* varieties are common in temperate region such as Northeastern Asia. *Aus* varieties were known as early maturing and drought tolerant upland rice, with restricted distribution in Bangladesh and West Bengal state of India. *Aromatic* rice predominates in the Indian subcontinent (KHUSH 1997).

The domestication history of Asian rice is intensely studied. Rice was domesticated from the wild species, *O. rufipogon* (KHUSH 1997). Recent studies indicate that there were at least two domestication centers: one in South China for *japonica* rice, and the other in south and southwest of the Himalayan mountain range for *indica* rice. *Aus* rice may be a third domestication event if it is considered as an independent rice variety group (Londo *et al*., 2006)(GARRIS *et al.* 2005). These multiple domestication events provide a unique opportunity for parallel evolutionary comparisons.

Domesticated species are usually different from their wild ancestors (DIAMOND 2002). These differences have been described as a distinct suite of traits, termed the "domestication syndrome" (HARLAN *et al.* 1973). Cereal crops share a suite of similar domestication syndromes, which includes the reduction in seed shattering and dormancy, synchronization of seed maturation, decrease in culm number and branches, increase in inflorescence and seed sizes (BURGER *et al.* 2008). These domestication syndromes are the result of directional selection by humans for effective seed harvest and planting and higher grain yield and quality (BURGER *et al.* 2008).

In addition to the well known domestication history and domestication syndromes, rice has a lot of available genetic information. A recent study surveyed the variation pattern across the rice genome in five cultivated rice variety groups and their ancestor, *O. rufipogon* (CAICEDO *et al.* 2007). This study sequenced 111 sequenced tagged sites (STS) distributed across the whole genome of rice. Each STS locus is about 500 bp long and includes both coding and non-coding regions. These data suggest the complex demographic history of rice. A simple bottleneck model, which has been the dominant model for domesticated species, can not explain the pattern of nucleotide polymorphism of the STS data in rice. Complex demographic models, which include a bottleneck model that incorporates selective sweeps, and a demographic model that includes subdivision and gene flow is more consistent with the STS data. Although no clear demographic model was available for rice by the STS study, these genome-wide STS data can serve as neutral expectation for neutrality tests (RAMOS-ONSINS *et al.* 2008).

Starch quality is one of the most important agronomic traits in cereal crops, and it has been the target of selection during domestication (WHITT *et al.* 2002). Starches, which account for approximately 90% of milled rice seed’s dry weight, are a major determinant of both rice yield and quality (MOHAPATRA *et al.* 1993). The starch synthesis pathway is thus one of the most important agronomic pathways. The starch synthesis pathway is also one of the best characterized pathways in plants. Until now, over 20 genes involved in the starch synthesis pathway have been identified in cereal crops. Among these genes, six are known to play major roles in rice endosperm starch synthesis: *Shrunken2 (Sh2), Brittle2 (Bt2), Waxy (Wx), Starch synthase IIa (SsIIa), Starch branching enzyme IIb (SbeIIb)* and *Isoamylase1 (Iso1)* (JAMES *et al.* 2003).

What is the relative role of genetic drift and selection in shaping the pattern genetic diversity at these six starch genes in five rice variety groups and their wild ancestor, *O. rufipogon*? Are there different targets of selection at these six starch genes in different rice variety groups? How does the genetic variation of these six starch genes contribute to the starch quality variation among rice variety groups and *O. rufipogon*? All these questions are fundamental for the understanding of starch phenotype evolution before and during rice domestication. These questions are addressed in Chapter One.

The other important goal of population genetics is to determine the association between genetic diversity and phenotypic diversity within a species. There currently are several approaches. One approach is a quantitative trait locus (QTL) study. QTL methods typically make crosses between two or more lines that differ genetically with regard to the trait of interest (LYNCH and WALSH 1998). The crosses are then genotyped using SNPs or other markers across the whole genome, and statistical associations of the linkage disequilibrium between genotype and phenotype are identified. QTL analysis usually identifies one or several genomic regions with dozens of genes and requires further investigation to find the genes for a particular phenotype. It is widely used in domesticated species or other model species which have genome-wide genetic markers (FULTON *et al.* 1997; PERRETANT *et al.* 2000; SPECHT *et al.* 2001; TURRI *et al.* 2001).

Another approach is a genome-wide association (GWA) study (AMUNDADOTTIR *et al.* 2009). A GWA study examines the genome-wide variation for gene regions or single nucleotide polymorphisms (SNPs) associated with observable traits. This approach is widely used in humans to search for the genetic basis of disease (AMUNDADOTTIR *et al.* 2009; HAROLD *et al.* 2009; WEISS and ARKING 2009).

The final approach, a candidate gene association study, examines genetic variation across candidate genes and seeks to identify the genes and/or SNPs for particular phenotypes (NACHMAN *et al.* 2003). This approach does not need genome-wide variation across a studied genome. It requires that candidate genes for a particular phenotype have been identified. With increasing knowledge of physiology and biochemistry, more and more candidate genes are being identified.

In domesticated crops, candidate genes for important agronomic traits, especially "domesticated traits" favored by early farmers (e. g. reduction of seed shattering and dormancy, increased yield) are becoming available through biochemical and molecular genetic studies (BENTSINK *et al.* 2006; HARLAN 1975; KONISHI *et al.* 2006; LI *et al.* 2006; LI and GILL 2006; LIN *et al.* 1998). Examining the pattern of variation of these candidate genes for a signal of selection, and searching for associations between candidate genes' variation and phenotypic variation, will allow us understand the genetic basis of phenotypic diversity within species.

Population structure within samples from association studies can generate spurious associations (MARCHINI *et al.* 2004). Without knowing the population structure, it is difficult to distinguish the real association between genotype and phenotype from the false association, which was caused by population structure. To avoid this problem, there are two available approaches. One is to include information about population structure as covariate in association analysis. However, this approach requires genome-wide markers to calculate the relative kinship matrix (which reflects population structure) of the sampled materials (BRADBURY *et al.* 2007). The other approach is to perform the association analysis within each subpopulation respectively if the population structure of the study system is known (MARCHINI *et al.* 2004). The demographic history and population structure of Asian rice is well studied (see above), which makes it an ideal model for genotype phenotype association study. The genotype phenotype association analysis can be performed separately within each known population/variety group to avoid the spurious effect of population structure.

Starch quality is one of the most important agronomic traits for rice. Starch is composed of amylose and amylopectin. Amylose is a linear molecule of (1→4) linked α-D-glucopyranosyl units. Amylopectin is the highly branched component of starch. It is formed through chains of α-D-glucopyranosyl residues linked together by 1→4 linkages but with 1→6 bonds at the branch points (BULÉON *et al.* 1998). Amylopectin molecules vary in fine structure by the length of branches and are classified into two types: L-type and S-type. The L-type amylopectin differs from the S-type amylopeciton in that the former has a dramatically lower proportion of short amylopectin chains with a degree of polymerization (DP) <=10 (NAKAMURA *et al.* 2006). Both the relative ratio of amylose to amylopectin content or different amylopectin types could cause starch quality variation in rice.

Starch disintegration level in alkali (1.5% KOH) solution is a standard method to characterize rice endosperm starch. Starch disintegration level variation in alkali among rice varieties has been first reported by Warth and Darabsett (WARTH and DARABSETT 1914). Recent studies suggest that the nonsysnonymous polymorphism at *SsIIa* exon 8 might be responsible for starch quality variation in rice (GAO *et al.* 2003; UMEMOTO and AOKI 2005; UMEMOTO *et al.* 2004; UMEMOTO *et al.* 2002). However, the relationship between the nonsynonymous variation and starch disintegration variation in alkali remain unclear.

In the second chapter of this thesis, I surveyed the relationship betweeen nonsysnonymous mutations of *SsIIa* exon 8 and starch disintegration varation in alkali in *O. rufipogon* and each *O. sativa* variety group. The molecular evolution of the nonsysnonymous mutations at *SsIIa* exon 8 was also surveyed in *O. rufipogon* and *O. sativa*. The primary hypothesis is that starch quality evolved during rice domesitication. In order to test this hypothesis, the phenotypic difference of starch disintegration level in alkai among *O. rufipogon* and 5 cultivated rice variety groups were surveyed.

The chapters that follow are intended as case studies of population genetics to understand the genetic basis of starch quality variation present between cultivated and wild rice, within cultivated rice or within wild rice. These chapters also shed some light on the issues and challenges that are involved in studying the evolution of functional genes in domesticated species and in studying genotype-phenotype association within species.

References:

AMUNDADOTTIR, L., P. KRAFT, R. Z. STOLZENBERG-SOLOMON, C. S. FUCHS, G. M. PETERSEN *et al.*, 2009 Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nature Genetics **41:** 986-U947.

BENTSINK, L., J. JOWETT, C. J. HANHART and M. KOORNNEEF, 2006 Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **103:** 17042-17047.

BISWAS, S., and J. M. AKEY, 2006 Genomic insights into positive selection. Trends in Genetics **22:** 437-446.

BRADBURY, P. J., Z. ZHANG, D. E. KROON, T. M. CASSTEVENS, Y. RAMDOSS *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The Hitchhiking Effect on the Site Frequency-Spectrum of DNA Polymorphisms. Genetics **140:** 783-796.

BULÉON, A., P. COLONNA, V. PLANCHOT and S. BALL, 1998 Starch granules: structure and biosynthesis. International Journal of Biological Macromolecules **23:** 85-112.

BURGER, J. C., M. A. CHAPMAN and J. M. BURKE, 2008 Molecular insights into the evolution of crop plants. American Journal of Botany **95:** 113-122.

CAICEDO, A., S. WILLIAMSON, R. D. HERNANDEZ, A. BOYKO, A. FLEDEL-ALON *et al.*, 2007 Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLoS Genetics **3:** e163.

DIAMOND, J., 2002 Evolution, consequences and future of plant and animal domestication. Nature **418:** 700-707.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

FU, Y. X., and W. H. LI, 1993 Statistical Tests of Neutrality of Mutations. Genetics **133:** 693-709.

FULTON, T. M., T. BECKBUNN, D. EMMATTY, Y. ESHED, J. LOPEZ *et al.*, 1997 QTL analysis of an advanced backcross of Lycopersicon peruvianum to the cultivated tomato and comparisons with QTLs found in other wild species. Theoretical and Applied Genetics **95:** 881-894.

GAO, Z. Y., D. L. ZENG, X. CUI, Y. H. ZHOU, M. YAN *et al.*, 2003 Map-based cloning of the ALK gene, which controls the gelatinization temperature of rice. Science in China Series C-Life Sciences **46:** 661-668.

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.

HARLAN, J. R., 1992 *Crops and Man*. American Society of Agronomy, Madison, WI.

HARLAN, J. R., J. M. J. DEWET and E. G. PRICE, 1973 Comparative evolution of cereals. Evolution; International Journal of Organic Evolution **27:** 311-325.

HAROLD, D., R. ABRAHAM, P. HOLLINGWORTH, R. SIMS, A. GERRISH *et al.*, 2009 Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nature Genetics **41:** 1088-U1061.

INNAN, H., and Y. KIM, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America **101:** 10667-10672.

JAMES, M. G., K. DENYER and A. M. MYERS, 2003 Starch synthesis in the cereal endosperm. Current Opinion in Plant Biology **6:** 215-222.

KHUSH, G. S., 1997 Origin, dispersal, cultivation and variation of rice. Plant Molecular Biology **V35:** 25-34.

KIM, Y., and R. NIELSEN, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167:** 1513-1524.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624-626.

KONISHI, S., T. IZAWA, S. Y. LIN, K. EBANA, Y. FUKUTA *et al.*, 2006 An SNP caused loss of seed shattering during rice domestication. Science **312:** 1392-1396.

LI, C. B., A. L. ZHOU and T. SANG, 2006 Rice domestication by reducing shattering. Science **311:** 1936-1939.

LI, H. P., and W. STEPHAN, 2005 Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics **171:** 377-384.

LI, W., and B. GILL, 2006 Multiple genetic pathways for seed shattering in the grasses. Functional & Integrative Genomics **6:** 300-309.

LIN, S. Y., T. SASAKI and M. YANO, 1998 Mapping quantitative trait loci controlling seed dormancy and heading date in rice, Oryza sativa L., using backcross inbred lines. Theoretical and Applied Genetics **96:** 997-1003.

LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Sunderland, MA.

MARCHINI, J., L. R. CARDON, M. S. PHILLIPS and P. DONNELLY, 2004 The effects of human population structure on large genetic association studies. Nature Genetics **36:** 512-517.

MATHER, K. A., A. L. CAICEDO, N. R. POLATO, K. M. OLSEN, S. MCCOUCH *et al.*, 2007 The extent of linkage disequilibrium in rice (Oryza sativa L.). Genetics **177:** 2223-2232.

MOHAPATRA, P. K., R. PATEL and S. K. SAHU, 1993 Time of flowering affects grain quality and spikelet partitioning within the rice panicle. Australian Journal of Plant Physiology **20:** 231-241.

NACHMAN, M. W., H. E. HOEKSTRA and S. L. D'AGOSTINO, 2003 The genetic basis of adaptive melanism in pocket mice. Proceedings of the National Academy of Sciences of the United States of America **100:** 5268-5273.

NAKAMURA, Y., A. SATO and B. O. JULIANO, 2006 Short-chain-length distribution in debranched rice starches differing in gelatinization temperature or cooked rice hardness. Starch-Starke **58:** 155-160.

OKA, H., and H. MORISHIMA, 1997 *Wild and cultivated rice*. Genetics, Nobunkyo, Tokyo.

PERRETANT, M. R., T. CADALEN, G. CHARMET, P. SOURDILLE, P. NICOLAS *et al.*, 2000 QTL analysis of bread-making quality in wheat using a doubled haploid population. Theoretical and Applied Genetics **100:** 1167-1175.

PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179-1189.

PRZEWORSKI, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164:** 1667-1676.

RAMOS-ONSINS, S. E., E. PUERMA, D. BALANA-ALCAIDE, D. SALGUERO and M. AGUADE, 2008 Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in Arabidopsis thaliana. Molecular Ecology **17:** 1211-1223.

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. P. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832-837.

SABETI, P. C., P. VARILLY, B. FRY, J. LOHMUELLER, E. HOSTETTER *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature **449:** 913-U912.

SATO, Y., Y. FUKUDA and H. Y. HIRANO, 2001 Mutations that cause amino acid substitutions at the invariant positions in homeodomain of OSH3KNOX protein suggest artificial selection during rice domestication. Genes & Genetic Systems **76:** 381-392.

SPECHT, J. E., K. CHASE, M. MACRANDER, G. L. GRAEF, J. CHUNG *et al.*, 2001 Soybean response to water: A QTL analysis of drought tolerance. Crop Science **41:** 493-509.

SWANSON, W. J., Z. H. ZHANG, M. F. WOLFNER and C. F. AQUADRO, 2001 Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proceedings of the National Academy of Sciences of the United States of America **98:** 2509-2514.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

TURRI, M. G., S. R. DATTA, J. DEFRIES, N. D. HENDERSON and J. FLINT, 2001 QTL analysis identifies multiple behavioral dimensions in ethological tests of anxiety in laboratory mice. Current Biology **11:** 725-734.

UMEMOTO, T., and N. AOKI, 2005 Single-nucleotide polymorphisms in rice starch synthase IIa that alter starch gelatinisation and starch association of the enzyme. Functional Plant Biology **32:** 763-768.

UMEMOTO, T., N. AOKI, H. X. LIN, Y. NAKAMURA, N. INOUCHI *et al.*, 2004 Natural variation in rice starch synthase IIa affects enzyme and starch properties. Functional Plant Biology **31:** 671-684.

UMEMOTO, T., M. YANO, H. SATOH, A. SHOMURA and Y. NAKAMURA, 2002 Mapping of a gene responsible for the difference in amylopectin structure between japonica-type and indica-type rice varieties. Theoretical and Applied Genetics **104:** 1-8.

WARTH, F., and D. DARABSETT, 1914 disintegration of rice grains by means of alkali. Bulletin of Agricultural Research Institute **38:** 1-9.

WEISS, L. A., and D. E. ARKING, 2009 A genome-wide linkage and association scan reveals novel loci for autism. Nature **461:** 802-U862.

WHITT, S. R., L. M. WILSON, M. I. TENAILLON, B. S. GAUT and E. S. BUCKLER, 2002 Genetic diversity and selection in the maize starch pathway. Proceedings of the National Academy of Sciences of the United States of America **99:** 12959-12962.

WONG, W. S. W., and R. NIELSEN, 2004 Detecting selection in noncoding regions of nucleotide sequences. Genetics **167:** 949-958.

WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The effects of artificial selection of the maize genome. Science **308:** 1310-1314.

WRIGHT, S. I., and B. S. GAUT, 2005 Molecular Population Genetics and the Search for Adaptive Evolution in Plants. Molecular Biology and Evolution **22:** 506-519.

YAMASAKI, M., M. I. TENAILLON, I. VROH BI, S. G. SCHROEDER, H. SANCHEZ-VILLEDA *et al.*, 2005 A Large-Scale Screen for Artificial Selection in Maize Identifies Candidate Agronomic Loci for Domestication and Crop Improvement. Plant Cell **17:** 2859-2872.

YAMASAKI, M., S. I. WRIGHT and M. D. MCMULLEN, 2007 Genomic Screening for Artificial Selection during Domestication and Improvement in Maize. Annals of Botany **100:** 967 - 973.

ZEDER, M. A., 2006 Central questions in the domestication of plants and animals. Evolutionary Anthropology **15:** 105-117.

Chapter 1

Evolution of Endosperm Starch Synthesis Pathway Genes in *Oryza Sativa* and its wild ancestor *O. rufipogon*

Introduction

A fundamental goal of population genetics is to quantify the roles of various forces of evolution, such as selection and drift, in shaping patterns of genetic variation (CLEGG 1997). Numerous studies have been conducted to understand their relative roles in evolution. However, most of the studies focus on individual genes or multiple neutral markers. There is a limited understanding about the evolution of the genetic basis for both complex traits and metabolic pathways where variation at one locus potentially affects and constrains the evolution of other genes connected within a network. In order to fully understand the evolution of metabolic pathways, we need to understand how evolutionary forces act on multiple interacting genes that are components of molecular pathways or networks (FRASER *et al.* 2002). Furthermore, metabolic pathways are true functional units in cellular metabolic systems. Most cellular processes and organismal phenotypes are determined by metabolic pathways and their regulatory cascades. Therefore, understanding fundamental questions about phenotypic diversification and adaptation requires answers to the following questions: What is the evolutionary history of genes in a metabolic pathway? Do the genes have similar evolutionary histories? How do genetic drift and selection shape the variation pattern of the component genes of a metabolic pathway? Finally, how do molecular changes within the genes of metabolic pathways contribute to the phenotypic changes? During the last decade, functional genomic studies have accumulated a vast amount of information about molecular interactions in a cellular context (COLLADO-VIDES and HOFESTADT 2002). Specific databases of metabolic pathways have been constructed for different organisms and are available via the Internet (e.g., for rice, www.gramene.org/pathway). These recent advances have identified genes and their biochemical function in the biochemical pathways, and now allow for evolutionary study of functional genes in the context of their metabolic pathways.

Domestication of crop plants from their wild ancestors has typically involved rapid phenotypic evolution in response to strong directional selection (HARLAN 1992). There is usually an increase of yield, nutrition and the reduction of toxins, which involves many metabolic and regulatory changes (DIAMOND 2002). Biologists including Darwin (DARWIN 1859) have asserted that these dramatic, human-mediated transformations provide an excellent model for studying phenotypic evolution. Advantages of a domestication model system include a well documented and short timescale as well as a suite of known traits that were under intense selection during domestication. These factors, combined with large accumulation of genetic information in many domesticated plants and animals, have made them unique models for studying the genetic and phenotypic consequences of strong directional selection.

Asian rice, *Oryza sativa,* is especially attractive for studying both genetic and phenotypic evolution owing to its complete sequenced genome, as well as its unique domestication history with at least two independent domestications (LONDO *et al.* 2006), which provides an opportunity for independent comparison. Moreover, after domestication, there is subsequent selective improvement beginning with traditional rice varieties (landraces) preserved by local indigenous farmers to the commercially bred ''elite'' cultivars (modern cultivars). This range of populations provides an opportunity to study selection on three different levels: the natural selection in *Oryza rufipogon,* rice's wild ancestor, artificial selection during domestication and selective improvement post domestication.

Detection of selection at the molecular level has been proven to be difficult in most species, mainly due to their complex demographic history (WRIGHT and GAUT 2005). Most available methods detect the signal of selection by comparing the variation level or pattern of studied loci with those expected by the standard neutral model (SNM). Deviations from the SNM at particular loci might reflect the species demographic history and not the action of positive selection. The SNM assumes a simple demographic history (e.g. constant population size), while most species, especially domesticated species, have experienced much more complicated demographic processes (e.g. bottleneck and/or population expansion). An alternative method is by using a most likely demographic model as the null hypothesis, a demographic model which reflects the most likely demographic scenarios for the studied species. Rice offers an additional advantage*,* which allows for discriminating between the locus-specific process of selection and genome-wide process of demography by using the genome-wide variation data as the reference of neutral evolution. Recent genome-wide variation survey for both *O. rufipogon* and *O. sativa* proved the complexity of the demographic history in *O. sativa* (CAICEDO *et al.* 2007). Although the most likely demographic model for *O. sativa* and *O. rufipogon* is still lacking, this genome-wide variation, which will mainly reflect the demographic histories in *O. rufipogon* and *O. sativa*, could be used as alternative neutral reference. In this study, we have used the genome-wide variation pattern (111 sequence tagged sites across rice genome) as neutral reference to detect positive selection before rice domestication and during rice domestication (CAICEDO *et al.* 2007). In order to detect the signal of positive selection before and during domestication, I have contrasted the pattern of diversity detected in the samples of *O. rufipogon* or *O. sativa* against those expected under the standard neutral model (SNM) as a first step. Then I compare particular summaries of statistics with the distributions of the statistics obtained from 111 genome-wide sequence tagged site (STS) loci.

The endosperm starch biosynthesis pathway is one of the best characterized metabolic pathways in plants (see “Study System” below for more details). Starch, which accounts for about 90% of dried milled rice seed’s weight, is a major determinant of both rice yield and quality (ZHANG *et al.* 2008). Strong selection by humans on starch traits is thus expected during domestication and/or subsequent improvement, and it is likely via selection on starch synthesis pathway genes (WHITT *et al.* 2002). In this study, I focus on six major starch genes which play critical roles in starch synthesis of rice endosperm. The following questions are addressed: 1) what is the pattern of nucleotide variation in these six starch genes in *O. sativa* and its wild ancestor species, *O. rufipogon*? 2) What is the relative level of nucleotide variation in these six starch genes in traditional landraces and modern cultivars? 3) What is the relationship of nucleotide variation and position of genes in the endosperm starch synthesis pathway? 4) What is the relative role of selection and genetic drift in shaping the variation pattern of these six starch pathway genes before rice domestication and during rice domestication? 5). How does the variation in the underlying genes of this pathway contribute to the adaptive shift of starch traits?

Study system

Asian rice is one of the world's most important food crops and feeds about one-third of the world's population. Rice is highly variable in phenotype with an estimated 120,000 varieties (KHUSH 1997). Most varieties of rice are placed into two subspecies or races, *Oryza sativa ssp.indica* and *Oryza sativa ssp. japonica*, based on morphological, physiological and ecological differences. These include, for example, the length of hull-hairs, the degree of seed dormancy, seedling cold tolerance, and disintegration of endorsperm starch granules in alkali solution (KHUSH 1997). Recent study by 169 nuclear microsatellite (SSR) markers and two chloroplast loci identified five rice variety groups including *aus, indica, tropical japonica, temperate japonica* and *aromatic* rice (GARRIS *et al.* 2005). Among these five variety groups, a close evolutionary relationship between *aus* and *indica*, between *tropical japonica*, *temperate japonica* and *aromatic* was supported by both chloroplast and nuclear markers (GARRIS *et al.* 2005).

Asian rice is one of the oldest domesticated species, domesticated in Asia at least 10000 years ago (CAICEDO *et al.* 2007). Previous studies have indicated that *O. sativa* is derived from its wild ancestor *O. rufipogon* (KHUSH 1997). Recent studies have indicated that there were at least two domestication centers: one in South China for *japonica* rice, another in the south and southwest of the Himalayan mountain range for *indica* rice. *Aus* rice may be a third domestication event if it is considered as an independent rice variety group (Londo *et al.*, 2006). After domestication, there has been a continuing selective improvement of the crop from traditional rice varieties preserved by local indigenous farmers to the commercially bred ‘‘elite’’ cultivars (KHUSH 1997).

The starch synthesis pathway is an ideal system for examining the evolution of biochemical pathways. Starch is the major component of yield in the world's most important cereal crop plants. It is composed of amylose and amylopectin. Amylose is a linear molecule of (1→4) linked α-D-glucopyranosyl units. Amylopectin is the highly branched component of starch. It is formed through chains of α-D-glucopyranosyl residues linked together by 1→4 linkages but with 1→6 bonds at the branch points (BULÉON *et al.* 1998).

Over 20 genes involved in the starch synthesis pathway have been identified so far (MYERS *et al.* 2000). Six of them are known to play major roles in rice endosperm starch synthesis: *Shrunken2 (Sh2), Brittle2 (Bt2), Waxy (Wx), Starch synthase IIa (SsIIa), Starch branching enzyme IIb (SbeIIb)* and *Isoamylase1 (Iso1)*. The position of these starch genes in the starch synthesis pathway is shown in Figure 1.1. These are the genes that were studied in the present study. All of these genes except *Wx* and *Iso1* are exclusively expressed in rice endosperm. *Sh2* and *Bt2* respectively encode the large and small subunits of ADP-glucose pyrophosphorylase (AGP), which converts glucose-1-phosphase into ADP-glucose. Mutants at either *Sh2* or *Bt2* locus give rise to shrunken, brittle seeds, greatly reducing starch level in endosperm (DKINSON and PREISS 1969; SMIDANSKY *et al.* 2003; TSAI and NELSON 1966). *Wx* and *SsIIa* encode starch synthases, which elongate linear chains by formation of 1→4 linkages. However, *Wx* is solely responsible for amylose production, and mutants at this locus cause drastically reduced amylose synthesis (SANO 1984). *SsIIa* is responsible for the elongation of short chains with the degree of polymerization (DP) of 10 or less that leads to intermediate chains for amylopectin. Previous studies suggested that *SsIIa* may account for the starch quality difference between *japonica* and *indica* rice (Umemoto, 2002; Waters, 2006; Nakamura, 2005). *SbeIIb* generates α-1,6 linkages by cleaving α-1,4 bonds and transferring the released reducing ends to C6 hydroxyls. Mutants at this locus in maize and rice have an apparent increase the relative portion of amylose to amylopectin (KIM *et al.* 1998; MIZUNO *et al.* 1993; TANAKA *et al.* 2004). *Iso1* hydrolyzes α-1,6 bonds, and its mutants in maize and rice exhibit the accumulation of phytoglycogen and reduced starch content (James, 1995; Pan, 1984; Nakamura, 1996).

Materials and methods

Plant materials

Both cultivated rice, *O. sativa*, and its wild progenitor, *O. rufipogon*, were sampled for population genetic study (Table 1.1). Two closely related *Oryz*a species, *O. barthii* and *O. meridionalis* were sampled as outgroup species. The samples of *O. sativa* and *O. rufipogon* were chosen to represent the diversity found within the species. The *O. sativa* samples include representatives of five variety groups identified by a previous study (GARRIS *et al.* 2005): 34 *indica* (19 landraces, 15 modern cultivars), 26 *tropical japonica* (13 landraces, 13 modern cultivars), 21 *temperate japonica* (19 landrace, 15 modern cultivar), 10 *aus* and 8 *aromatic* accessions. The *O. rufipogon* samples include 70 accessions representing its whole geographic distribution range except Australia. Most of its samples are from its geographic diversity centers such as India, Thailand and China. Most *O. rufipogon* samples from China were collected in the field by previous Schaal lab member Yuchung Chiang. All other samples of *O. rufipogon* and *O. sativa* and outgroup species were ordered from International Rice Research Institute (IRRI) (Manila, Philippines). Detailed information for these samples is listed in Table 1.1. IRRI is the largest nonprofit agricultural research center in Asia, which provides free rice materials for research. In order to include additional *O. rufipogon* samples from China (one of *O. rufipogon* diversity centers), field collection for more *O. rufipogon* samples were performed. However, only the *O. rufipogon* leaf materials were collected in the field due to political reason and the difficulty of getting the seeds.

DNA extraction, polymerase chain reaction (PCR) and sequencing

Five genes (*Sh2, Bt2, Iso1, SbeIIb and SsIIa*) were sequenced in all samples. The gene *Wx* was not sequenced here because its sequences in most of these samples have been published by previous study (OLSEN *et al.* 2006). These published sequences are enough for most of the analyses and were provided in genbank (http://www.ncbi.nlm.nih.gov/).

DNA was extracted from dried leaves through CTAB method with minor modifications (DOYLE and DOYLE 1990). Except for the *O. rufipogon* leaf materials collected from the field, samples from IRRI were grown for leaf materials in the greenhouse at Washington University in St. Louis. Samples of *O. rufipogon* from IRRI were self-fertilized in the greenhouse for two generations to decrease the degree of heterozygosity.

Primers were designed by the software Primer3 (http://frodo.wi.mit.edu/primer3/) from the Nipponbare genomic sequence available from Gramene (http://www.gramene.org/). Primers were designed to amplify around 1 kilobase (FULTON *et al.*) with approximately 100 base pair (bp) overlap between neighboring amplified regions for each gene. PCRs were conducted in a thermal cycler TX2 or PTC-100. The PCR solution is 20 µl, which includes 1X Tag buffer, 2mM dNTP mix, 1 µM primers, 1 u / 20 Taq polymerase, 2.5 mM $MgCl_{2,}$ 1 µg tempelate DNA and sterile deionized water. PCRs were conducted under the following conditions: 95°C for 5 minutes; 40 cycles of 95°C for 50 seconds, 52-60 °C for 1 minute, and 72 °C for 2.5 minutes; 10 minutes of extension at 72 °C. The annealing temperature for PCRs differs by primers. The primers and the annealing temperature for PCRs are listed in Table 1.2a. PCR products were cleaned using Exo1-SAP commercial kits, then cycle-sequenced using BigDye Terminator chemistry (Applied Biosystems) and analyzed on an ABI 3130 capillary sequencer (Applied Biosystems). The primers for the five newly sequenced starch genes (excluding *Wx*) are listed in Table 1.2b.

Data analyses

Sequences were aligned and manually edited with the software Biolign version 4.0.6.2 (http://en.bio-soft.net/dna/BioLign.html). The published sequences for the *Wx* gene were obtained from Olsen *et al.* (OLSEN *et al.* 2006). The sequences of Nipponbare were from Genbank and included in the analyses for *temperate japonica* (http://www.ncbi.nlm.nih.gov/Genbank/). Determination of exons was based on previous annotation with known protein information (Gramene database at http://www.gramene.org/).

Cultivated rice, *O. sativa,* is a predominantly selfing species, and no heterozygous SNPs were found in our samples. However, the wild species, *O. rufipogon,* is a predominantly outcrossing species. Some samples from IRRI and almost all the samples from the field showed multiple heterozygous SNPs in the sequenced starch genes. The haplotypes of these heterozygous samples were determined via the Excoffier-Laval-Balding algorithm in Arlequin version 3.11 (EXCOFFIER *et al.* 2005). The algorithm was run with default parameters, except burnin steps (500000) and sampling interval (500).

Level of nucleotide variation

Only samples with less than 50 bp missing data were included in analysis. The missing data are due to the heterozygosity of indels in some samples or other various reasons during the sequencing process. Therefore, slightly different numbers of samples were used for different studied starch genes. Statistics for levels of variation (number of polymorphic synonymous, nonsynonymous and silent sites; average pairwise nucleotide diversity, θ_π; average number of segregating sites, θ_W) (TAJIMA 1983; WATTERSON 1975) were performed in DnaSP v5.0 (LIBRADO and ROZAS 2009). Only silent sites were used for estimation of θ_π and $\theta_{W.}$ Differences in the level of nucleotide variation between landrace and modern cultivars were tested by Wilconxon signed-rank tests based on five genes (no *Wx*).

Because transposable elements have been observed in *SsIIa* snd *Wx* (see Results), sliding window analyses were conducted across *SsIIa* and *Wx* in *O. rufipogon* to examine the contribution of transposable elements to the level of diversity. The window size is 100 bp, and the step size is 20 bp.

Association between nucleotide variation and position in the metabolic pathway

In order to search for an association between either the level of variation (as measured by θ_π and θ_w) or pattern of variation (as measured by Tajima's D and Fay and Wu's H) and the position within the metabolic pathway, Kendall's rank correlation tests (SOKAL and ROHLF 1995) were performed in *O. rufipogon*.

Population Recombination Rate

The population recombination parameter ρ ($4N_er$, where Ne is the effective population size, r is the recombination rate per site per generation) was estimated for each starch gene by a composite-likelihood method (HUDSON 2001). Nonparametric permutation and the maximum composite-likelihood tests were performed for each ρ estimate to test for evidence of recombination. The minimum number of recombination events was estimated for each starch gene (HUDSON and KAPLAN 1985). All these parameters were estimated in LDhat version 2.1 (MCVEAN *et al.* 2002).

***McDonald-Kreitman* tests**

The McDonald-Kreitman (MK) test (MCDONALD and KREITMAN 1991) is a test of selection that compares the ratio of nonsynonymous to synonymous variation within and between species. The MK test was performed for each starch gene in *O. rufipogon*. Fisher's exact tests were used to determine statistical significance. Neutrality indices (NI) were calculated as R_pS_f/R_fS_p (RAND and KANN 1996), where R and S refer to counts of nonsynonymous and synonymous SNPs, and P and F refer to polymorphic and fixed sites, respectively. An excess of fixed nonsynonymous SNPs relative to synonymous SNPs will lead to values of NI lower than 1, which is often considered as evidence of positive selection, while an excess of polymorphic nonsynonymous SNPs may suggest purifying selection. The counts of fixed and polymorphic SNPs were used to estimate the population selection parameter, gamma (2NS), in the MKPRF software (BUSTAMANTE *et al.* 2002), where N is the effective population size and S is the selection coefficient. The estimation was performed for each gene separately. MKPRF was run with default parameters, except that both burnin and sampling were extended to 5000 steps.

Hudson-Kreitman-Aguadé tests

The multi locus Hudson-Kreitman-Aguadé test (HKA test) (HUDSON *et al.* 1987) has been proven as a robust test to find genes affected by selection by either simulation or empirical studies (INNAN and KIM 2004; MOORE and PURUGGANAN 2003). It assesses the ratio of polymorphism within species to the divergence between species, and compares this ratio among multiple loci. Loci that are significantly different from the known neutral loci are considered to be under selection. HKA tests were performed in the program HKA with 10000 simulations (The program HKA is available at website http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/HKA/HKA_Documentation.htm). Chi-square tests were used to determine statistical significance. Statistical significance occurs when the chi-square distribution probability or proportion of runs is lower than 2.5%.

HKA tests were performed in 111 published genome-wide STS loci to determine if the STS data could serve as a neutral reference for the starch genes (CAICEDO *et al.* 2007). It showed significant distribution of chi-square probability in all the rice variety groups and wild rice (see Table 1.3a), which suggests the existence of non-neutral evolution at some STS loci. Therefore, the STS loci are not good neutral reference for HKA test (CAICEDO *et al.* 2007). HKA tests were then conducted for five studied starch genes alone (no *Wx*). *Wx* was excluded because of its different sampling (see above). In order to include *Wx* in the HKA analysis, we matched the sampling of *Wx* by sampling the accessions with the similar geographic origins for the other five starch genes.

Since the HKA test showed significant result in *tropical japonica* rice, which suggests selection in some loci (Table 1.3b), maximum likelihood HKA (MLHKA) tests were used to determine which genes are under selection. MLHKA is a modified version of the HKA test. It allows for explicit tests of selection at individual loci by comparing the neutral model and the model with hypothesized selection at certain loci (WRIGHT and CHARLESWORTH 2004). The comparisons were performed by likelihood ratio test using a chi-square test to determine statistical significance. The MLHKA tests were performed in the program MLHKA with chain lengths of 100,000. The program of MLHKA is available at http://www.yorku.ca/stephenw/StephenI.Wright/Programs.html.

Comparing the observed data with a standard neutral model (SNM)

In order to compare the pattern of diversity of the starch genes with that expected by standard neutral model, Tajima's D (TAJIMA 1989) and Fay and Wu's H (FAY and WU 2000) tests of neutrality were performed in DNAsp version 4.50 (FAY and WU 2000; ROZAS *et al.* 2003; TAJIMA 1989). The species *O. meridionalis* served as the outgroup species for calculating Fay and Wu's H. Significant deviations from a standard neutral model were determined by coalescent simulation based on 5% significance level.

Comparing the observed data set with genome-wide data

Although the HKA test has been proved to be a robust test, it only utilizes the information of nucleotide polymorphism within species and divergence between species, not the information of allele frequency spectrum. Therefore, I also compared the allele frequency spectrum estimated by Tajima's D and Fay and Wu's H for these starch genes with 111 genome-wide STS loci. In order to compare the starch genes with 111 STS loci, I matched the sampling of 111 STS loci by using the accessions with similar geographic origin for the other five studied genes. The gene *Wx* has the same sampling as do the 111 STS loci. The Tajima's D and Fay and Wu's H value for the five starch genes were recalculated after resampling.

Within the 111 STS loci, six loci have no outgroup sequences to calculate Fay and Wu's H value. And among the remaining 105 loci, 76.2% to 94.3% of STS loci have less than 3 polymorphic nucleotide sites in rice variety groups, which means large range of error for Tajima's D and Fay and Wu's H statistics due to small number of polymorphic sites in each STS locus (see Table 1.4 for the summary of polymorphism in STS data). To overcome this problem, I randomly combined five STS loci together for 1000 replicates, which makes 1000 randomly combined STS loci (RCSTS). Tajima's D and Fay and Wu's H values for each RCSTS locus were calculated.

Coalescent simulations with 10,000 replicates were conducted for each RCSTS locus to determine whether the RCSTS locus deviates from SNM or not. Coalescent simulation simulates the samples of DNA sequences with many replicates (usually 10000 replicates) under certain evolutionary scenarios, which is the SNM in our case (INNAN *et al.* 2005; KINGMAN 1982). It provides the distribution of particular statistics of the simulated samples under SNM, which is Tajima's D and Fay and Wu's H here. Tajima's D and Fay and Wu's H value of each RCSTS locus were then compared respectively with the Tajima's D and Fay and Wu's H distribution of simulated samples under SNM to determine if the diversity pattern of RCSTS locus deviates from those simulated under SNM.

Tajima's D and Fay and Wu's H values of each studied starch gene were then compared with those of RCSTS loci. The gene with Tajima's D or Fay and Wu's H values lower than 2.5% of the RCSTS loci or greater than 97.5% of the RCSTS loci is considered to be significantly different from RCSTS loci. Bonferroni correction was used to correct the significant level for multiple tests.

The above analyses were conducted by the scripts written in Perl. In order to make the comparison meaningful, I equated my sampling with that for STS loci using the accession from the similar geographic region (See above). Samples of *aus* and *aromatic* rice for the studied starch genes completely include the samples for STS loci. The samples of other rice groups for starch genes only partially overlap with the samples for STS loci.

Neighbor-Joining and maximum parsimony trees

Neighbor joining (NJ) and maximum parsimony (MP) trees were constructed for five studied genes jointly (*Wx* was excluded because of its different sampling; see above) in Mega version 4 (TAMURA *et al.* 2007). Missing data and gaps were deleted. Bootstrap tests with 1000 replicates were performed to determine statistical support of the trees. For the NJ tree, I included both transition and transversion substitutions and set the nucleotide substitution model as maximum composite likelihood model, with the pattern among lineages as same, and the rates among sites as uniform. For the MP tree, I set the search method as close-neighbor-interchange (CNI) with the search level at 1. The initial trees for CNI search were 10 random trees.

Haplotype networks

Haplotype networks were constructed by median joining method in the software NETWORK version 4.5.1.0 (BANDELT *et al.* 1999). The program was run under default parameters. Haplotype networks were constructed for *O. sativa* to examine the relationship between rice variety groups. Haplotype networks were also constructed for both *O. rufipogon* and *O. sativa* to identify the relationship between cultivated rice and its ancestor species. The haplotype networks of *O. sativa* were constructed for each starch gene. Wilcoxon nonparametric signed rank tests were used to quantify the haplotype frequency difference between rice varieties groups. Because of recombination and/or recurrent mutations (homoplasy) in *O. rufipogon*, the haplotype network of each starch gene for both cultivated and wild rice is not readable. In order to make these haplotype networks readable, each gene was divided into several regions. The separated pieces are listed in Table 1.5. Two rules were followed to divide the studied genes into pieces: 1). The genes were divided into pieces at the site of missing data or nonsequenced region, which is at least 100 bp. 2) The gene was further divided to separate the recombined loci until the haplotype network had little or no homoplasy. The recombination breakpoints were detected in DnaSP v5.0 (LIBRADO and ROZAS 2009)

Results

Level of diversity

Five genes which code for the components of the starch synthesis pathway were sequenced in 70 wild rice accessions of *O. rufipogon* and 99 cultivated rice accessions. The closely related *Oryza* species *O. meridionalis* was used as an outgroup for Fay and Wu's H (FAY and WU 2000) statistics because *O. barthii* was grouped with the wild and cultivated rice based on NJ and MP tree of the starch genes (Figure 1.3 and 1.4). The length of the region sequenced for each gene varied from 3.12 to 4.63 kb, with a total length of 19.24 kb sequenced for each individual in this study. The location and size of the sequenced region for each gene is shown in Figure 1.2.

Estimates of nucleotide polymorphism of the starch genes in *O. rufipogon* are presented in Table 1.6. The number of nonsynonymous mutations ranges from 1 to 8. The number of synonymous mutations is higher than nonsynonymous mutations for all the loci except *Iso1* and ranges from 5 to 11. Among all the starch genes, the lowest level of variation was found at *Sh2* (θ_π=0.00146 and θ_w=0.00334) and the highest at *Wx* (θ_π=0.02127 and θ_w=0.01728). The level of diversity of the starch genes in *O. rufipogon* is not significantly different from that of RCSTS loci except for *Wx* (Figure 1.5), which has significantly higher variation.

Estimates of nucleotide polymorphism of starch genes in cultivated rice are presented in Table 1.7. Both landrace and modern cultivar samples were used for estimates of variation in *indica*, *tropical japonica* and *temperate japonica* in order to compare with *aromatic* and *aus* rice since *aromatic* and *aus* rice have both landrace and modern cultivar samples. No nonsynonymous mutations were discovered in the sequenced regions of *SbeIIb* and *Sh2*. Other genes had 1 to 3 nonsynonymous mutations in the sequenced regions, which is slightly lower than the number of synonymous mutations (1-6). The lowest θ_π and θ_w values were found in *tropical japonica* at *Wx* (θ_π=0.00018, θ_w= 0.00033) and the highest in *indica* at *SsIIa* (θ_π=0.00544, θ_w= 0.00529).

Among all the rice variety groups, the level of silent polymorphism based on θ_w is the highest in *indica*, and the lowest in *temperate japonica* at most starch genes. However, the genes *SbeIIb* and *Bt2* showed the highest diversity in *aromatic* and *tropical japonica* respectively, and the genes *Wx*, *Sh2* and *Bt2* had the lowest diversity in *tropical japonica*, *aromatic* and *aus* respectively. Among the six starch genes, the level of silent polymorphism based on θ_w is the highest at *SsIIa* in all cultivated rice groups except *indica*, which had the highest level of silent polymorphism at *Wx*. The lowest θ_w value was observed at *Bt2* for *aus*, *SbeIIb* for *indica*, *Wx* for *tropical japonica*, *Iso1* for *temperate japonica* and *Sh2* for *aromatic* rice (Table 1.7).

The level of diversity in *O. rufipogon* is significantly higher than that of any rice variety group at any of the starch genes (Figure 1.6) (Wilcoxon signed-rank test, P<0.05). Compared to *O. rufipogon*, the reduction of diversity in *tropical japonica* at *Wx* is extreme, only about 2% of that of *O. rufipogon*. In contrast, for all the other rice variety groups at any of the starch genes, it is 12-74% of that of *O. rufipogon*.

Estimates of the levels of nucleotide polymorphism in landraces and modern cultivars of three major rice variety groups are shown in Table 1.8. The genetic diversity between landrace and modern cultivars was compared for five of the six starch genes (*Wx* was excluded because of sampling; see Methods). In most cases, modern cultivars showed slightly higher number of synonymous and/or nonsynonymous mutations than did landrace rice. According to Wilcoxon signed rank tests, genetic diversity of silent sites estimated by θ_π based on five genes is significantly higher in modern cultivars of *indica* and *temperate japonica* than landrace rice of the same variety group. Genetic diversity of silent sites estimated by θ_w is significantly higher in modern cultivars of *tropical japonica* and *temperate japonica*, and marginally significantly higher in modern cultivars of *indica* than that in their respective landraces (Figure 1.7).

Recombination

Estimates of population recombination rate (ρ), minimum recombination events (Rm), and permutation tests are shown in Table 1.9. Wild rice, *O. rufipogon*, had the higher estimates of Rm (2-13) than any cultivated rice variety groups (mainly 0-2) at all studied starch genes except *Wx* in *indica. O. rufipogon* also had higher ρ values (4.040 to 19.192 per kb) at the starch genes except for *Iso1* and *Wx* in *indica*. Significant recombination was detected in *O. rufipogon* at *SbeIIb*, *SsIIa* and *Wx* genes. Among rice variety groups, *indica* and *tropical japonica* has slightly higher estimation of Rm or ρ values than do the other three rice variety groups at all starch genes. No significant recombination was detected by permutation tests in a*us*, *aromatic* and *temperate japonica* at five out of six studied starch genes. However, *indica* and *tropical japonica* showed significant recombination at three out of six starch genes.

MK tests

MK tests were performed separately for the six starch genes in *O. rufipogon*. The results are given in Table 1.10. Only *Sh2* yielded a significant result by Fisher's exact test (P=0.04). The ratio of nonsynonymous to synonymous mutation at *Sh2* is 7:5 in *O. rufipogon*, 0:6 between *O. rufipogon* and *O. meridionalis*. The ratios of nonsynonymous to synonymous mutation within

O. rufipogon range from 0:8 to 7:5. The ratios of nonsynonymous to synonymous mutation between *O. rufipogon* and *O. meridionalis* range from 0:6 to 1:0. The neutral indexes range from 0 to 2.45. The neutral indexes are not available for *Sh2* and *SbeIIb* due to the lack of nonsynonymous mutations between *O. rufipogon* and *O. meridionalis*. Population selection coefficients (gamma values) for all the six starch genes in *O. rufipogon* are positive and range from 1.31 to 3.02.

HKA tests

The results of HKA tests in *O. rufipogon* and the *O. sativa* variety groups are shown in Table 1.3. All the HKA tests of STS loci are significant, suggesting that some STS loci might be under selection. This result is consistent with the previous study by coalescent simulation (CAICEDO *et al.* 2007). Since STS loci are not a good neutral reference for HKA tests, HKA tests were performed in the starch genes alone. In order to include all the samples for this study, HKA test were performed in the five starch genes (No *Wx* because of its sampling). No significance was detected. After equating the samples of the five starch genes with *Wx,* HKA tests including with *Wx* were performed and showed significant result in *tropical japonica,* suggesting selection at some starch loci in *tropical japonica*.

MLHKA tests were then performed to detect selection in *tropical japonica* at individual loci. The results showed significant difference between a neutral model and the model with selection at *Wx* (P<0.05), which suggests selection at *Wx* in *tropical japonica*. The selection parameter (k value) at *Wx* in *tropical japonica* is 0.13, which indicates that the level of diversity at *Wx* is decreased to 0.13 of the neutral expectation due to selection. All other selection models are not significantly different from the neutral model in *tropical japonica*, which suggests neutral evolution at other five starch genes in *tropical japonica*. The selection parameter is greater than 1 at all the other starch genes, which suggests the elevation of diversity over neutral expectation at these loci in *tropical japonica*.

Association between nucleotide variation and position in the metabolic pathway

No significant association between either the level of variation (as measured by θ_π and θ_w) or pattern of variation (as measured by Tajima's D and Fay and Wu's H) and the position within the metabolic pathway was detected in *O. rufipogon* (Table 1.11). However, the result should be viewed with caution because only a subset of the starch synthesis pathway genes (six genes) were included for the test and the analysis simply considers the gene position in the pathway and ignored the other part of the pathway reticulation.

Contrasting the observed data against a standard neutral model (SNM)

Values of Tajima's D for *O. rufipogon* are given in Table 1.6. These statistics showed a broad range of values across the studied starch genes. In *O. rufipogon, Sh2* showed a significantly negative Tajima's D value, which reflects an excess of low frequency alleles.

Since I am interested in the evolution of starch genes during domestication and later improvement after domestication, analyses were performed separately for landraces and modern cultivars. Both Tajima's D and Fay and Wu's H showed a broad range of values among the rice variety groups (Table 1.12 and 1.13). Both Tajima's D and Fay and Wu's H values deviated significantly from SNM at *Wx* and *SsIIa* in landraces of *temperate japonica*, and at *SbeIIb* and *Wx* in *aromatic* rice. However, no gene showed significant values of both Tajima's D and Fay and Wu's H in *aus*, *indica*, *tropical japonica* and any modern cultivar variety groups. Significant positive values of Tajima's D were detected in *aus* at *Wx*, and in landraces of *tropical japonica* at *SbeIIb*. Significant negative values of Fay and Wu's H or Tajima's D were detected in landraces or modern cultivars of some rice variety groups (for example, Tajima's D at *SbeIIb* in the landraces of *tropical japonica*). Significant deviation from SNM by estimation of Tajima's D or Fay and Wu's H might be caused by selection or demographic events such as bottlenecks.

In order to compare the starch genes with RCSTS loci, θ_w and Tajima's D value were calculated in the resampled *O. rufipogon* (Table 1.14). Both Tajima's D and Fay and Wu's H values were also calculated in the resampled cultivated rice variety groups (Table 1.15) (see Method). Except for *SbeIIb*, consistent patterns of deviation from SNM were found at all other studied starch genes before resampling and after resampling in *O. rufipogon* and rice variety groups (see Tables 1.6, 1.12, 1.14, 1.15). A different pattern was found at *SbeIIb* before and after resampling our samples in *indica*, *tropical japonica* and *aromatic* varieties. *SbeIIb* showed significant positive Tajima's D values in *tropical japonica*, significant negative Fay and Wu's H in *aromatic* varieties before resampling our samples but not after resampling. *SbeIIb* indicated significant positive Tajima's D in *indica* after resampling but not before resampling.

Contrasting the observed data against genome-wide RCSTS data

I conducted comparisons of Tajima's D or Fay and Wu's H values for six genes in six rice groups. Therefore, I divided 5% significant level by six times six. The significant level after Bonferrroni correction is 0.001389.

The values for Tajima's D and θ_w of the starch genes in the resampled *O. rufipogon* were compared with that of the RCSTS data (Table 1.14, Figure 1.5). Based on 5% significant level, *Sh2* and *Wx* deviated significantly from the genome-wide RCSTS data by Tajima's D value. However, none of these values is significant after Bonferroni correction. *Wx* showed significantly higher θ_w value in *O. rufipogon* over RCSTS data after Bonferroni correction.

Both Tajima's D and Fay and Wu's H in the resampled rice variety groups were compared with that of RCSTS data (Table 1.15, Figure 1.5). All rice variety groups showed one or two genes with Fay and Wu's H and/or Tajima's D which significantly deviated from the genome-wide RCSTS data based on 5% significance level. However after Bonferroni correction, only *Wx* in a*romatic* and *temperate japonica* continued to have significantly lower values of both Tajima's D and Fay and Wu's H than that of RCSTS data. *SbeIIb* in *aromatic* rice indicated significant lower value of Fay and Wu's H than that of RCSTS data after Bonferroni correction. *SbeIIb* also has lower value of Tajima's D in *aromatic* varieties than that of the RCSTS data although it is not a significant deviation after Bonferroni correction (P = 0.0098). *SbeIIb* in *aus* showed a significantly lower value of Fay and Wu's H than that of RCSTS data after Bonferroni correction. *Bt2* in *tropical japonica* and *temperate japonica*, *SsIIa* in *aromatic* have significantly higher values of Fay and Wu's H than that of the RCSTS data.

Derived allele frequency distribution

The distributions of derived allele frequency are shown in Figure 1.8. There is an excess of low frequency derived alleles in *O. rufipogon*, which is consistent with the estimation of Tajima's D (Table 1.6). Five of the starch genes (*Wx* excluded) showed negative Tajima' D values in *O. rufipogon*, which also suggested an excess of rare alleles. This pattern suggests selective constraint on these genes in *O. rufipogon*.

In the cultivated rice groups (Figure 1.8), both the pattern of excess of rare derived alleles or high frequency of derived allele were observed, which is consistent with Tajima's D values (Tables 1.12, 1.13). The pattern of the excess high frequency derived alleles is also observed in the 111 STS loci (CAICEDO *et al.* 2007). This pattern can be the result of positive selection or demographic events (bottleneck and population expansion) during domestication. However it is unlikely that all the cases were explained by positive selection.

Haplotype networks

Haplotype networks were constructed separately for each gene in *O. sativa* to examine the relationship between different rice groups. The list of haplotypes and haplotype networks in *O. sativa* is shown in Table 1.16. Different rice variety groups share haplotypes, but differ in haplotype frequency. The significant difference in haplotype frequency between rice variety groups was determined by Wilcoxon nonparametric signed rank test (Table 1.17). Only one case showed a significant difference (*indica* vs. *aromatic* at *SbeIIb* gene, P = 0.04).

The haplotypes for wild rice are listed in Table 1.18. The haplotype networks of both *O. sativa* and *O. rufipogon* are shown in Figure 1.9. There is more haplotype diversity in wild rice than in cultivated rice. However there are haplotypes which have a high frequency in cultivated rice but a low frequency in wild rice (for example: haplotype A2 of *Sh2*; A1 of *SbeIIb*; B2 and D5 of *Iso1*; A4 and C7 of *SsIIa*). Haplotypes which have a high frequency in cultivated rice but were not detected in wild rice were also found (for example: B10 and B13 of *Bt2*; B4 and D2 of *SbeIIb*; A2 of *Iso1*; B10 and C8 of *SsIIa*).

Transposable elements in SsIIa and Wx

By blasting against the *Oryza* repeat database, I detected a miniature inverted-repeat transposable element (MITE) with a size approximately 360 bp in intron 7 of *SsIIa* (http://rice.plantbiology.msu.edu/blast.shtml). A different MITE with similar size might have been inserted at this location in *O. meridionalis*, which makes it unalignable to the rest of our samples. This MITE is the member of the stowaway family which is associated with the genes in Angiosperms and is AT rich (around 72%) (BUREAU and WESSLER 1994). Previous studies indicate that there is a short interspersed element (SINE) (about 125 bp) in intron 1 and a MITE (about 75 bp) in intron 13 of *Wx* gene (UMEDA *et al.* 1991). Nucleotide variation analysis indicates that there is higher variation in this transposable region than in the rest gene regions of

O. rufipogon in both *SsIIa* and *Wx* (Figure 1.10), which indicates the possible contribution of transposable elements to the level of diversity at *SsIIa* and *Wx*.

Discussion

Pattern of diversity

Domestication is a process of strong selection for desirable traits favored by humans (DIAMOND 2002). During domestication, domesticated species also experienced severe bottleneck events, which can result in lower diversity in cultivated species compared with that in their wild ancestors (CAICEDO *et al.* 2007). Genetic changes associated with domestication have been documented in previous studies in crop species such as rice, corn and wheat, as well as in domesticated animal species (BRIGGS and GOLDMAN 2006; EYRE-WALKER *et al.* 1998; HAUDRY *et al.* 2007; HYTEN *et al.* 2006; KAVAR and DOVC 2008; MILLER and SCHAAL 2006; ZHU *et al.* 2007). This study in starch genes has also revealed a pattern of lower diversity in cultivated rice groups than in the wild ancestor, *O. rufipogon.* The lower level of diversity most likely reflects bottleneck events during rice domestication (ZHU *et al.* 2007). The possible selective sweep effect on some studied genes during rice domestication might also contribute to this pattern. However, it is unlikely that selective sweeps affect all the starch genes.

Cultivated rice, *O. savtiva*, was domesticated from *O. rufipogon* around 10,000 – 12,000 years ago and experienced severe bottleneck events during its domestication (CAICEDO *et al.* 2007; ZHU *et al.* 2007). It is expected that haplotypes of cultivated rice would be a subset of haplotypes in *O. rufipogon* because it is unlikely that cultivated rice accumulated some new mutations and increased them to high frequency in such a short period of time. However, the haplotypes which exist in cultivated rice but not in wild ancestor were found. These are most likely the result of their low frequency or restricted distribution range in wild rice and our sampling (see the results for haplotype network). It is highly possible that some haplotypes which have low frequency or restricted distribution range in wild rice were passed on to cultivated rice and our sampling did not include these haplotypes. Haplotypes with high frequency in cultivated rice but low frequency in wild rice were also discovered, which is consistent with the cultivated rice demographic history of bottleneck events and population expansion during domestication. Some alleles in cultivated rice would be expected to increase in frequency due to population expansion events or selection. However, it is unlikely that the pattern in all these studied genes can be explained by selection because selective forces will only affect alleles that are responsible for the favored traits during domestication or those linked with them.

My study of *indica*, *tropical japonica*, and *temperate japonica* rice showed slightly higher diversity in modern cultivars compared with traditional landraces. Both the patterns of slightly higher or lower diversity in modern cultivars compared with that of their landraces have been documented in previous studies. The study of *indica* rice from Tamil Nadu (8 landraces and 12 modern cultivars) based on 664 AFLP markers showed slightly higher diversity in modern cultivars compared to landraces (PRASHANTH *et al.* 2002). However, a study of Indonesian *indica* (168 landraces vs 63 modern cultivars) based on 30 SSR loci indicated lower diversity in modern cultivars (THOMSON *et al.* 2007). During the process of late improvement in 1960s or more recently, rice experienced the so-called "green revolution" and experienced strong selection for high nutrients absorbing efficiency, short stems, high yield and disease resistance (KHUSH 2001). Associated with the strong selection, modern cultivars experienced bottleneck events and genetic introgression (MCNALLY *et al.* 2009). The bottleneck events will decrease diversity while genetic introgression from other rice variety groups or wild rice will increase diversity. Different genomic regions might experience different extents of genetic introgression. Moreover, the amount of effect on diversity by bottleneck events and genetic introgression might be comparable during rice late improvement stage. Therefore, these two contrasting effects can result in different relative levels of diversity between landraces and their modern cultivars if different genomic regions were sampled or different samples were used. Our results might only indicate that genetic introgression plays a relatively more important role in evolution of the studied starch genes than bottleneck events, not the entire genome of rice. However, our results can also be the results of slightly different sampling range of modern cultivars and their landraces.

Recombination rate

A previous study shows a higher population recombination rate (ρ value) in *O. rufipogon*, followed by *indica*, *tropical japonica* and *temperate japonica* (Mather, 2007). My results also show higher ρ value in *O. rufipogon* than in rice variety groups. However, within rice variety groups, the pattern is not consistent among six genes. This inconsistency might be due to the short sequence region (3.12-4.65 kb), which could cause large statistical error for estimation of ρ values. Effective population size, outcrossing rate, domestication, and demographic history all play a role in shaping population recombination rates (MATHER *et al.* 2007). It was suggested that high outcrossing rate and large population size will result in high recombination rate (MATHER *et al.* 2007). In all six genes, *O. rufipogon* appears to have more recombination than the rice variety groups, consistent with its greater outcrossing rates (GAO *et al.* 2007; OKA 1988) and a larger effective population size (CAICEDO *et al.* 2007).

MK tests

Significant excess of nonsynonymous mutations in *O. rufipogon* was observed only at *Sh2*, which might suggest the presence of positive selection that favors amino-acid replacements in the protein product or the relaxation of purifying selection at *Sh2*. However, this result should be viewed with caution because of the low number of replacement and synonymous differences between *O. rufipogon* and its outgroup species (No. of fixed replacement mutations: 0-4, No. of fixed synonymous mutations: 0-13) (ANDOLFATTO 2008). Furthermore, if *Sh2* is under positive selection, a high frequency of the derived alleles is expected among the seven nonsynonymous mutations within *O. rufipogon*. However, very low frequency of the derived alleles was observed in these seven nonsynonymous mutations (one is 0.008, three is 0.02 and the rest three is 0.03). Among these seven nonsynonymous mutations, one showed its derived allele in one out of 63 *O. rufipogon* individuals as heterozygous state; three showed their derived alleles in three individuals and the other three showed their derived alleles in two individuals. This suggests that all the derived alleles at these seven nonsynonymous mutations were under strong purifying selection or were recently derived. However, it is unlikely that all these nonsynonymous mutations are recently derived and had no enough time to increase to high frequency. Finally, the population selection coefficient is the highest at *Sh2* among the six starch genes in *O. rufipogon.* The level of diversity at *Sh2* based on silent sites is the lowest among six starch genes. It is unlikely that selection pressure was relaxed at *Sh2* compared to other studied genes.

Evolution of the studied starch genes in O. rufipogon before domestication

According to all of the tests that we used for detecting selection, no strong evidence of selection was discovered in five *O. rufipogon* starch genes (no *Sh2*) before rice domestication. In contrast, *Sh2* might have experienced stronger selective constraint than other studied genes. First, among all the genes, *Sh2* has the lowest diversity, and its level of diversity based on θ_{ω} value is lower than 95.17% of RCSTS loci. Second, significant negative Tajima's D value was observed at *Sh2*, and this value is significantly deviated from genome-wide RCSTS data at the 5% significant level. The strong selective constraint on gene *Sh2* was also documented in *Zea Mays* (MANICACCI *et al.* 2007; WHITT *et al.* 2002). The strong selective constraints at *Sh2* in *O. rufipogon* might be due to its large effect on starch phenotype and its position in the starch synthesis pathway. Mutation in *Sh2* can result in shrunken seeds (BHAVE *et al.* 1990). Both *Sh2* and *Bt2* encode subunits of ADP-glucose pyrophosphorylase (AGPase), and they catalyze a rate-limiting step in the synthesis of both amylose and amylopectin. However, the enzyme coded by *Wx* only controls the amylose production, and enzymes coded by *Iso1*, *SbeIIb* and *SsIIa* only control the amylopectin production (Fig 1.1).

The pattern of neutrality at five *O. rufipogon* starch genes (no *Sh2*) suggests that genetic drift plays more important roles than selection at these five starch genes before rice domestication. The diversity pattern of *Sh2* in *O. rufipogon* suggests the important role of purifying selection on *Sh2* evolution. These results show no evidence of directional selection at six starch genes in *O. rufipogon*, which suggests that these starch genes did not contribute to the adaptive shift of starch traits in *O. rufipogon*, or these starch genes contributed to the adaptation and lost the signal of directional selection. This might also suggest that there is no adaptive shift of starch traits in *O. rufipogon*, or other starch genes were responsible the adaptive shift in *O. rufipon*. These hypotheses require further study with more starch genes or characterization of starch quality in *O. rufipogon* and its close relatives.

ADP-glucose pyrophosphorylase (AGPase) is a heterotetramer with two small and two large subunits. Smith-White and Preiss (SMITHWHITE and PREISS 1992) suggested that the small subunit is more selectively constrained than is the large subunit among species in angiosperms.

In rice, the large subunits of AGPase are encoded by *Sh2*, the small subunits are encoded by *Bt2*. Under Smith-White and Preiss's expectation, *Sh2* should have higher diversity than gene *Bt2* in *O. rufipogon*. However, our results showed the opposite, lower diversity at *Sh2* than that at *Bt2* in *O. rufipogon* (*Sh2*, 0.00334; *Bt2*, 0.00464). Previous study in *Zea mays* also showed that the diversity of *Sh2* is half of that in gene *Bt2* (WHITT *et al.* 2002). This might suggest that gene *Sh2* and *Bt2* have different relative patterns of evolution at the macroevolutionary and microevolutionary level. However, it might suggest nothing since no statistical evidence was provided.

The level of diversity at *Wx* is significantly higher than that of RCSTS data. The high diversity at *Wx* is probably due to high diversity of transposable elements region of the gene. *Wx* has two transposable elements, and the transposable element regions showed the highest diversity in *O. rufipogon*. The high diversity in the transposable element regions may be explained by the high mutation rate within the region. High mutation rate in transposable elements has been documented in several species (Souames, 2003; Koga, 2006). Furthermore, all the tests for selection including HKA, MK, Tajima's D, Fay and Wu's H, the comparison with genome wide RCSTS data indicated that *Wx* was under neutral evolution in *O. rufipogon* before rice domestication. High diversity at *Wx* is probably also due to the lack of strong selection.

Evolution of starch genes during domestication

Strong evidence of directional selection was found at *Wx* in *tropical japonica*, *temperate japonica* and *aromatic* rice. However, the evidence was different for these rice variety groups. The evidence of directional selection at *Wx* in *tropical japonica* is based on the level of diversity. The MLHKA test in *tropical japonica* suggested selection at *Wx* (Table 1.19); level of diversity is the lowest at *Wx* in *tropical japonica* among all the rice variety groups at any studied genes;

and the reduction of diversity at *Wx* in *tropical japonica* relative to *O. rufipogon* is the most extreme (Figure 1.6). In contrast, evidence of directional selection at *Wx aromatic* and *temperate japonica* was from the pattern of allele frequency spectrum. *Wx* showed significant deviation in *aromatic* and *temperate japonica* by both comparison with a standard neutral model and comparison to genome wide RCSTS data both for Tajima's D and Fay and Wu's H measures.

Evidence of directional selection was also found at *SbeIIb* in *aromatic* rice. For Tajima's D and Fay and Wu's H values, S*beIIb* in *aromatic* deviates significantly from genome wide RCSTS data. *SbeIIb* also deviates significantly from SNM based on Tajima's D. However it is not significantly deviated from standard neutral model (SNM) according to Fay and Wu's H statistics although its Fay and Wu's H value is significantly lower (-10.400) than that of RCSTS data. This inconsistency is also observed for *SbeIIb* in *aus*. These results may be due to the low sample size of *aromatic* after resampling.

SsIIa has been suggested as the gene responsible for the starch quality difference between *indica* and *japonica* rice (Umemoto, 2005; Umemoto, 2002; Waters, 2006). As expected, it might be under directional selection in *indica* or *japonica* rice. However, no strong evidence of directional selection in *indica* or *japonica* rice is provided in this study. *SsIIa* significantly deviates from SNM by both Tajima's D and Fay and Wu's H statistics in *temperate japonica* rice. And only a very low proportion of RCSTS loci showed lower Tajima's D and Fay and Wu's H value than *SsIIa* (P(D_RCSTS<D_starch)= 0.06488, P(H_RCSTS<H_starch) = 0.00895), which, however, are not significantly different from genome-wide RCSTS loci. The pattern of *SsIIa* in *temperate japonica* might be explained as the result of directional selection. This is because the comparison with RCSTS loci is very conservative for detecting directional selection (CAICEDO *et al.* 2007). Some RCSTS loci might be also under directional selection themselves in *temperate*

japonica, which, however, are considered as neutral reference (CAICEDO *et al.* 2007). Moreover, Bonferroni correction is very conservative and results in a greatly diminished power to detect selection. A gene will be considered as significant deviation only if the gene is completely deviated from RCSTS loci (p<0.00069 or p>0.9993 after Bonferroni correction). Therefore, studies with better neutral reference or a demographical model for rice are required to determine the evolutionary pattern at *SsIIa* in rice. Since no strong evidence of directional selection at *SsIIa* is shown in *temperate japonica*. *SsIIa* could also be explained as under neutral evolution in *temperate japonica* rice, which might suggest that there are other genes responsible for the starch quality difference between *indica* and *japonica* rice, which however have not been identified yet.

The pattern of *Bt2* in *temperate japonica*, *tropical japonica* and *SsIIa* in *aromatic* might be explained as neutral evolution rather than balancing selection although they showed significantly higher Fay and Wu's H value over genome wide RCSTS data. However, their Tajima's D value does not deviate significantly from SNM and genome wide STS data set. Their Fay and Wu's H value also does not deviate significantly from SNM. Furthermore, previous study indicated that these STS loci show an excess of high frequency derived alleles, which are better explained by the bottleneck plus selective sweep model rather than neutral demographic model in *tropical japonica* rice (CAICEDO *et al.* 2007). This suggests that not all STS loci are neutral, and the frequency distribution of Fay and Wu's H values of RCSTS loci skew to negative value. It is risky to consider that the gene is not under neutral evolution in *tropical japonica* rice only because it has significantly higher Fay and Wu's H value compared to the RCSTS loci.

Evidence of directional selection was found at some starch genes in *tropical japonica*, *temperate japonica* and *aromatic* but not in *aus* or *indica* rice. Previous studies suggest independent domestication events for *aus, indica* and *japonica* (LONDO *et al.* 2006). Therefore, it is likely that starch quality is a trait that is under selection during the domestication for *aromatic*, *japonicas.* The origin of *aromatic* rice is still unclear. However, as in *japonicas*, *Wx* is also under selection in *aromatic* varieties. This could be explained as a single selection event at *Wx* during the domestication of *japonicas* and *aromatic*. However this requires the evidence of a single origin for *japonicas* and *aromatic*. It is also possible that gene *Wx* is independently selected in *aromatic* if there is independent origin for *aromatic* and *japonicas*. Furthermore, *SbeIIb* is under selection in *aromatic,* not in other variety groups. This might suggests that it was under directional selection for unique starch quality in *aromatic* rice. These questions require further research in the evolutionary history of *aromatic* and finer starch quality survey.

Although the same gene (*Wx*) was selected in both *tropical* and *temperate japonica* and domestication for *japonicas* is a single event, it is unlikely that selection at *Wx* in *japonicas* is a single event. *Wx* is believed to be a gene that affects amylose content in rice seed endosperm. However, *tropical japonica* is characterized with high amylose content (~20-30%) while *temperate japonica* has low amylose content (~10-20%). Therefore, it is unlikely that the same allele was selected for both *japonica* groups for the opposite starch quality. Furthermore, a mutant that is under strong selection in *temperate japonica* has been documented (OLSEN *et al.* 2006). This mutation is a G to T mutation at the 59 splice site of *Wx* intron 1, which leads to incomplete post-transcriptional processing of the pre-mRNA and cause undetectable levels of spliced mRNA in glutinous *temperate japonica* individuals (Bligh, 1998; Wang, 1995; Cai, 1998; Hirano, 1998; Isshiki, 1998). The size of selective sweep caused by selection for this mutant in *temperate japonica* is about 250 kb, which suggests a strong selection (OLSEN *et al.* 2006). A selective sweep is the reduction or elimination of variation among the nucleotides in a neighboring DNA region of a mutation as the result of selection. It was suggested that this G to T mutation originally arose in *tropical japonica* (OLSEN and PURUGGANAN 2002). This mutant has been increased to a high frequency in *temperate japonica* (17/20) but had a very low frequency in *tropical japonica* (1/18), which suggests that a different mutant or location is under selection at *Wx* in *tropical japonica* (OLSEN *et al.* 2006). Further investigation is required to understand the exact target of selection at *Wx* in *tropical japonica*. However it could be concluded here that selection at *Wx* in *tropical and temperate japonica* might be independent events and might have occurred after the divergence of these two *japonica* groups.

The strong evidence of positive selection at *SbeIIb*, *Wx* in aromatic, at *Wx* in *tropical japonica*, *temperate japonica* rice suggests that selection plays a major role at these genes in these cultivated rice groups during rice domestication. This also suggests that *SbeIIb* and *Wx* contribute to the adaptive shift of starch traits in *aromatic* rice, *Wx* contributes to the adaptive shift of starch traits in *tropical* and *temperate japonica*. All the other starch candidate genes within cultivated rice group show the pattern of neutrality, which suggests the important role of genetic drift within these cases. No evidence of directional selection were discovered in *aus* and *indica* among six candidate starch genes, which either suggests that there is no adaptive shift of starch traits among these two cultivated rice groups, or other unsampled starch starch genes were involved in the adaptive shift of starch traits in aus and indica. These two hypotheses could be tested with further studies of more candidate starch genes and starch traits comparison between O. rufipogon and cultivated rice groups.

Genes that were under directional selection during domestication were all located downstream of the starch synthesis pathway (see Figure 1.1). This may not simply be the result of evolutionary stochasticity. The upstream genes did show a lower level of diversity before rice domestication although no significant association between nucleotide variation and position in the metabolic pathway was detected. This might suggest strong selective constraint on these genes before domestication, which caused the escape of directional selection at these upstream genes during domestication. Similar case at *Sh2* has been documented in *Zea mays*. However, selection at downstream genes of starch synthesis pathway might be simply the efficient response to selection for starch quality during domestication. The starch quality difference is caused by the ratio of amylose to amylopectin or amylopectin structure. The downstream genes directly affect the production of amylose and amylopectin as their position in starch synthesis pathway suggested (Fig 1.1). Therefore it might be more efficient for selection to act on downstream genes than on upstream genes for certain starch quality. However, more information about the target nucleotides for selection and their effects on starch quality in rice is required to test the above hypothesis.

Benefits and Challenges of detecting selection in domesticated species

Ever since Darwin, domesticated species have been used as model species for the study of evolution. Domestication of plants or animals from their wild ancestors has typically involved rapid phenotypic evolution in response to strong directional selection (HARLAN 1992). These dramatic, human-mediated transformations provide an excellent model for studying directional selection. The advantages of a domestication model system also include a well documented and short timescale for domestication, a suite of known traits that were under intense selection during domestication and the accumulation of genetic information (see introduction of the dissertation for detail benefits of the study system).

However, there is also some challenges to distinguish the pattern of directional selection from neutrality in domesticated species (HAMBLIN *et al.* 2006; TENAILLON *et al.* 2004). One difficulty is that domestication is a complicated evolutionary process which includes not only strong artificial selection for traits important to farmers and breeders but also demographic events such as bottlenecks, population expansion (EYRE-WALKER *et al.* 1998; TENAILLON *et al.* 2004). Most current available standard methods for detecting selection can not distinguish the pattern of selection from bottleneck events (FAY and WU 2000; FU and LI 1993; TAJIMA 1989). These methods compare the studied genes with the standard neutral model, which has the assumption of constant population size. However, the pattern of deviation from the standard neutral model can be explained not only as selection and but also as demographic events such as a bottleneck or population expansion. Almost all domesticated species experienced severe bottleneck and population expansion events. Therefore, the power of most standard methods for detecting selection in domesticated species is limited.

An alternative strategy to detect selection in domesticated species or other natural species with unknown or complicated demographic history is to compare a gene with genome-wide variation (RAMOS-ONSINS *et al.* 2008; WRIGHT and GAUT 2005). This strategy assumes that the pattern of genome-wide variation should mainly reflect species' demographic history. Therefore, deviation from genome-wide variation pattern means deviation from the true neutral model, which should suggest selection. This approach is limited by the availability of genome-wide variation data. The high throughput sequencing techniques will continue to increase the quantity of genome-wide variation data at multiple species. However, the pattern of genome-wide variation might also be significantly affected by selective force if selection affects a large part of the genome. It is especially possible in selfing domesticated species. The reason is that high selfing rates will decrease the recombination rates, thus amplify the signal of selection and affect the genome-wide variation pattern. Rice is an example. The pattern of genome-wide STS loci in *tropical japonica* and *indica* is better explained by demography plus selective sweep model instead of neutral demographic model, which might be caused by the strong selection during rice domestication and the mating system of rice (CAICEDO *et al.* 2007). Therefore, this approach might be too conservative to detect directional selection in rice since the genome-wide variation pattern was also affected by selective sweep and could not be explained by its demographic history alone.

Although both strategies of detecting selection have limitations, the combination of these two approaches increases the liability of detecting directional selection. For example, we found out that several genes in this study, which showed the pattern of significant deviation from SNM, are not significantly deviated from genome-wide variation pattern. Some genes in this study (*Bt2* in *temperate japonica* and *tropical japonica rice*, *Wx* in *aus*, *SbeIIb* in *indica*), which significantly deviate from genome-wide variation distribution, do not significantly deviate from standard neutral model. The situation of inconsistency by these two strategies is difficult to be explained. One solution to overcome this difficulty is to compare the studied loci with the evolutionary model of a species to determine selection. The evolutionary model should reflects the most likely demographic history of the species (RAMOS-ONSINS *et al.* 2008), which is possible with the increasing availability of genome-wide sequence data at species level and the development of modeling.

Another challenge of detecting selection is that domesticated plants may have experienced introgression to/from wild relatives (Sweeney, 2007). Interspecies gene flow are common between wild species (GASKIN and SCHAAL 2002; SOBRAL *et al.* 1994; WANG *et al.* 1992; WHITTEMORE and SCHAAL 1991), and between wild species and cultivated species (ALDRICH and DOEBLEY 1992; WILLIAMS and STCLAIR 1993). Natural gene flow between *Oryza* wild species to cultivated rice has been frequently reported in literatures (MAJUMDER *et al.* 1997; SONG *et al.* 2003; SONG *et al.* 2006). In addition, in order to increase cultivated rice yield, quality or disease resistance, breeders frequently bring alleles from wild rice to cultivated rice (BRAR and KHUSH 1997). Furthermore, It was suggested that there was limited introgression between divergent cultivated rice gene pools, which transferred key domestication alleles (KOVACH *et al.* 2007). As a result of genetic introgression, some genomic regions of domesticated plants might show the pattern of an excess of high frequency derived alleles because cultivated individuals my carry "wild" alleles or the outgroup species may carry "cultivated" allele. Statistics such as Fay and Wu's H (FAY and WU 2000), which require an outgroup species to determine derived alleles, are quite sensitive to effect of introgression. One solution is to use statistics such as Tajima's D (TAJIMA 1989), which do not require outgroup species to infer derived alleles.

Conclusions

I have shown here the diversity level of six starch genes in five rice variety groups and their ancestor species, *O. rufipogon*. The diversity of the starch genes is significantly higher in *O. rufipogon* than that in any rice variety groups. The level of diversity of starch genes is slightly higher in modern cultivars than traditional landrace in *indica, tropical* and *temperate japonica*, which might be the result of the genetic introgression during modern improvement. No association between nucleotide variation and position in the metabolic pathway was found in *O. rufipogon*. However, upstream genes *Sh2* did show low diversity and significant deviation from SNM by Tajima's D value, which might simply suggest strong selective constraints at this gene before domestication. The level of diversity is significantly higher at *Wx* in *O. rufipogon*, which might be due to high diversity of the transposable elements.

Evidence of directional selection was detected at *Wx* in *tropical japonica*, *temperate japonica,* and at *Wx* and *SbeIIb* in *aromatic*, but not in *aus* and *indica*, which suggests that starch quality might be a trait under selection during the domestication for *aromatic* and *japonicas*. Although the same gene (*Wx*) was selected in *aromatic*, *tropical* and *temperate japonica* rice, it appears to be not a single selection event. The origin of *aromatic* is unknown and might be an independent event. Although *temperate japonica* was derived from *tropical japonica*, it is likely that selection at *Wx* in both *japonica* groups occurred after their divergence because they have different target of selection and different amylose content in rice seed endosperm (which is controlled by *Wx* in rice). Furthermore, our study also suggests further investigation at *SbeIIb* in *aromatic* for the detail target of selection and their contribution to the starch quality difference between rice variety groups.

References

ALDRICH, P. R., and J. DOEBLEY, 1992 Restriction Fragment Variation in the Nuclear and Chloroplast Genomes of Cultivated and Wild Sorghum-Bicolor. Theoretical and Applied Genetics **85:** 293-302.

AMUNDADOTTIR, L., P. KRAFT, R. Z. STOLZENBERG-SOLOMON, C. S. FUCHS, G. M. PETERSEN *et al.*, 2009 Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nature Genetics **41:** 986-U947.

ANDOLFATTO, P., 2008 Controlling Type-I Error of the McDonald-Kreitman Test in Genomewide Scans for Selection on Noncoding DNA. Genetics **180:** 1767-1771.

BANDELT, H. J., P. FORSTER and A. ROHL, 1999 Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution **16:** 37-48.

BENTSINK, L., J. JOWETT, C. J. HANHART and M. KOORNNEEF, 2006 Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **103:** 17042-17047.

BHAVE, M. R., S. LAWRENCE, C. BARTON and L. C. HANNAH, 1990 Identification and Molecular Characterization of Shrunken-2 cDNA Clones of Maize. Plant Cell **2:** 581-588.

BISWAS, S., and J. M. AKEY, 2006 Genomic insights into positive selection. Trends in Genetics **22:** 437-446.

BRADBURY, P. J., Z. ZHANG, D. E. KROON, T. M. CASSTEVENS, Y. RAMDOSS *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635.

BRAR, D. S., and G. S. KHUSH, 1997 Alien introgression in rice. Plant Molecular Biology **35:** 35-47.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The Hitchhiking Effect on the Site Frequency-Spectrum of DNA Polymorphisms. Genetics **140:** 783-796.

BRIGGS, W. H., and I. L. GOLDMAN, 2006 Genetic variation and selection response in model breeding populations of Brassica rapa following a diversity bottleneck. Genetics **172:** 457-465.

BULÉON, A., P. COLONNA, V. PLANCHOT and S. BALL, 1998 Starch granules: structure and biosynthesis. International Journal of Biological Macromolecules **23:** 85-112.

BUREAU, T. E., and S. R. WESSLER, 1994 Stowaway - a New Family of Inverted Repeat Elements Associated with the Genes of Both Monocotyledonous and Dicotyledonous Plants. Plant Cell **6:** 907-916.

BURGER, J. C., M. A. CHAPMAN and J. M. BURKE, 2008 Molecular insights into the evolution of crop plants. American Journal of Botany **95:** 113-122.

BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature **416:** 531-534.

CAICEDO, A., S. WILLIAMSON, R. D. HERNANDEZ, A. BOYKO, A. FLEDEL-ALON *et al.*, 2007 Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLoS Genetics **3:** e163.

CLEGG, M. T., 1997 Plant genetic diversity and the struggle to measure selection. Journal of Heredity **88:** 1-7.

COLLADO-VIDES, J., and R. HOFESTADT, 2002 Gene Regulation and Metabolism: Post-Genomic Computational Approaches. MIT Press**:** 103-128.

DARWIN, C., 1859 *On the Origin of Species*. Harvard University Press, Cambridge, MA.

DIAMOND, J., 2002 Evolution, consequences and future of plant and animal domestication. Nature **418:** 700-707.

DKINSON, D. B., and J. PREISS, 1969 Presence of ADP-glucose pyrophosphorylase in Shrunken-2 and Brittle-2 mutants of maize endosperm. Plant physiology **44:** 1058-1062.

DOYLE, J. J., and J. L. DOYLE, 1990 Isolation of plant DNA from fresh tissue. Focus **12:** 13-15.

EXCOFFIER, L., G. LAVAL and S. SCHNEIDER, 2005 Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics **1:** 47-50.

EYRE-WALKER, A., R. L. GAUT, H. HILTON, D. L. FELDMAN and B. S. GAUT, 1998 Investigation of the bottleneck leading to the domestication of maize. Proceedings of the National Academy of Sciences of the United States of America **95:** 4441-4446.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

FRASER, H. B., AARON E. HIRSH, LARS M. STEINMETZ, C. SCHARFE and M. W. FELDMAN, 2002 Evolutionary Rate in the Protein Interaction Network. Science **296:** 750.

FU, Y. X., and W. H. LI, 1993 Statistical Tests of Neutrality of Mutations. Genetics **133:** 693-709.

FULTON, T. M., T. BECKBUNN, D. EMMATTY, Y. ESHED, J. LOPEZ *et al.*, 1997 QTL analysis of an advanced backcross of Lycopersicon peruvianum to the cultivated tomato and comparisons with QTLs found in other wild species. Theoretical and Applied Genetics **95:** 881-894.

FUSHENG WEI, 2,3 ED COE,4,5 WILLIAM NELSON,2,3,6 ARVIND K BHARTI,7 FRED ENGLER,2,3,6 ED BUTLER,1,2,3 HYERAN KIM,1,2,3 JOSE LUIS GOICOECHEA,1,2,3 MINGSHENG CHEN,1,2,3 SEUNGHEE LEE,1,2,3 GALINA FUKS,7 HECTOR SANCHEZ-VILLEDA,4 STEVEN

SCHROEDER,4 ZHIWEI FANG,4 MICHAEL MCMULLEN,4,5 GEORGIA DAVIS,4 JOHN E BOWERS,8 ANDREW H PATERSON,8 MARY SCHAEFFER,4,5 JACK GARDINER,4 KAREN CONE,9 JOACHIM MESSING,7 CAROL SODERLUND,2,3,6* AND ROD A WING1,2,3*, 2007 Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History. Plos Genetics **3:** 123.

GAO, H., S. WILLIAMSON and C. D. BUSTAMANTE, 2007 A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics **176:** 1635-1651.

GAO, Z. Y., D. L. ZENG, X. CUI, Y. H. ZHOU, M. YAN *et al.*, 2003 Map-based cloning of the ALK gene, which controls the gelatinization temperature of rice. Science in China Series C-Life Sciences **46:** 661-668.

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.

GASKIN, J. F., and B. A. SCHAAL, 2002 Hybrid Tamarix widespread in US invasion and undetected in native Asian range. Proceedings of the National Academy of Sciences of the United States of America **99:** 11256-11259.

HAMBLIN, M. T., A. M. CASA, H. SUN, S. C. MURRAY, A. H. PATERSON *et al.*, 2006 Challenges of detecting directional selection after a bottleneck: Lessons from Sorghum bicolor. Genetics **173:** 953-964.

HARLAN, J. R., 1992 *Crops and Man*. American Society of Agronomy, Madison, WI.

HARLAN, J. R., J. M. J. DEWET and E. G. PRICE, 1973 Comparative evolution of cereals. Evolution; International Journal of Organic Evolution **27:** 311-325.

HAROLD, D., R. ABRAHAM, P. HOLLINGWORTH, R. SIMS, A. GERRISH *et al.*, 2009 Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nature Genetics **41:** 1088-U1061.

HAUDRY, A., A. CENCI, C. RAVEL, T. BATAILLON, D. BRUNEL *et al.*, 2007 Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication. Molecular Biololgy and Evolution **24:** 1506-1517.

HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805-1817.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA-Sequences. Genetics **111:** 147-164.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics **116:** 153-159.

HYTEN, D. L., Q. J. SONG, Y. L. ZHU, I. Y. CHOI, R. L. NELSON *et al.*, 2006 Impacts of genetic bottlenecks on soybean genome diversity. Proceedings of the National Academy of Sciences of the United States of America **103:** 16666-16671.

INNAN, H., and Y. KIM, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America **101:** 10667-10672.

INNAN, H., K. Y. ZHANG, P. MARJORAM, S. TAVARE and N. A. ROSENBERG, 2005 Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. Genetics **169:** 1763-1777.

JAMES, M. G., K. DENYER and A. M. MYERS, 2003 Starch synthesis in the cereal endosperm. Current Opinion in Plant Biology **6:** 215-222.

KAVAR, T., and P. DOVC, 2008 Domestication of the horse: Genetic relationships between domestic and wild horses. Livestock Science **116:** 1-14.

KHUSH, G. S., 1997 Origin, dispersal, cultivation and variation of rice. Plant Molecular Biology **V35:** 25-34.

KHUSH, G. S., 2001 Green revolution: the way forward. Nature Reviews Genetics **2:** 815-822.

KIM, K. N., D. K. FISHER, M. GAO and M. J. GUILTINAN, 1998 Molecular cloning and characterization of the amylose-extender gene encoding starch branching enzyme IIB in maize. Plant Molecular Biology **38:** 945-956.

KIM, Y., and R. NIELSEN, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167:** 1513-1524.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624-626.

KINGMAN, J., 1982 The coalescent. Stochastic Processes and their Applications **13:** 235-248.

KONISHI, S., T. IZAWA, S. Y. LIN, K. EBANA, Y. FUKUTA *et al.*, 2006 An SNP caused loss of seed shattering during rice domestication. Science **312:** 1392-1396.

KOVACH, M. J., M. T. SWEENEY and S. R. MCCOUCH, 2007 New insights into the history of rice domestication. Trends in Genetics **23:** 578-587.

LI, C. B., A. L. ZHOU and T. SANG, 2006 Rice domestication by reducing shattering. Science **311:** 1936-1939.

LI, H. P., and W. STEPHAN, 2005 Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics **171:** 377-384.

LI, W., and B. GILL, 2006 Multiple genetic pathways for seed shattering in the grasses. Functional & Integrative Genomics **6:** 300-309.

LIBRADO, P., and J. ROZAS, 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25:** 1451-1452.

LIN, S. Y., T. SASAKI and M. YANO, 1998 Mapping quantitative trait loci controlling seed dormancy and heading date in rice, Oryza sativa L., using backcross inbred lines. Theoretical and Applied Genetics **96:** 997-1003.

LONDO, J. P., Y. C. CHIANG, K. H. HUNG, T. Y. CHIANG and B. A. SCHAAL, 2006 Phylogeography of Asian wild rice, Oryza rufipogon, reveals multiple independent domestications of cultivated rice, Oryza sativa. Proceedings of the National Academy of Sciences of the United States of America **103:** 9578-9583.

LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Sunderland, MA.

MAJUMDER, N. D., T. RAM and A. C. SHARMA, 1997 Cytological and morphological variation in hybrid swarms and introgressed population of interspecific hybrids (Oryza rufipogon Griff x Oryza sativa L) and its impact on evolution of intermediate types. Euphytica **94:** 295-302.

MANICACCI, D., M. FALQUE, S. LE GUILLOU, B. PIEGU, A. M. HENRY *et al.*, 2007 Maize Sh2 gene is constrained by natural selection but escaped domestication. Journal of Evolutionary Biology **20:** 503-516.

MARCHINI, J., L. R. CARDON, M. S. PHILLIPS and P. DONNELLY, 2004 The effects of human population structure on large genetic association studies. Nature Genetics **36:** 512-517.

MATHER, K. A., A. L. CAICEDO, N. R. POLATO, K. M. OLSEN, S. MCCOUCH *et al.*, 2007 The extent of linkage disequilibrium in rice (Oryza sativa L.). Genetics **177:** 2223-2232.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive Protein Evolution at the Adh Locus in Drosophila. Nature **351:** 652-654.

MCNALLY, K. L., K. L. CHILDS, R. BOHNERT, R. M. DAVIDSON, K. ZHAO *et al.*, 2009 Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proceedings of the National Academy of Sciences **106:** 12273-12278.

MCVEAN, G., P. AWADALLA and P. FEARNHEAD, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231-1241.

MILLER, A. J., and B. A. SCHAAL, 2006 Domestication and the distribution of genetic variation in wild and cultivated populations of the Mesoamerican fruit tree Spondias purpurea L. (Anacardiaceae). Molecular Ecology **15:** 1467-1480.

MIZUNO, K., T. KAWASAKI, H. SHIMADA, H. SATOH, E. KOBAYASHI *et al.*, 1993 Alteration of the Structural-Properties of Starch Components by the Lack of an Isoform of Starch Branching Enzyme in Rice Seeds. Journal of Biological Chemistry **268:** 19084-19091.

MOHAPATRA, P. K., R. PATEL and S. K. SAHU, 1993 Time of flowering affects grain quality and spikelet partitioning within the rice panicle. Australian Journal of Plant Physiology **20:** 231-241.

MOORE, R. C., and M. D. PURUGGANAN, 2003 The early stages of duplicate gene evolution. Proceedings of the National Academy of Sciences **100:** 15682-15687.

MYERS, A. M., M. K. MORELL, M. G. JAMES and S. G. BALL, 2000 Recent progress toward understanding biosynthesis of the amylopectin crystal. Plant Physiology **122:** 989-997.

NACHMAN, M. W., H. E. HOEKSTRA and S. L. D'AGOSTINO, 2003 The genetic basis of adaptive melanism in pocket mice. Proceedings of the National Academy of Sciences of the United States of America **100:** 5268-5273.

NAKAMURA, Y., A. SATO and B. O. JULIANO, 2006 Short-chain-length distribution in debranched rice starches differing in gelatinization temperature or cooked rice hardness. Starch-Starke **58:** 155-160.

OKA, H., and H. MORISHIMA, 1997 *Wild and cultivated rice*. Genetics, Nobunkyo, Tokyo.

OKA, H. I., 1988 Origin of cultivated rice.

OLSEN, K. M., A. L. CAICEDO, N. POLATO, A. MCCLUNG, S. MCCOUCH *et al.*, 2006 Selection under domestication: Evidence for a sweep in the rice Waxy genomic region. Genetics **173:** 975-983.

OLSEN, K. M., and M. D. PURUGGANAN, 2002 Molecular evidence on the origin and evolution of glutinous rice. Genetics **162:** 941-950.

PERRETANT, M. R., T. CADALEN, G. CHARMET, P. SOURDILLE, P. NICOLAS *et al.*, 2000 QTL analysis of bread-making quality in wheat using a doubled haploid population. Theoretical and Applied Genetics **100:** 1167-1175.

PRASHANTH, S. R., M. PARANI, B. P. MOHANTY, V. TALAME, R. TUBEROSA *et al.*, 2002 Genetic diversity in cultivars and landraces of Oryza sativa subsp indica as revealed by AFLP markers. Genome **45:** 451-459.

PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179-1189.

PRZEWORSKI, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164:** 1667-1676.

RAMOS-ONSINS, S. E., E. PUERMA, D. BALANA-ALCAIDE, D. SALGUERO and M. AGUADE, 2008 Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in Arabidopsis thaliana. Molecular Ecology **17:** 1211-1223.

RAND, D. M., and L. M. KANN, 1996 Excess amino acid polymorphism in mitochondrial DNA: Contrasts among genes from Drosophila, mice, and humans. Molecular Biology and Evolution **13:** 735-748.

ROZAS, J., J. C. SÀNCHEZ-DELBARRIO, X. MESSEQUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496-2497.

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. P. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832-837.

SABETI, P. C., P. VARILLY, B. FRY, J. LOHMUELLER, E. HOSTETTER *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature **449:** 913-U912.

SANO, Y., 1984 differential regulation of waxy gene expression in rice endosperm. Theoretical and Applied Genetics **68:** 467-473.

SATO, Y., Y. FUKUDA and H. Y. HIRANO, 2001 Mutations that cause amino acid substitutions at the invariant positions in homeodomain of OSH3KNOX protein suggest artificial selection during rice domestication. Genes & Genetic Systems **76:** 381-392.

SMIDANSKY, E. D., J. M. MARTIN, L. C. HANNAH, A. M. FISCHER and M. J. GIROUX, 2003 Seed yield and plant biomass increases in rice are conferred by deregulation of endosperm ADP-glucose pyrophosphorylase. Planta **216:** 656-664.

SMITHWHITE, B. J., and J. PREISS, 1992 Comparison of Proteins of Adp-Glucose Pyrophosphorylase from Diverse Sources. Journal of Molecular Evolution **34:** 449-464.

SOBRAL, B. W. S., D. P. V. BRAGA, E. S. LAHOOD and P. KEIM, 1994 Phylogenetic Analysis of Chloroplast Restriction Enzyme Site Mutations in the Saccharinae Griseb Subtribe of the Andropogoneae Dumort Tribe. Theoretical and Applied Genetics **87:** 843-853.

SOKAL, R. R., and F. J. ROHLF, 1995 *Biometry: the principles and practice of statistics in biological research. 3rd edition.* W. H. Freeman and Co.

SONG, Z. P., B. R. LU, Y. G. ZHU and J. K. CHEN, 2003 Gene flow from cultivated rice to the wild species Oryza rufipogon under experimental field conditions. New Phytologist **157:** 657-665.

SONG, Z. P., W. Y. ZHU, J. RONG, X. XU, J. K. CHEN *et al.*, 2006 Evidences of introgression from cultivated rice to Oryza rufipogon (Poaceae) populations based on SSR fingerprinting: implications for wild rice differentiation and conservation. Evolutionary Ecology **20:** 501-522.

SPECHT, J. E., K. CHASE, M. MACRANDER, G. L. GRAEF, J. CHUNG *et al.*, 2001 Soybean response to water: A QTL analysis of drought tolerance. Crop Science **41:** 493-509.

SWANSON, W. J., Z. H. ZHANG, M. F. WOLFNER and C. F. AQUADRO, 2001 Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proceedings of the National Academy of Sciences of the United States of America **98:** 2509-2514.

TAJIMA, F., 1983 Evolutionary Relationship of DNA-Sequences in Finite Populations. Genetics **105:** 437-460.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

TAMURA, K., J. DUDLEY, M. NEI and S. KUMAR, 2007 MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution **24:** 1596-1599.

TANAKA, N., N. FUJITA, A. NISHI, H. SATOH, Y. HOSAKA *et al.*, 2004 The structure of starch can be manipulated by changing the expression levels of starch branching enzyme llb in rice endosperm. Plant Biotechnology Journal **2:** 507-516.

TENAILLON, M. I., J. U'REN, O. TENAILLON and B. S. GAUT, 2004 Selection versus demography: A multilocus investigation of the domestication process in maize. Molecular Biology and Evolution **21:** 1214-1225.

THOMSON, M. J., E. M. SEPTININGSIH, F. SUWARDJO, T. J. SANTOSO, T. S. SILITONGA *et al.*, 2007 Genetic diversity analysis of traditional and improved Indonesian rice (Oryza sativa L.) germplasm using microsatellite markers. Theoretical and Applied Genetics **114:** 559-568.

TSAI, C. Y., and O. E. NELSON, 1966 Starch-deficient maize mutant lacking adenosine diphosphate glucose pyrophosphorylase activity. Science **151:** 341-343.

TURRI, M. G., S. R. DATTA, J. DEFRIES, N. D. HENDERSON and J. FLINT, 2001 QTL analysis identifies multiple behavioral dimensions in ethological tests of anxiety in laboratory mice. Current Biology **11:** 725-734.

UMEDA, M., H. OHTSUBO and E. OHTSUBO, 1991 Diversification of the Rice Waxy Gene by Insertion of Mobile DNA Elements into Introns. Japanese Journal of Genetics **66:** 569-586.

UMEMOTO, T., and N. AOKI, 2005 Single-nucleotide polymorphisms in rice starch synthase IIa that alter starch gelatinisation and starch association of the enzyme. Functional Plant Biology **32:** 763-768.

UMEMOTO, T., N. AOKI, H. X. LIN, Y. NAKAMURA, N. INOUCHI *et al.*, 2004 Natural variation in rice starch synthase IIa affects enzyme and starch properties. Functional Plant Biology **31:** 671-684.

UMEMOTO, T., M. YANO, H. SATOH, A. SHOMURA and Y. NAKAMURA, 2002 Mapping of a gene responsible for the difference in amylopectin structure between japonica-type and indica-type rice varieties. Theoretical and Applied Genetics **104:** 1-8.

WANG, Z. Y., G. SECOND and S. D. TANKSLEY, 1992 Polymorphism and Phylogenetic-Relationships among Species in the Genus Oryza as Determined by Analysis of Nuclear Rflps. Theoretical and Applied Genetics **83:** 565-581.

WARTH, F., and D. DARABSETT, 1914 disintegration of rice grains by means of alkali. Bulletin of Agricultural Research Institute **38:** 1-9.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biololgy **7:** 256-276.

WEISS, L. A., and D. E. ARKING, 2009 A genome-wide linkage and association scan reveals novel loci for autism. Nature **461:** 802-U862.

WHITT, S. R., L. M. WILSON, M. I. TENAILLON, B. S. GAUT and E. S. BUCKLER, 2002 Genetic diversity and selection in the maize starch pathway. Proceedings of the National Academy of Sciences of the United States of America **99:** 12959-12962.

WHITTEMORE, A. T., and B. A. SCHAAL, 1991 Interspecific Gene Flow in Sympatric Oaks. Proceedings of the National Academy of Sciences of the United States of America **88:** 2540-2544.

WILLIAMS, C. E., and D. A. STCLAIR, 1993 Phenetic Relationships and Levels of Variability Detected by Restriction-Fragment-Length-Polymorphism and Random Amplified

Polymorphic DNA Analysis of Cultivated and Wild Accessions of Lycopersicon-Esculentum. Genome **36:** 619-630.

WONG, W. S. W., and R. NIELSEN, 2004 Detecting selection in noncoding regions of nucleotide sequences. Genetics **167:** 949-958.

WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The effects of artificial selection of the maize genome. Science **308:** 1310-1314.

WRIGHT, S. I., and B. CHARLESWORTH, 2004 The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. Genetics **168:** 1071-1076.

WRIGHT, S. I., and B. S. GAUT, 2005 Molecular Population Genetics and the Search for Adaptive Evolution in Plants. Molecular Biology and Evolution **22:** 506-519.

YAMASAKI, M., M. I. TENAILLON, I. VROH BI, S. G. SCHROEDER, H. SANCHEZ-VILLEDA *et al.*, 2005 A Large-Scale Screen for Artificial Selection in Maize Identifies Candidate Agronomic Loci for Domestication and Crop Improvement. Plant Cell **17:** 2859-2872.

YAMASAKI, M., S. I. WRIGHT and M. D. MCMULLEN, 2007 Genomic Screening for Artificial Selection during Domestication and Improvement in Maize. Annals of Botany **100:** 967 - 973.

ZEDER, M. A., 2006 Central questions in the domestication of plants and animals. Evolutionary Anthropology **15:** 105-117.

ZHANG, Z. C., S. F. ZHANG, J. C. YANG and J. H. ZHANG, 2008 Yield, grain quality and water use efficiency of rice under non-flooded mulching cultivation. Field Crops Research **108:** 71-81.

ZHU, Q., X. ZHENG, J. LUO, B. S. GAUT and S. GE, 2007 Multilocus Analysis of Nucleotide Variation of Oryza sativa and Its Wild Relatives: Severe Bottleneck during Domestication of Rice. Molecular Biology and Evolution **24:** 875-888.

Table 1.1), Collections of *O. rufipogon* and *O. sativa* from IRRI and field used in the study

IRRI/USDA #[a]	Cultivar Name[b]	Species	Race[c]	Status[d]	Origin[e]	Label[f]	Resample[g]
CIor 12374	P166	*O. sativa*	aus	cultivar	China, Sichuan	C_CN_AI_04	
3397	Hashikalmi	*O. sativa*	aus	landrace	Suriname	L_SR_AI_01	
20461	ARC 7046	*O. sativa*	aus		India	U_IN_AI_01	Yes
22739	Bei Khe	*O. sativa*	aus	landrace	Cambodia	L_KH_AI_04	
64771	Chikon Shoni	*O. sativa*	aus	landrace	Bangledesh	L_BD_AI_01	
CIor 5987	Ramgarh	*O. sativa*	aus	landrace	India, Bihar	L_IN_AI_02	Yes
21289	ARC 11287	*O. sativa*	aus		India	U_IN_AI_02	Yes
66765	Asha	*O. sativa*	aus	landrace	Bangledesh	L_BD_AI_06	Yes
66828	Tepi Borua	*O. sativa*	aus	landrace	Bangledesh	L_BD_AI_09	Yes
67700	Bijri	*O. sativa*	aus	landrace	India	L_IN_AI_06	Yes
9466	Amarelao	*O. sativa*	indica	cultivar	Brazil	C_BR_I_02	
3384	Arroz en Granza	*O. sativa*	indica	cultivar	Guatemala	C_GT_I_01	
51231	CO 39	*O. sativa*	indica	cultivar	India, Orissa	C_IN_I_03	Yes
PI 584560	Gerdeh	*O. sativa*	indica	cultivar	Iran	C_IR_I_01	
15935	Balislus	*O. sativa*	indica	cultivar	Senegal	C_SN_I_02	
15113	Tunsart	*O. sativa*	indica	cultivar	Vietnam	C_VN_I_01	
66770	Bamura	*O. sativa*	indica	landrace	Bangledesh	L_BD_I_07	
67859	Zakha	*O. sativa*	indica	landrace	Bhutan	L_BT_I_07	
66463	Younoussa	*O. sativa*	indica	landrace	Guinea	L_GW_I_01	
74625	Gembira Kuning	*O. sativa*	indica	landrace	Indonesia	L_ID_I_11	Yes
21778	ARC 11956	*O. sativa*	indica	landrace	India	L_IN_I_04	Yes
5926	Aikoku	*O. sativa*	indica	landrace	Japan, Okinawa	L_JP_I_02	
12110	Phcar Tien	*O. sativa*	indica	landrace	Cambodia	L_KH_I_01	
8948	Pokkali	*O. sativa*	indica	landrace	Sri Lanka	L_LK_I_02	Yes
78916	Let Yone Gyi	*O. sativa*	indica	landrace	Myanmar	L_MM_I_02	
71508	Batu	*O. sativa*	indica	landrace	Malaysia	L_MY_I_04	
11443	Ramdulari	*O. sativa*	indica	landrace	Nepal	L_NP_I_02	
CIor 4637	Lupa	*O. sativa*	indica	landrace	Philippines	L_PH_I_02	
201	Doc Phung	*O. sativa*	indica	landrace	Vietnam	L_VN_I_01	
14503/11355	IR20	*O. sativa*	indica	cultivar	Philippines, Luzon	C_PH_I_09	Yes
39292	IR36	*O. sativa*	indica	cultivar	Philippines, Luzon	C_PH_I_10	
15058	KU188	*O. sativa*	indica	cultivar	Thailand	C_TH_I_04	Yes
	Chiem Chanh	*O. sativa*	indica	cultivar	Vietnam	C_VN_I_02	
5868	Doc Phung Lun	*O. sativa*	indica	cultivar	Vietnam	C_VN_I_03	
5803		*O. sativa*	indica	cultivar	Thailand	C_TH_I_05	Yes
6663		*O. sativa*	indica	cultivar	India	C_IN_I_04	Yes
26872		*O. sativa*	indica	Landrace	Philippines	L_PH_I_07	Yes
27513		*O. sativa*	indica	Landrace	Bangladesh	L_BD_I_10	Yes

43545		*O. sativa*	indica	Landrace	Indonesia	L_ID_I_17	Yes
51250		*O. sativa*	indica	cultivar	China	C_CN_I_05	Yes
56036		*O. sativa*	indica	Landrace	Vietnam	L_VN_I_16	Yes
58930		*O. sativa*	indica	Landrace	Nepal	L_NP_I_04	Yes
7755		*O. sativa*	indica	Landrace	Sri Lanka	L_LK_I_03	Yes
8240		*O. sativa*	indica	cultivar	Taiwan	C_TW_I_06	Yes
CIor 1496	Wanica	*O. sativa*	temp. japonica	cultivar	Suriname	C_SR_JP_01	Yes
PI 564580	Pare Riri	*O. sativa*	temp. japonica	landrace	Indonesia, Celebes	L_ID_JP_01	Yes
PI 419449	Silewah	*O. sativa*	temp. japonica	landrace	Indonesia, Sumatra	L_ID_JP_02	Yes
66647	Si Gepai	*O. sativa*	temp. japonica	landrace	Indonesia	L_ID_JP_08	Yes
14945	Kolleh	*O. sativa*	temp. japonica	landrace	Liberia	L_LR_JP_01	Yes
14383	Padi Babas	*O. sativa*	temp. japonica	landrace	Malaysia	L_MY_JP_02	Yes
13375	Jumula 2	*O. sativa*	temp. japonica	landrace	Nepal	L_NP_JP_03	Yes
CIor 4602	Kinabugan	*O. sativa*	temp. japonica	landrace	Philippines	L_PH_JP_01	Yes
78364	Nep Lao Hoa Binh	*O. sativa*	temp. japonica	landrace	Vietnam	L_VN_JP_10	Yes
	63-83	*O. sativa*	temp. japonica	cultivar	Cote D'Ivoire	C_CI_JP_02	Yes
PI 584575	Canella De Ferro	*O. sativa*	temp. japonica	cultivar	Brazil	C_BR_JP_01	Yes
64878	Bjanam	*O. sativa*	temp. japonica	landrace	Bhutan	L_BT_JP_01	Yes
77619	Pare Pulu Lotong	*O. sativa*	temp. japonica	landrace	Indonesia	L_ID_JP_14	Yes
CIor 12145	Kinabugan selection	*O. sativa*	temp. japonica	landrace	Philippines, Palawan	L_PH_JP_04	Yes
9669	Yacca	*O. sativa*	temp. japonica	landrace	West Africa	L_WR_JP_01	Yes
	IRAT 13	*O. sativa*	temp. japonica	cultivar	Cote D'Ivoire	C_CI_JP_03	Yes
	Adair	*O. sativa*	temp. japonica	cultivar	USA	C_US_JP_10	Yes
	Bluebonnet 50	*O. sativa*	temp. japonica	cultivar	USA	C_US_JP_12	Yes
CIor 9277	Smooth Zenith	*O. sativa*	temp. japonica	cultivar	USA, Texas	C_US_JP_13	Yes
	OS4	*O. sativa*	temp. japonica	cultivar	West Africa	C_WR_JP_01	Yes
	Koshihikari	*O. sativa*	temp. japonica	cultivar	Japan, Fukui	C_JP_JP_03	Yes
	Azucena	*O. sativa*	trop. japonica	cultivar	Philippines	C_PH_JV_03	Yes
70990	Silla	*O. sativa*	trop. japonica	cultivar	Italy	C_IT_JV_01	
39174	RD1	*O. sativa*	trop. japonica	cultivar	Thailand	C_TH_JV_01	
3493	Tung Ting Yellow	*O. sativa*	trop. japonica	cultivar	China, Jiangsu	C_CN_JV_02	
10810	Earlirose	*O. sativa*	trop. japonica	cultivar	USA, California	C_US_JV_14	Yes
33984	Baber	*O. sativa*	trop. japonica	cultivar	India, Kashmir	C_IN_JV_01	
43325	Arias	*O. sativa*	trop. japonica	Landrace	Indonesia, Java	C_ID_JV_01	Yes
PI 596815	376	*O. sativa*	trop. japonica	cultivar	Cambodia	C_KH_JV_01	
12908	Deng Mak Tek	*O. sativa*	trop. japonica	Landrace	Laos	C_LA_JV_01	
14371	Padi Siam	*O. sativa*	trop. japonica	landrace	Malaysia	L_MY_JV_01	
15814	Kalor	*O. sativa*	trop. japonica	cultivar	Senegal	C_SN_JV_01	
67759	Sathiya	*O. sativa*	trop. japonica	landrace	India	L_IN_JV_08	
77638	Aeguk	*O. sativa*	trop. japonica	landrace	Korea	L_KR_JV_01	Yes

66756		*O. sativa*	trop. japonica	cultivar	Texas	C_US_JV_01	Yes
38698		*O. sativa*	trop. japonica	cultivar	Pakistan	C_PK_JV_01	Yes
2545		*O. sativa*	trop. japonica	cultivar	Japan	C_JP_JV_01	Yes
16428*		*O. sativa*	trop. japonica	Landrace	Indonesia	L_ID_JV_01	Yes
24225		*O. sativa*	trop. japonica	Landrace	Thailand	L_TH_JV_01	Yes
25901		*O. sativa*	trop. japonica	Landrace	Bangladesh	L_BD_JV_01	Yes
43372		*O. sativa*	trop. japonica	Landrace	Indonesia (Bali)	L_ID_JV_02	Yes
8244		*O. sativa*	trop. japonica	Landrace	Philippines	L_PH_JV_02	Yes
15046		*O. sativa*	trop. japonica	Landrace	Thailand	L_TH_JV_02	Yes
12104		*O. sativa*	trop. japonica	Landrace	Vietnam	L_VN_JV_02	
12922		*O. sativa*	trop. japonica	Landrace	Laos	L_LA_JV_02	
19552		*O. sativa*	trop. japonica	Landrace	Maylasia	L_MY_JV_02	
22796		*O. sativa*	trop. japonica	Landrace	Cambodia	L_KH_JV_01	
CIor 5816	Basmati	*O. sativa*	aromatic		India, Assam	Basmati_1	Yes
CIor 8982	Basmati 3	*O. sativa*	aromatic		India, Delhi	Basmati_2	Yes
CIor 12524	Basmati	*O. sativa*	aromatic		India, Punjab	Basmati_3	Yes
PI 173923	Basmati	*O. sativa*	aromatic		India, Uttar Pradesh	Basmati_4	Yes
PI 584556	Basmati I	*O. sativa*	aromatic		Pakistan	Basmati_5	Yes
PI 412774	Basmati 5875	*O. sativa*	aromatic		Pakistan, N-W Front	Basmati_6	Yes
PI 385403	Basmati	*O. sativa*	aromatic		Pakistan, Punjab	Basmati_7	
PI 430924	Basmati	*O. sativa*	aromatic		Pakistan, Sind	Basmati_8	
81990		*O. rufipogon*	wild		Myanmar	W_MM_07	
100189		*O. rufipogon*	wild		Malaysia	W_MY_01	Yes
100588		*O. rufipogon*	wild		Taiwan	W_TW_02	
100916		*O. rufipogon*	wild		China	W_CN_01	Yes
103404		*O. rufipogon*	wild		Bangladesh	W_BD_01	
104599		*O. rufipogon*	wild		Sri Lanka	W_LK_02	
104624		*O. rufipogon*	wild		China	W_CN_02	Yes
104815		*O. rufipogon*	wild		Thailand	W_TH_13	
104618		*O. rufipogon*	wild		Thailand	W_TH_15	
104833		*O. rufipogon*	wild		Thailand	W_TH_17	
104857		*O. rufipogon*	wild		Thailand	W_TH_18	
104871		*O. rufipogon*	wild		Thailand	W_TH_21	
105567	Padi Hijang	*O. rufipogon*	wild		Indonesia	W_ID_04	
105568		*O. rufipogon*	wild		Philippines	W_PH_02	
105711	Kozhinelli	*O. rufipogon*	wild		India	W_IN_29	Yes
105956	Padi Padian	*O. rufipogon*	wild		Indonesia	W_ID_05	
106036	Padi Hantu	*O. rufipogon*	wild		Malaysia	W_MY_03	
106086	Uri Dan	*O. rufipogon*	wild		India	W_IN_32	Yes
106122		*O. rufipogon*	wild		India	W_IN_35	Yes

106168		*O. rufipogon*	wild		Vietnam	W_VN_14	
106321		*O. rufipogon*	wild		Cambodia	W_KH_11	
106078		*O. rufipogon*	wild		India	W_IN_31	Yes
106262		*O. rufipogon*	wild		Papau New Guinea	W_PG_04	
105898	Uri Dan	*O. rufipogon*	wild		Bangladesh	W_BD_04	Yes
105471		*O. rufipogon*	wild		India	W_IN_27	Yes
NSGC 5936		*O. rufipogon*	wild		India	W_IN_02	Yes
NSGC 5940		*O. rufipogon*	wild		India	W_IN_05	Yes
106103		*O. rufipogon*	wild		India	W_IN_33	Yes
106453		*O. rufipogon*	wild		Indonesia	W_ID_06	
106346		*O. rufipogon*	wild		Myanmar	W_MM_09	
104714		*O. rufipogon*	wild		Thailand	W_TH_06	
81994	Semo	*O. rufipogon*	wild		Papau New Guinea	W_PG_01	
100904		*O. rufipogon*	wild		Thailand	W_TH_30	
105855		*O. rufipogon*	wild		Thailand	W_TH_29	
106166		*O. rufipogon*	wild		Vietnam	W_VN_13	
105720		*O. rufipogon*	wild		Cambodia	W_KH_01	
106163		*O. rufipogon*	wild		Laos	W_LA_06	Yes
106523		*O. rufipogon*	wild		Papau New Guinea	W_PG_08	
105295		*O. barthii*	wild			barthii	
104119		*O. meridionalis*	wild			meridionalis	

Field collection

Population	Number[h]	Species	Status	Origin	Label	Resampled
1	2	*O. rufipogon*	wild	Guangdong, China	GD_E1_05, 12	
2	2	*O. rufipogon*	wild	Guangdong, China	GD_E2_12, 16	GD_E2_12
3	2	*O. rufipogon*	wild	Guangdong, China	GD_E3_04, 07	GD_E3_07
4	2	*O. rufipogon*	wild	Guangdong, China	GD_E4_09, 16	
5	1	*O. rufipogon*	wild	Guangdong, China	GD_GS_01	
6	2	*O. rufipogon*	wild	Guangdong, China	GD_W1_12, 17	
7	2	*O. rufipogon*	wild	Guangdong, China	GD_W2_06, 17	GD_W2_06
8	2	*O. rufipogon*	wild	Guangdong, China	GD_W3_06, 13	
9	1	*O. rufipogon*	wild	Guangxi, China	GX_1_08	
10	1	*O. rufipogon*	wild	Guangxi, China	GX_GS_01	
11	4	*O. rufipogon*	wild	Hainan, China	HA_N_09, 11, 16, 18	HA_N_09
12	4	*O. rufipogon*	wild	Hainan, China	HA_S_05, 12, 19, 29	HA_S_05
13	1	*O. rufipogon*	wild	Hunan, China	HN_1_02	
14	4	*O. rufipogon*	wild	Jianxi, China	JX_E_13, 18, 25, 28	JX_E_13
15	1	*O. rufipogon*	wild	Guangxi, China	W_CN_03	
16	1	*O. rufipogon*	wild	Hunan, China	W_CN_HN02	

[a] The identification number of the samples from IRRI or USDA.

[b] common name of the cultivated rice variety

[c] the racial designation of cultivated rice based on IRRI documentation and phenol reaction (tested by Jason Londo in Schaal lab)

[d] degree of cultivated rice based on IRRI supplied publication

[e] Country and region from which the original germplasm was donated or collected

[f] the label that I used in the study

[g] the samples used to compare with genome-wide STS data

[h] the number of individuals sampled from a population

The lists of primers used for PCRs and sequencing

Table 1.2a), The list of primers for PCRs and annealing temperature
Table 1.2b), The list of sequencing primers

Table 1.2a), The list of PCR primers and annealing temperature

Gene	Region	Forward (5'-3')	Reverse (5'-3')	Annealing Temperature (°C)
Sh2	1	GAGTACCCATGCAATTCATGATG	GCACCCTCTGGTTTCTCAAG	58
	2	CCGAGCTTCTGACTATGGAC	CGCCTATACCAATTGGGAC	58
Bt2	1	GTGGCAATCGTAGGATTGTTAG	GGCAGGGCAGCAACAGTAATATC	58
	2	CCGCATGGACTATGAAAAGTTC	CTGGTACAGGCTTTTTCG	57
	3	GGTGGCATTCCCATTGGTATTG	GCCCCATCACATATTACGCAG	57
Iso1	1	CTTCACCCCCGACGATCTCGA	CTCCAGCTCATTGAATTCATG	58
	2	CTGGGTGACAGAAATGCATG	CCAGTGAGGAAATTGACCTAC	60
	3	GGGAGGCCTCTATCAAGTAG	GTGGACTTCCACACAAACATTC	57
	4	CCGGGACATTGTTCGTCAATTC	CCAGCTGAGGTTATGATTTTC	57
	5	CCTTGGCACAGTATCAACTTTG	GCGGAATTTGGTCATAAGAG	58
	6	GGCCATACAAAAGGAGGCAAC	GGACCGCACAACTTCAACATATAC	55
SsIIa	1	GGCGAGGAGAGGACATCGTGTTATG	CCCCCCATAGATGTCATCCTG	60
	2	CCGTCACCGTCAGGATGACATC	CCTGGTAAGCGATATTATGT	53
	3	GCACTCCTGCCTGTTTATCTG	CGAGGCCACGGTGTAGTTG	53
	4	CGGGAGAACGACTGGAAGATGAAC	CAGACACGAGAGCTAATGAAG	58
SbeIb	1	GGACCCCAGAAGATTTAAAG	CTTAGAACCCAGAGGCCAATG	59
	2	CTGGTGGACAACAACTAGAAC	CAGGTGGTCACTAATCTTTG	57
	3	GTCGCTGGTTCAAGAAATTG	CCCCACCATCTTGAACAGGAAG	60
	4	GTCAGTGGAATGCCTACATTTG	GCTATTCCACGATCAATGCTAGGTG	60
	5	CCACCAGTACATATCTCGAAAGCA	GGCAGTTCGAACGGTACCAACA	60
Wx	1	CCATTCCTTCAGTTCTTTGTC	CCTTCACACTGAATTCTTGCAC	60

Table 1.2b), The list of sequencing primers

Gene	Segment	(5'-3')	(5'-3')
Sh2	1	GGCACAATGCACTTATTCAAG	GAGGGAAAGTCAACTGAAGGA
		GGCTGCTACACAAATGCCTG	GGGGTATTTTATGATCCATG
		GGGGCCTAGACTAATGAGACAC	
	2	CACCCATTTGTGATAGTTC	AACCCGGATTACAGTTGTTC
		CGCGTGAAGAGATTCCAATG	CCTCACCTCGATACTTACCTC
Bt2	1	GCGCAATCAGATCAGGTG	PCR primers
		CCAAGTCCAGTCATGAAC	
	2	CACCACATACTTCCATATC	CCCTCATTCCGATGTTTG
	3	CCCCATTGTTCTTGGAATG	CTGCGTAACTGTATGTAAC
		CCGCGAAGCACACATAGATG	
Iso1	1	CTTCACCCCCGACGATCTCGA	GGTAGGTCACCTTGCCAATC
		GGCAAGGTGACCTACCTCTG	CTCCAGCTCATTGAATTCATG
		Reverse PCR primer	
	2	PCR primers	
	3	PCR primers	
	4	PCR primers	
	5	CATGGTTGCTGGTTATATG	GGTCCTGGTTCATGCAGTTC
	6	GACTTCCCAACAGCTCAAC	CTCGCCTTTCGTTTCATCTTTC
SsIIa	1	CGCGTGCAGCGTGTCCATATG	CGGATCAGGCCCATACAATTAG
	2	PCR primers	
	3	PCR primers	
	4	AACGGCATCGTGAACGGCATC	AGCACAACAGCAAGGTGCGCGGGTG
		TATAGCTATAGCCTCCCTGAAG	
SbeIIb	1	PCR primers	
	2	PCR primers	
	3	PCR primers	
	4	Reverse PCR primer	GCTGTGGTTTTCATACCGTTC
	5	CGCCGTTCGACAGCCATAGAT	CATGAACGCGCTGCAAAACTG
		ATCTCGAAAGCATGAAGAGGA	

Tables for the results of HKA test for STS loci and starch genes in *O. rufipogon* and *O. sativa*

Table 1.3a), HKA test results of genome-wide STS loci in *O. rufipogon* and *O. sativa*
Table 1.3b), HKA test results of five starch genes (*Wx* excluded) in *O. rufipogon* and *O. sativa*
Table 1.3c), HKA test results of six studied starch genes (including *Wx*) in *O. rufipogon* and *O. sativa*

Table 1.3a), HKA test results of genome-wide STS loci in *O. rufipogon* and *O. sativa*

Rice group	**N**[a]	**Deviation**[b]	**degree of freedom**	**P_{chi}**[c]
O. rufipogon	21	162.3103	103	0.00017
aus	6	87.7802	50	0.0002
indica	21	139.2729	69	0.00001
tropical japonica	18	87.7802	46	0.0002
temperate japonica	21	56.0731	31	0.00381
aromatic	6	55.8619	34	0.01047

Table 1.3b), HKA test results of five starch genes (*Wx* excluded) in *O. rufipogon* and *O. sativa*

Rice group	**Deviation**[b]	**P_{chi}**[c]
O. rufipogon	3.0165	0.55507
aus	3.0031	0.55731
indica	0.4363	0.84783
tropical japonica	0.6402	0.77011
temperate japonica	1.3286	0.58170
aromatic	2.1566	0.70698

Table 1.3c), HKA test results of six studied starch genes (including *Wx*) in *O. rufipogon* and *O. sativa*

Rice group	**Deviation**[b]	**P_{chi}**[c]
O. rufipogon	7.3093	0.19863
aus	4.3105	0.50563
indica	2.1123	0.83340
tropical japonica	17.4076	0.00379*
temperate japonica	3.5296	0.61891
aromatic	2.8832	0.71799

[a] number of sample size for each rice group

[b] the sum of deviations in chi-square test

[c] chi-square distribution probability

*P<0.05

Table 1.4), Summary of polymorphism of 111 STS loci

Rice group	No. of polymorphmic STS loci	Percentage of STS loci with less than 3 polymorphic nucleotide sites
O. rufipogon	105	29.5
aus	57	93.3
indica	76	76.2
tropical japonica	54	85.7
temperate japonica	38	94.3
aromatic	39	91.4

Table 1.5), List of the start and end sequence of the segments for the haplotype networks in *O. rufipogon* and *O. sativa*

Regions	**Start (5'-3')**	**End ((5'-3'))**
Sh2 segment 1	TGAGTACCCATGCAATTCAT	TTTCGTTGTCAACTGTGAAT
Sh2 segment 2	AATAATCTAAATTCTATTTTT	GGTCCCAATTGGTATAGGC
Bt2 segment 1	AGGATTGTTAGTGGTTGAGG	AGTATAATTGTTTCTCAGTA
Bt2 segment 2	AAGAAAAAACAAATCATGAT	GAATAACGAAAAAGCCTGT
Bt2 segment 3	GGCATTCCCATTGGTATTGG	TGCGTAATATGTGATGGGG
Iso1 segment 1	AATCCGATGATTAGTTGAGG	ATAATAGCCTGTTTATATTTT
Iso1 segment 2	TAATGACCAGAGGATGCAG	TCATAGCACTGACTTATATT
Iso1 segment 3	ATGTGCTTGTTTCTCCAGTA	GGGGAAAACCATGTGAGC
Iso1 segment 4	CATTGACTAGGATTTGCGCT	TGATGATTGAAAGAACAAA
SbeIIb segment 1	GTTGTTCACAGGTAATTAAT	TTCTATTCAATTATTGAATTT
SbeIIb segment 2	CTGGGTTCTAAGCCCTTTTG	AGTTGATCAAATTATGTCGC
SbeIIb segment 3	ACCATCAATGTTATACGAGT	ATTTCAACTGTTCTGTGGTTA
SbeIIb segment 4	ATCACTCATAGGGTTATAGG	TGTGTATCAGTGTATCACCG
SbeIIb segment 5	CTGGTATTTGTGTTCAACTTC	TGCTTTTGTGCTTTGCGCTCC
SSIIa segment 1	CGAGGAGAGGACATCGTGT	AGGAATGAGATGTATTTGTT
SSIIa segment 2	TGGTTTGAGGATTGGTCAAA	GCTGGGGCCTCCACGACAT
SSIIa segment 3	CCTGCAGTCCGACGGCTACG	ATGTTACTTCTTCATTAGCTC

Table 1.6), Population statistics of starch genes in *O. rufipogon*

Statistics	***Sh2***	***Bt2***	***Iso1***	***SbeIIb***	***SsIIa***	***Wx***
sample size	63	63	66	62	45	9
nonsynonymous	7	1	8	2	3	4
synonymous	5	9	7	6	11	7
silent	47	62	99	112	86	168
$\boldsymbol{\theta_\pi}$	0.00146	0.00285	0.00267	0.00296	0.00566	0.02127
$\boldsymbol{\theta_W}$	0.00334	0.00464	0.00543	0.00519	0.00738	0.01728
Tajima's D	-1.80957*	-1.22540	-1.55277	-1.41417	-0.82375	1.18644

*P <0.05

Table 1.7), Nucleotide Polymorphism of starch genes in cultivated rice variety groups

				tropical	***temperate***	
		aus	***indica***	***japonica***	***japonica***	***aromatic***
Sh2	**sample size**	10	34	26	21	8
	nonsynonymous	0	0	0	0	0
	synonymous	1	4	4	4	1
	silent	9	24	20	20	6
	$\theta\pi$	0.00102	0.00151	0.00160	0.00143	0.00083
	θ_w	0.00123	0.00225	0.00201	0.00214	0.00089
Bt2	**sample size**	10	34	26	21	8
	nonsynonymous	0	1	1	0	0
	synonymous	1	3	4	3	1
	silent	7	30	29	23	7
	$\theta\pi$	0.00122	0.00146	0.00287	0.00194	0.00117
	θ_w	0.00101	0.00297	0.00308	0.00259	0.00109
Iso1	**sample size**	10	34	26	21	8
	nonsynonymous	1	3	2	2	1
	synonymous	2	4	4	3	1
	silent	17	32	23	17	13
	$\theta\pi$	0.00184	0.00159	0.00151	0.00053	0.00193
	θ_w	0.00179	0.00234	0.00180	0.00141	0.00150
SbeIIb	**sample size**	10	33	26	21	8
	nonsynonymous	0	0	0	0	0
	synonymous	2	2	2	2	2
	silent	22	29	27	22	22
	$\theta\pi$	0.00252	0.00222	0.00250	0.00252	0.00136
	θ_w	0.00194	0.00181	0.00179	0.00155	0.00210
SsIIa	**sample size**	8	30	23	20	8
	nonsynonymous	2	3	3	3	1
	synonymous	3	6	3	6	4
	silent	22	50	36	30	31
	$\theta\pi$	0.00387	0.00529	0.00310	0.00276	0.00527
	θ_w	0.00365	0.00544	0.00420	0.00364	0.00514
Wx	**sample size**	6	21	17	22	6
	nonsynonymous	0	2	0	1	1
	synonymous	0	3	0	1	2
	silent	27	102	7	57	66
	$\theta\pi$	0.00279	0.00594	0.00018	0.00109	0.00372
	θ_w	0.00209	0.00546	0.00033	0.00328	0.00489

Table 1.8), Nucleotide Polymorphism of five starch genes in rice landrace and modern cultivar

		indica		*tropical japonica*		*temperate japonica*	
Gene	**Statistics**	**L**[a]	**C**[b]	**L**	**C**	**L**	**C**
Sh2	**sample size**	19	15	13	13	12	9
	nonsynonymous	0	0	0	0	0	0
	synonymous	1	4	1	4	1	4
	silent	10	23	9	19	9	17
	$\theta\pi$	0.00099	0.00216	0.00139	0.00187	0.00120	0.00179
	θ_w	0.00110	0.00272	0.00111	0.00235	0.00114	0.00240
Bt2	**sample size**	19	15	13	13	12	9
	nonsynonymous	1	0	0	1	0	1
	synonymous	3	0	2	4	2	3
	silent	20	14	20	25	12	25
	$\theta\pi$	0.00132	0.00232	0.00271	0.00312	0.00155	0.00312
	θ_w	0.00232	0.00232	0.00261	0.00326	0.00174	0.00326
Iso1	**sample size**	19	15	13	13	12	9
	nonsynonymous	1	2	1	2	1	2
	synonymous	2	4	2	4	1	2
	silent	21	25	14	19	8	10
	$\theta\pi$	0.00129	0.00199	0.00171	0.00121	0.00044	0.00066
	θ_w	0.00179	0.00230	0.00135	0.00183	0.00079	0.00110
SbeIIb	**sample size**	19	14	12	13	12	9
	nonsynonymous	0	0	0	0	0	0
	synonymous	2	2	2	2	2	2
	silent	24	25	21	26	21	21
	$\theta\pi$	0.00184	0.00264	0.00168	0.00247	0.00168	0.00256
	θ_w	0.00174	0.00196	0.00176	0.00211	0.00176	0.00195
SsIIa	**sample size**	17	13	12	11	12	8
	nonsynonymous	3	2	2	3	3	3
	synonymous	2	6	2	3	2	6
	silent	27	48	24	30	15	28
	$\theta\pi$	0.00432	0.00616	0.00239	0.00401	0.00108	0.00496
	θ_w	0.00344	0.00666	0.00342	0.00441	0.00214	0.00465

[a] landrace

[b] modern cultivars

Table 1.9), Population Recombination Rate and minimum recombination events (Rm) estimates in starch genes

				tropical	*temperate*	
Gene	***O. rufipogon***	***aus***	***indica***	***japonica***	***japonica***	***aromatic***
Sh2	4.040(2)	2.020(0)	0(0)	0(0)	0(0)	0(0)
Bt2	9.091(5)	3.030(0)	0(0)	3.030****(1)	0 (0)	0(0)
Iso1	11.111(8)	0(1)	27.273****(6)	8.081(2)	0(0)	0(0)
SbeIIb	12.121*****(6)	3.061****(2)	5.051**(2)	3.030(3)	2.020(2)	NA
SsIIa	19.192***(13)	0(0)	6.061(8)	6.061*****(2)	5.051*(2)	16.162****(2)
Wx	4.040****(8)	1.010(0)	3.030*(10)	5.051****(1)	15.152(1)	NA

Rm are shown in parentheses;
Test of recombination by permutation test are shown by *; *P <0.05; **P <0.01; ***P<0.005; ****P<0.001

Table 1.10), MK tests of starch genes in *O. rufipogon*

	Polymorphism		**Fixed differences**				
Gene	**R[a]**	**S[b]**	**R**	**S**	**P[c]**	**NI[d]**	**Gamma[e]**
Sh2	7	5	0	6	0.04	NA[f]	3.02
Bt2	0	8	1	0	0.11	0.00	2.65
Iso1	8	7	4	4	1.00	1.14	1.31
SbeIIb	2	6	0	4	0.52	NA	2.92
SsIIa	3	11	1	9	0.61	2.45	2.50
Wx	4	7	4	13	0.67	1.86	2.29

[a] No. of replacement sites
[b] No. of synonymous sites
[c] probability of two tailed Fisher's exact test
[d] Neutrality index
[e] Gamma, population selection coefficient
[f] not available because the number of the fixed replacement sites is 0

Table 1.11), the results of association between nucleotide variation and position in starch synthesis pathway in *O. rufipogon* by Kendall’s rank correlation tests

Statistics	T	P(T' < T)	P(T' > T)
θ_π	-0.07	0.43	0.57
θ_w	0.07	0.57	0.43
Tajima's D	-0.20	0.29	0.71

Table 1.12), Population Allele frequency statistics of starch genes in rice variety groups

				tropical	***temperate***	
Gene	**Statistics**	***aus***	***indica_L***[a]	***japonica_L***	***japonica_L***	***aromatic***
Sh2	**Tajima's D**	-0.739	-0.372	0.971	0.213	-0.345
	Fay and Wu's H	-2.933	-0.433	0.731	-1.182	-2.571
Bt2	**Tajima's D**	0.900	-1.698	0.164	-0.479	0.340
	Fay and Wu's H	0.089	0.942	3.718	2.394	0.643
Iso1	**Tajima's D**	-1.553	-1.046	1.229	**-1.830***	1.509
	Fay and Wu's H	-4.533	-6.363	2.013	-2.182	-0.857
SbeIIb	**Tajima's D**	1.419	0.225	**2.080***	-0.192	**-1.840****
	Fay and Wu's H	-1.778	-7.029	-2.641	-8.333	**-14.143****
SsIIa	**Tajima's D**	0.138	0.929	-1.181	**-2.165****	0.145
	Fay and Wu's H	1.071	-0.559	**-6.485***	**-9.242****	4.214
Wx	**Tajima's D**	**2.087***	0.373	-1.590	**-2.615*****	**-1.537***
	Fay and Wu's H	0.533	-1.190	0.221	**-20.745*****	**-16.533****

[a]L is landrace
[*]P<0.05
[**]P<0.01
[***]P<0.001

Table 1.13), Population allele frequency statistics of starch gene in rice landrace and modern cultivars

		indica		***tropical japonica***		***temperate japonica***	
Gene	**Statistics**	**L**[a]	**C**[b]	**L**	**C**	**L**	**C**
sh2	**Tajima's D**	-0.37154	-0.8511	0.9714	-0.8706	0.2126	-1.2361
	Fay and Wu's H	-0.43275	-1.1905	0.7308	-2.0897	-1.1818	-2.6667
bt2	**Tajima's D**	-1.69754	-0.2989	0.1642	-0.2765	-0.4788	-0.7595
	Fay and Wu's H	0.94152	-5.3429	3.7180	-3.7692	2.3939	-3.2222
ISO1	**Tajima's D**	-1.04627	-0.6423	1.2294	-1.3812	-1.8304*	1.8764*
	Fay and Wu's H	-6.36257	-3.8000	2.0128	1.5128	-2.1818	-1.1667
SBEIIb	**Tajima's D**	0.22544	1.4933	2.0795*	0.7490	-0.1917	1.5218
	Fay and Wu's H	-7.02924	1.4066	-2.6410	-0.6923	-8.3333	-2.2778
SSIIa	**Tajima's D**	0.92871	-0.3805	-1.1806	-0.2829	-2.1649**	0.2098
	Fay and Wu's H	-0.55882	1.9103	-6.4849*	-4.4182	-9.2424**	-0.5714

[a] landrace
[b] modern cultivars
*P<0.05
**P<0.01
***P<0.001

Table 1.14), Nucleotide diversity pattern of the studied starch gene in the resampled *O. rufipogon*

Gene	θ_w	P(θ_w_RCSTS <θ_w_starch)[a]	FW-H[b]	P(D_RCSTS< D_starch)[c]
Sh2	0.00314	0.04829	-1.926*	0.00201
Bt2	0.00365	0.11167	-0.500	0.40946
SsIIa	0.00665	0.59356	-0.544	0.38632
SbeIIb	0.00412	0.16583	-1.212	0.08753
Iso1	0.00421	0.17907	-1.085	0.12475
Wx	0.01728	1.00000	1.186	0.97990

[a] P(θ_w _RCSTS< θ_w _starch) is the percentage of RCSTS loci which has θ_w value lower than the studied starch gene.

[b] Fay and Wu's H

[c] P(D_RCSTS<D_starch) is the percentage of RCSTS loci which has Tajima's D value lower than the studied starch gene.

[*] P<0.05; It means significant deviation from standard neutral model

Table 1.15), Population allele frequency statistics of starch gene in the resampled rice variety groups

Rice variety group	**Gene**	**Tajima's D**	**P(D_RCSTS <D_starch)[a]**	**FW-H[b]**	**P(H_RCSTS <H_starch)[c]**
aus	***Sh2***	-0.013	0.45053	-0.533	0.29684
	Bt2	0.956	0.74316	-0.267	0.38632
	SsIIa	1.331	0.84737	-1.600	0.06947
	SbeIIb	-1.223	0.06737	-8.000	0.00000
	Iso1	0.740	0.69263	-1.333	0.12421
	Wx	2.087*	0.98947	0.533	0.71684
indica	***Sh2***	-0.605	0.45546	-0.417	0.65966
	Bt2	-1.547	0.07107	0.733	0.95095
	SsIIa	0.867	0.91892	0.044	0.81181
	SbeIIb	2.221*	0.99900	-0.167	0.84084
	Iso1	-0.999	0.28228	-6.933	0.02302
	Wx	0.373	0.82382	-1.190	0.44244
tropical japonica	***Sh2***	0.484	0.78808	-0.654	0.65011
	Bt2	0.368	0.76932	2.449	1.00000
	SsIIa	-1.708	0.15784	-6.933*	0.04084
	SbeIIb	0.491	0.79139	-6.038	0.06291
	Iso1	-1.107	0.38300	-0.218	0.70088
	Wx	-1.590	0.18653	0.221	0.70088
temperate japonica	***Sh2***	0.213	0.90716	-1.182	0.69799
	Bt2	-0.479	0.76510	2.394	1.00000
	SsIIa	-2.165**	0.06488	-9.242**	0.00895
	SbeIIb	-0.192	0.83221	-8.333	0.02573
	Iso1	-1.830*	0.18680	-2.182	0.45414
	Wx	-2.615***	0.00001	-20.745***	0.00001
aromatic	***Sh2***	-0.206	0.54254	-2.133	0.22734
	Bt2	0.708	0.72524	0.533	0.81032
	SsIIa	-0.135	0.56067	4.533	1.00000
	SbeIIb	-1.495*	0.00976	-10.400	0.00000
	Iso1	0.974	0.79219	0.048	0.72385
	Wx	-1.537*	0.00001	-16.533**	0.00001

[a] P(D_RCSTS<D_starch) is the percentage of RCSTS loci which has Tajima's D value lower than the studied starch gene.

[b] Fay and Wu's H

[c] P(H_RCSTS<H_starch) is the percentage of RCSTS loci which has Tajima's D value lower than the studied starch gene.

* means significantly deviation from the standard neutral model; *P<0.05; **P<0.01; ***P<0.001

Tables of haplotypes and haplotype networks of starch genes in *O. sativa*; The heads of tables shows the location of the polymorphic nucleotide. I3 means intron 3, E11 means exon 11. The star (Fusheng Wei) indicates nonsynonymous mutation; the first sequence is the reference sequence; the following is the list of haplotypes. The dot (.) means the nucleotide is the same as reference sequence. The dash (-) and the plus (+) means deletion and insertion. The number following them is the number of nucleotide deletion or insertion.

Varieties and tax are shown as follows: *aus* (□), *indica* (🟦), *tropical japonica* (🟩), *temperate japonica* (🟪), *aromatic* (▩), *O. barthii* (■). The circles represent the nodes. The size the circle is proportional to the number of individuals. The line connecting the nodes is roughly proportional to the number of mutations.

.

Table 1.16a), Haplotypes of *SsIIa* in *O. sativa*

```
      IIIIIIIIIIIIIIIIIIEEEIIIIIIIIIIIIIIIIIIIIIIEEEEEEEEEEE
      333333333333333333444445666677777777777777788888888888
                                                  ** * *
      AGGCTGCGCACGCCGCAGCTGGAGTCACCCCCGTCAATGGGCGCGGGGGCGGGG
H1    ...A.C.....A............A.....T...G.....A.T...........
H2    ...A.C.A...A............A.....T...G.....A.T...........
H3    ...ATC.....A............A.....T...G.....A.T...........
H4    ...A.C.....A............A.....T...G.....A.T.A.........
H5    ...A.C.....A............A.....TA..G.....A.T.A.T.......
H6    ...A.C.....A............A.....TA..G.....A.T...........
H7    ...A.C.....A............A..T..T..CA.....A.T...........
H8    ...A.C.....A............A..T..T..CA.....A.T........AA.
H9    ....T.T.T.......G.......A..T..T..CA.....A.T...........
H10   ...A.C.....A............A...A......C..................
H11   ......T.T.......G.........C.A......C..................
H12   ......T.T.......G.........C.A......C................A.
H13   ......T.T.......G.........C.A......C............TT.AA.
H14   ......T.T.......G.........C.A......C...............AA.
H15   ......T.T.......G.........C.AG.....C........A.T.TT.AA.
H16   ......T.T.......G.........C.A......C........A.T.TT.AA.
H17   ......T.T.......G.........C.A......C........A.T....AA.
H18   ......T.T.......G.........C.A......C........ACT....AA.
H19   ......T.T.......G.........C.A......C........A.TA...AA.
H20   ......T.T.T.....G.........C........C...............AA.
H21   ......T.T.......G.........C.A......C............A..AA.
H22   ......T.T.......G.........C.........GA......A.T....AA.
H23   ......T.T.......G.........C.........GAAA....A.T....AA.
H24   ..A......G........C..A........T...G.....A.T...........
H25   ..A......G...........A..............GA................
H26   ..A......G...........A..............GA......A.T.......
H27   .A.......GT.T.....TCA.T.............GA.A.............A
H28   .........G....CT..TCA.T.............GAAA..............
H29   ....A....GT...CT..TCA.T.............GAAA.....C........
H30   .........G...T.........A.A......A........T.TA.T....AA.
H31   G........G...T...T.....A.A......A........T.TA.T...A...
H32   G........GC..T...T.....A.A......A........T.T......A...
```

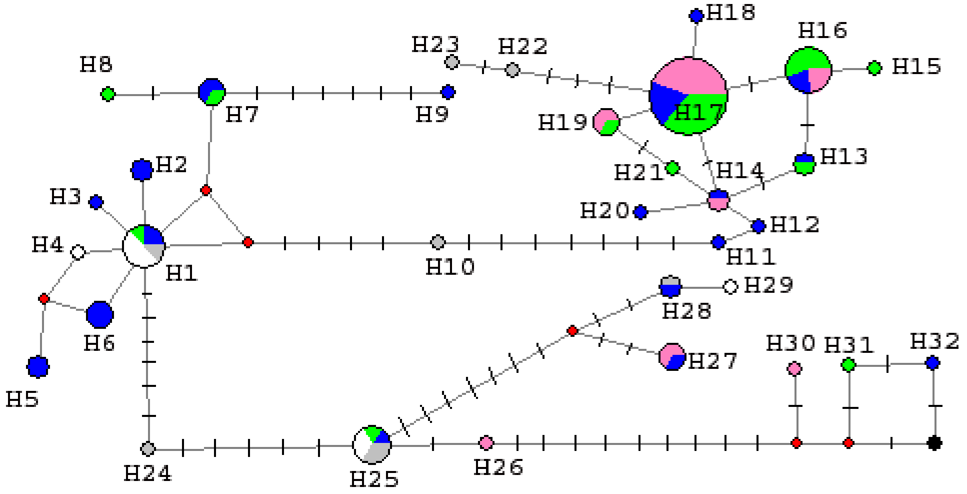


Table 1.16b), Haplotypes of *Sh2* in *O. sativa*

	E	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	E	E	E	I
														1	1	1	1	1	1	1	1	1	1	1	1	1
	5	5	5	5	5	5	6	6	6	6	6	6	7	0	0	1	1	1	1	2	2	3	4	4	4	6
	T	C	T	G	C	C	G	C	A	A	T	C	C	G	C	C	C	G	C	G	C	T	C	C	T	G
H1	.	.	.	T	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
H2	.	.	.	T	.	T	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.
H3	.	.	.	T	.	T	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.
H4	.	.	.	T	.	T	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
H5	.	.	.	T	.	T	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
H6	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
H7	.	.	.	.	.	.	.	.	.	G	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.
H8	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.
H9	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	T	.	.	T	.	.	.	.	.	.	.
H10	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.
H11	.	.	.	.	.	.	A	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.
H12	C	A	C	.	T	.	.	.	.	.	.	.	.	A	G	.	.	.	.	A	T	C	T	.	A	.
H13	C	A	C	.	T	.	.	.	.	.	.	.	.	A	G	.	.	.	.	A	T	C	T	.	A	T

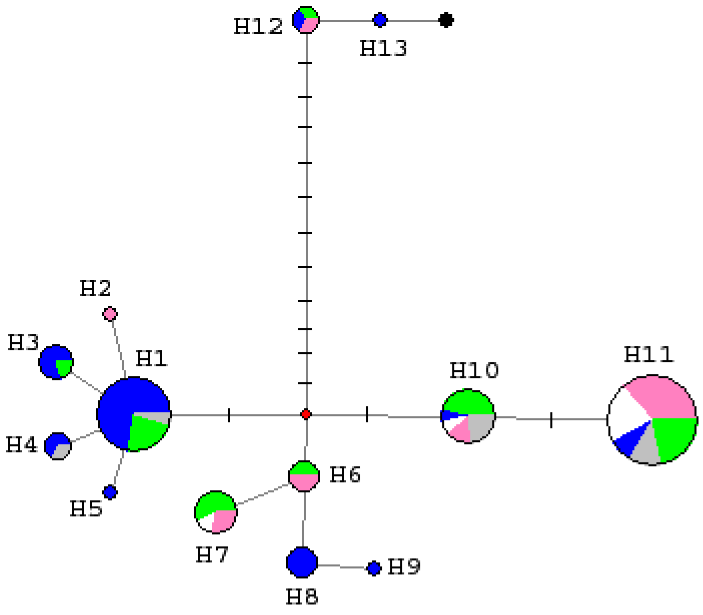


Table 1.16c), Haplotypes of *Bt2* in *O. sativa*

```
        IIIIIIIIEIIIIIIIIIIIIIIIIIIIEEEEEIIEEIIIIIIIE
        222222223333333333333333334444444556688888889
                                       *    *
        TTCCAATGCCAATGGTGGAATAAGGCCACCGGGGTATTGAACGCC
H1      ..........G...T.......G......................
H2      ..........G.C.T.......G......................
H3      ..........G...T.......G.........A............
H4      .........GG...T.......G.........A............
H5      .............................................
H6      ...........................C.................
H7      ...............C..G....A.....................
H8      ...................................GAA.......
H9      ..........................T..................
H10     .................A...........................
H11     ........................A........T...........
H12     ...........G.....................T...........
H13     ...........G.................................
H14     ....G..T...G.............A...................
H15     .......................................C.T..T
H16     C..T.......G...................A.......C.TT..
H17     C..T.......G....C......................C.T...
H18     C..T.......G....C..G.G........A........C.T..T
H19     C..T.......G....C..G.G........A........C.T.TT
H20     .GT..GC....G.A......A.............A...A.G....
H21     .GT..GC.A..G.A......A.............A...A.G....
H22     .GT..GC....G................TT....A.....G....
```

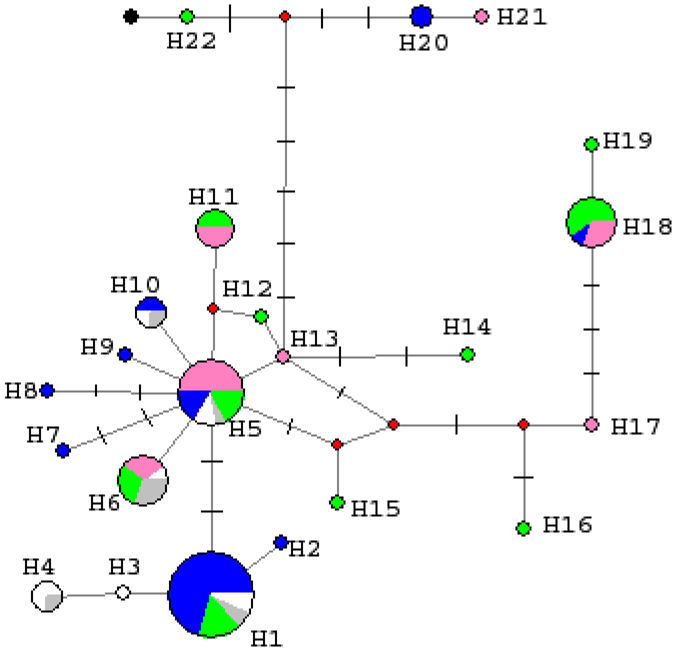


Table 1.16d), Haplotypes of *SbeIIb* in *O. sativa*

	IIIIIIIIEEIIIIIIIIIIIIIIIIII	IIIIIIE	E
	1111111111111111111111111111	1222222	2
	1111111122234444444444444446	6000001	2
	GTTGTTAGAGGTCTAATGGCTGGTTCG+65	CACGA+37	C
H1	...........C..............A.	......	.
H2	...........C..............A.	.....-	.
H3	...........C..............A.	.T...-	.
H4	...........C..............A.	.T....	.
H5	...........C..............A.	...C..	.
H6	...........C..............A.	....C.	.
H7	...........CT.............A.	......	.
H8	...........C........C.....A.	......	.
H9	...........C..............A.	..A...	.
H10	...........C.............T..	......	.
H11	...........C................	......	G
H12	...........C................	T.....	.
H13	.............C...........T..	......	.
H14	.............C............A.	......	.
H15	.............C............A-	......	.
H16	......G......C............A-	......	.
H17	.............C............A-	.....-	.
H18	.............C..............	T....-	.
H19	.CA.....C.....GGG.AG.A...T..	......	.
H20	.CA.....C.A...GGG.AG.A......	......	.
H21	.CA.....C.....GGG.AG.A......	T.....	G
H22	ACA....AC........TAG.....T..	......	.
H23	ACAACG.ACT....G.GTAG..T..T..	..A...	G
H24	ACAACG.ACT....G.GTAG..T..T..	..A..-	G
H25	ACAAC...CT....G.GTAG..T..T..	..A...	G
H26	ACAACG.ACT....G.GTAG..T..T..	..A.C.	G
H27	ACAACG.ACT....G.GTAG..TG....	..A...	G
H28	ACAACG.ACT....G.GTAG..T.C...	..A...	G
H29	ACAACG.ACT....G.GTAG..T...A.	..A...	G
H30	ACAACG.ACT....G.GTAG..T...A-	..A...	G
H31	ACAACG.ACT....G.GTAG..T..T..	..A...	.
H32	ACAACG.ACT....G.GTAG..T.....	......	.

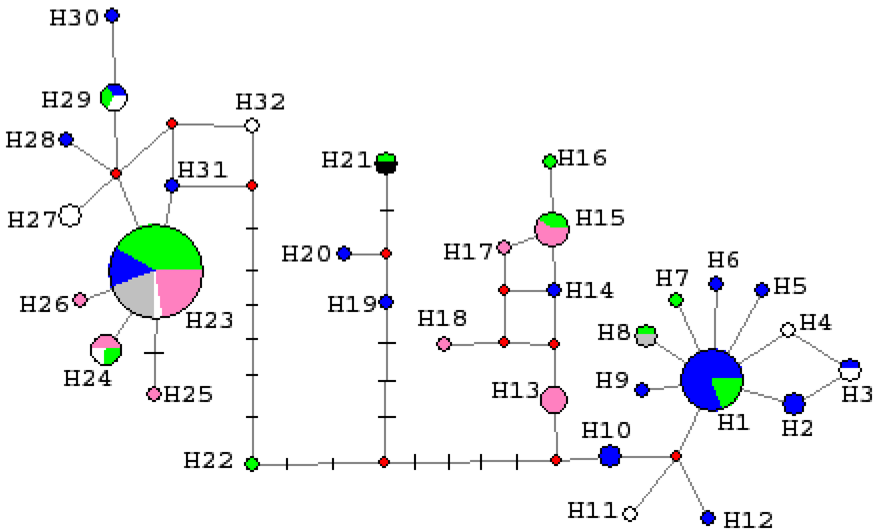


Table 1.16e), Haplotypes of *Iso1* in *O. sativa*

```
        EEEIIIIIIIIIIIIIIIIIIIIIIIIIIIIIEIIEEE
                   111111111111111111111111111
        22333334444000000011111112222222333678

        * *                                 *
        ACAGGGCCGACATGGCCAGCAAGCACGTTCAACGTTTG
H1      ...T......T.C..............G.....AGCC.
H2      ...T......T.C............T.G.....AGCC.
H3      ...T........C..............G.....AGCC.
H4      ...T...A....C............T.G.....AGCC.
H5      ...T......T.C..............G........C.
H6      ..........T....................G......
H7      ...............................G......
H8      .....C......................A.CG......
H9      ...............A...............G......
H10     ..........................C....G......
H11     .......A.......................G......
H12     ......................................
H13     .....................................T
H14     ...............................G.....T
H15     .................G.............G.....T
H16     ..........T....................G.....T
H17     ...........G......CT..................
H18     .......A...G......CT..................
H19     .......A...G.A....CT..................
H20     .......A...G......CT..............GC..
H21     ....A..A...G......CT..................
H22     .......A.G.G......C...................
H23     ......TA......A..............T.......T
H24     ......TA..T...A..............T.......T
H25     ......TA..T...A.....CTTTG....T.......T
H26     GG......A.T.C...T...CTTTG..G....TAGC..
H27     GGC.........C...T...CTTTG..G....TAGC..
```

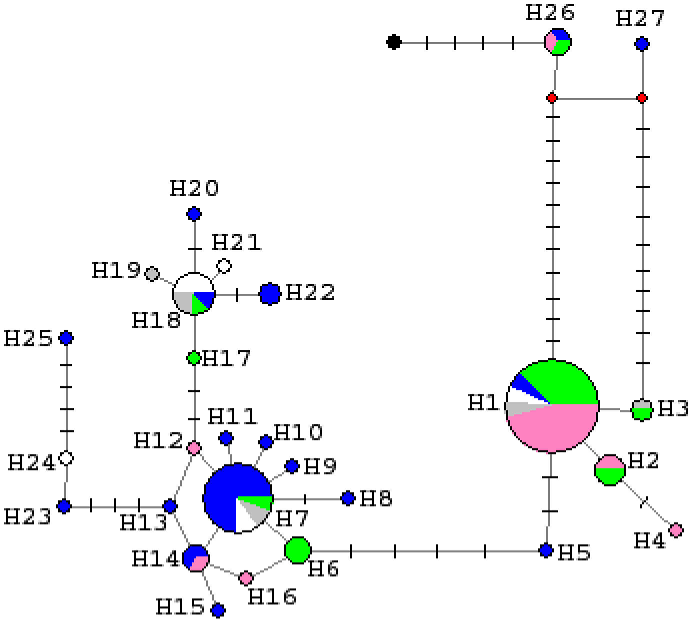


Table 1.17), Haplotype frequency difference between rice variety groups

	aus	***indica***	***tropical japonica***	***temperate japonica***
Bt2				
indica	0.69			
tropical japonica	0.64	0.33		
temperate japonica	0.86	0.47	0.73	
aromatic	0.75	1.00	0.86	0.88
Iso1				
indica	0.26			
tropical japonica	0.75	0.88		
temperate japonica	0.58	0.08	0.27	
aromatic	1.00	0.27	0.88	0.58
SbeIIb				
indica	0.57			
tropical japonica	0.57	0.98		
temperate japonica	0.88	0.93	0.33	
aromatic	0.44	0.04*	0.44	0.28
Sh2				
indica	0.72			
tropical japonica	0.48	0.88		
temperate japonica	0.74	0.92	0.78	
aromatic	0.89	0.33	0.74	0.87
SsIIa				
indica	0.17			
tropical japonica	0.70	0.67		
temperate japonica	0.93	0.42	0.57	
aromatic	0.78	0.76	0.64	0.27

Haplotype of studied starch genes in *O. rufipogon*; The heads of tables show the location of the polymorphic nucleotide. I3 means intron 3, E11 means exon 11. The star (FUSHENG WEI) means nonsynonymous mutation; the first sequence is the reference sequence; the following is the list of haplotypes. The dot (.) means the nucleotide is the same as reference sequence. The dash (-) and the plus (+) means deletion and insertion. The number following them is the number of nucleotide deletion or insertion. S is the repeat of AGA, 1 S means AGA, 2S means AGAAGA.

Table 1.18a), Haplotype of *Bt2* in *O. rufipogon*
Table 1.18b), Haplotypes of *Iso1* in *O. rufipogonf*
Table 1.18c), Haplotype of *SbeIIb* in *O. rufipogon*
Talbe 1.18d), Haplotypes of *SsIIa* in *O. rufipogon*
Table 1.18e), Haplotypes of *Sh2* in *O. rufipogon*

Table 1.18a), Haplotype of *Bt2* in *O. rufipogon*

```
        IIIIIIIIIEEEEIIIIIIIIIIIIIIIIIIIIIIIEEEIIIEEIIIIIIIIIIIIIIIIIIEIII
        222222222333333333333333333333333333444444555888888888888888889999
                   *
        TGCACA-AGGGGGACCTGAGGTTGGACATGAAGGCCAACGCAGTGAACCATT-AGCGCACGCCGAG
Hap_1   .............................A.......C............................
Hap_2   .......................C.....A.......C............C...............
Hap_3   .....G...........A........T............A...........A.T...........A
Hap_4   C.T....................C......................C........T..........
Hap_5   .............G...A.T.............................................A
Hap_6   C.T....................C...G..............A...C.......AT......T...
Hap_7   C.T....................C...G..............A...C........T......T...
Hap_8   .................A...C...G......A.................................
Hap_9   .............G...A.T...........G..................................
Hap_10  .................A.........................................G......
Hap_11  .................A......................T...........A.............
Hap_12  .................A...............A.......................A.....A..
Hap_13  .................A..........................................A.....
Hap_14  .................A.................................C..............
Hap_15  .T...G...........A........T........T...............A.T............
Hap_16  .................A.................................C............T.
Hap_17  .................A.T................................A.............
Hap_18  ...........................................C......................
Hap_19  ...........................................C...T..........T.......
Hap_20  ..........................................A...C........T......T...
Hap_21  ..........................................A.......................
Hap_22  C.........................T..........................T...........A
Hap_23  C.........................T...................C........T..........
Hap_24  .................A................................................
Hap_25  .............................A.............C......................
Hap_26  .............................A.............C................A.....
Hap_27  .....G........A..A........T........................A.T...........A
Hap_28  .................A..A.C...........................................
Hap_29  ......T.....................................................A.....
Hap_30  ...G....T.........................A...............................
Hap_31  ....T....A..........................T........G...-..........A.....
Hap_32  ....T....A...................A......T........G...-................
Hap_33  ...G....T.........................A.............T.................
Hap_34  .................A.........................C......................
Hap_35  ................C............A.............C..................T...
Hap_36  .............................A........T....C......................
Hap_37  C.T.........C...........A..G...........................T..........
Hap_38  ..T........CC...........A..G................A.C........T......T...
Hap_39  ............CG...A.............G..................................
Hap_40  .............G...A.............G..................................
Hap_41  ............CG...A.T...........G..................................
Hap_42  C.T....G...............C...G..............A...C........T......T...
Hap_43  C.T...........................................C........T..........
Hap_44  C.T.......A.C.................................C........TT.........
Hap_45  C.T.......A...................................C........TT.........
Hap_46  ...........CC..G..G........................C................A.....
Hap_47  ...........CC..............................C................A.....
Hap_48  C..........CC..........................................T....A.....
Hap_49  ...........CC..........................................T....A.....
Hap_50  C...........C..........................................T....A.....
```

```
Hap_51  .......................................................T....A.....
Hap_52  C...........C..........................................T....A.T...
Hap_53  ...........................G..............A...C........T......T...
Hap_54  C...........C..............G.............GA...C........T......T...
Hap_55  C.T.........................G..........................T..........
Hap_56  ..T.........................G............G.............T..........
Hap_57  C.T........................G..............A...C........T......T...
Hap_58  C.T.......................................A...C........T......T...
Hap_59  ......................................T....C......................
Hap_60  .................A......A...................A.................T...
Hap_61  .................A......A...................A.....................
Hap_62  C.T.........C..............G...........................T......T...
Hap_63  ..T.....................A..G................A..........T......T...
Hap_64  C.T....................C......................C........T.....G....
Hap_65  C.T....................C...G..G...............C........T......T...
Hap_66  C.T....................C...G..G...............C........TT.........
Hap_67  C.T...........................................C........TT.........
Hap_68  C.T........................G..............A...C........T..........
```

Table 1.18b), Haplotypes of *Iso1* in *O. rufipogon*

```
     IIIIIIIIIIIIII   I    IIIIIIIIIIIIIIIIIIIIIIIIIIEEEIIIIIIIIIIII    IIIIIIIIIIIIIIIIIIIIIIIIIIEIIIIIIIIIIIIII   IIIIIIEIIIIIIIIEEEIIIIIIEEEEEEE
                                                     111111111111111    11111111111111111111111111111111111111111   1111111111111111111111111111111
     12333333333333   3    33333333444444444444999999000000000111111    11111111111111122222222222333333333333333   3333334444444446677777778888888
                                                     **                                           *                       *          *        * **
H1   GTT-CCGTGTT-A(0S)(54-)TTTT--ACTGCACGTGCTTACGACCCGCCAACACACCGGA(11-)CCCTACGCGACCATAACGGTTTGATACTCACAATCCGTGT(5-)GGGCGTTAG-TTTTTCCTCCA-TCGTGGACA
H2   ..--.......-T(0S)(54-)....--..G.TG........T........G..G.......(11-).................T...G.........G........(5-)..A..C...T...........-A........
H3   ...T.......-.(0S)(54-)....--G..............................A..(11-).....................G......T..G........(5-).........-.......C...-......C..
H4   ..--.......-T(0S)(54-)....--..G.TG........T........G..G.......(11-)....-............T...G.........G........(5-)..A..C...-...........-A........
H5   ...-....A..-.(0S)(54-)....--...........................T...A..(11-).....................G........TG........(5-)..A......-...........-.........
H6   ...-..A....-.(0S)(54-)....--.............................A....(11-).....................G....T....G......A.(5-).........T..........G-..A......
H7   ...T.......-.(0S)(54-)...---..................................(11-)..............G......G.........G........(5-).........T.......C.-GA......C..
H8   ...T.......-.(0S)(54-)...---.........A..................T.....(11-)..............G......G.........G........(5-).........T.........-.A.........
H9   ...T.......-.(0S)(54-)...--C............C............T........(11-).......................-.G...G.G........(5-).T.......-.....T.....-.........
H10  ...-.......-.(3S)(54-)...---.........A.........T........T.....(11-)..............G......GA........G........(5-)A........-.-........G-.........
H11  ...T.......-.(0S)(54-)...--C......A.....C...........GT......C.(11-).......T...............-.....G.G........(5-).T.......TG....T.....-........G
H12  ...T.......-.(0S)(54-)...--C......A.....C...........GT......C.(11-).......................-.....G.G........(5-).T.T.....T.....T.....-.........
H13  ...-.......-.(1S)(54-)...---.........A..................T.....(11-)....-.........G......GA........G........(5-)A........-.-........G-.T.......
H14  ...-..A....-.(0S)(54-)....--.............................A....(11-)...C.................G....T....G......A.(5+).........T..........G-..A......
H15  ...T......G-T(0S)(54+)....--..G..............T................(11-).................TT..G.........G........(5-).........T...........-A........
H16  ...T.......-.(1S)(54-)....--.T.............T..................(11-).....................G.........G.AT-.C..(5-).........-.......C...-.........
H17  ...T.......-.(0S)(54-)...--C............C............T........(11-).......................-.G...G.G........(5-).T.......-.....T.....-.....T...
H18  ...-.......-.(2S)(54-)....--.........A...........T...T..T.....(11-)..............G.....GG.........G........(5-).........T.-........G-.T.......
H19  ...-.......-.(2S)(54-)....--.....................T...T..T.....(11-)..............G.....GG.........G........(5-).........T.-........G-.T.......
H20  ...-.......-.(2S)(54-)....--..................................(11-)G..............................G........(5-).........T......T....-....A..A.
H21  ...T.......-.(2S)(54-)...---.........A..................T.....(11-)..............G......GA........G........(5-)A........-.-........G-.........
H22  ...T.......-.(0S)(54-)..---C...A.....A..................T.....(11-)..............GCT....G.........G........(5-).........T.-........G-.........
H23  ...T.......-.(0S)(54-)..---C..................................(11-)..............GCT....G.........G........(5-).........T.-........G-.........
H24  ...-.......-.(0S)(54-)....--............C.....................(11-).T.........T...................G........(5+).........T...-.......-.........
H25  ...T.......-.(0S)(54-)...---.........A..................T.....(11-)......A.......G......G.........G........(5-).........-.-........G-.........
H26  ...T.......-.(2S)(54-)...---.........A...........T......T.....(11-)..............G.....GG.........G........(5-).........-.-........G-.T.......
H27  ...T.......-.(2S)(54-)...---.........A...........T......T.....(11-)..............G.....GG.........G........(5-)........A-.-........G-.T.......
H28  ...T.......-.(0S)(54-)...---.........A..................T.....(11-)......A.......G......GA........G........(5+)A.....C..T.-........G-.........
H29  ...-.......-T(0S)(54-)....T-..G..........C..G.................(11-)....-T...........T...G..C......G........(5-).........-...........-A........
H30  ...-.......-T(0S)(54-)....T-..G..........C..G.....T...........(11-)....-T...........T...G..C......G........(5-).........-.-.........-A........
H31  ...T.......-.(2S)(54-)...---.........A...........T......T.....(11-)..............G.....GG.........G........(5+).........T.-........G-.T.......
H32  ...-.....G.-.(0S)(54-)....--...........................T...A..(11-).....................G.........G........(5-).........T......T....-.........
H33  A..-...G...T.(0S)(54-)-...T-.........A.C......................(11+).....................G.....C..G........(5-).........T.........TG-.........
H34  ...-.......-.(1S)(54-)...---.........A..................T.....(11-)..............G......GA........G.......A(5-)A........T.-........G-.T.......
H35  ...T.......-.(0S)(54-)...---.........A..................T.....(11-)......A.......G......G.........G....A...(5-).........T.-........G-.........
H36  ...-.......-.(0S)(54-)...---.........A..................T..A..(11-).....................GA........G........(5-)A........T.-......G.G-.........
H37  ...-.......-.(0S)(54-)....--............C..................A..(11-)...............................G........(5-).........T...-..T....-.........
H38  ...-....A..-.(0S)(54-)....--.........A.....C............T.....(11-)................T....G.........G........(5-).........T.-........G-.........
H39  ...-....A..-.(0S)(54-)....--.......A.A.....C............T.....(11-)................T....G.........G........(5-).........T.-........G-.........
H40  ...-.......-.(0S)(54-)...---...A.....A..................T.....(11-)...............CT....G.........G........(5-).......-.T.-........G-.........
H41  ...-.......-.(0S)(54-)...---...A....AAT.................T.....(11-)................T....G.........G........(5+).........T.-........G-...C.....
H42  ...-.......-.(0S)(54-)...---.........A..................T..A..(11-).....................G.........G....A...(5-).........T.-........G-.........
H43  ...-.......-.(0S)(54-)....--...............................A..(11-).....................G......T..G........(5-).........T...........-.........
H44  ...-.......-.(0S)(54-)....--G..............................A..(11-).....................G......T..G........(5-).........T.......C...-.........
H45  ...-.......-.(0S)(54-)....--G..............................A..(11-).....................G......T..G........(5+).........T...........-.........
H46  ...-.......-.(0S)(54-)....--.........................T.....A..(11-).............................G.G........(5+).T.......T.....T.....-.........
```

```
H47  ...-.......-.(0S)(54-)....--....................A....T.....A..(11-)...............................G........(5+)....A....T...........-.....T...
H48  ...-.......-.(0S)(54-)....--G..............................A..(11-).....................G......T..G........(5-).........T.......C...-......C..
H49  ...-.......-.(0S)(54-)....--G..............................A..(11-).....................G......T..G........(5+).........T.......C...-......C..
H50  ...-.......-.(0S)(54-)...---G..............................A..(11-).....................G......T..G........(5+).........T...........-......C..
H51  ...-.......-.(0S)(54-)...---G..............................A..(11-).....................G......T..G........(5+).........T.......C...-......C..
H52  ...-.......-.(0S)(54-)....--.........................T........(11-)..............G......G.......G.G........(5+).T.......T.....T.....-......C..
H53  ...-.......-.(0S)(54-)....--G........................T........(11-)..............G......G.-....T..G........(5+).........T.......C...-......C..
H54  ...-.......-.(0S)(54-)....--G.................................(11-).....................G......T..G........(5+).........T..G....C...-......C..
H55  ...-.......-.(0S)(54-)....--G.................................(11-).....................G......T..G........(5+).........T.......C...-......C..
H56  ...-.......-.(0S)(54-)....--G..............................A..(11-).....................G......T..G........(5+).........T..G....C...-......C..
H57  ...-.......-.(0S)(54-)....--..................................(11-).....................G......T..G........(5-).........T...A...C...-......C..
H58  ...-.......-.(0S)(54-)....--G.................................(11-).....................G......T..G........(5-).........T.......C...-......C..
H59  .C.-A...A..-.(0S)(54-)...---.........................T..T.....(11-).....................G.......G.G........(5+).T.......T...........-.........
H60  ...-A...A..-.(0S)(54-)...---.........................T..T.....(11-).....................G.-.......G........(5+).T.......T...........-.....T...
H61  ...-.......-.(0S)(54-)....--..................T.........T..A..(11-).....................G.........G........(5+).........T....-..C...-.........
H62  ...-.......-.(0S)(54-)....--............................T.....(11-).....................G.........G........(5+).........T....-..C...-.........
H63  ...-.......-.(0S)(54-)....--G.............................T..G(11-)..T.....A.T.CC.......G......T..G........(5+).........T...........-.........
H64  ...-.......-.(0S)(54-)....--G.............................T...(11-)..T.....A.T.CC.......G......T..G........(5+).........T...........-.........
H65  ...-.......-.(0S)(54-)....--G..............................A..(11-).........C...........G......T..G........(5+).........T...........-.........
H66  ...-.......-.(0S)(54-)....--....................A..........A..(11-).........C.....................GG.......(5+)....A....T...........-.......A.
H67  ...-.......-.(0S)(54-)....--............C............T.....A..(11-).........C.............-.......GG.......(5+).T..A....T.....T.....-.........
H68  ...-.......-.(0S)(54-)....--....................A....T.....A..(11-).........C.............-.......G........(5+).T..A....T...........-.....T...
H69  ...-.......-.(0S)(54-)....--...............................A..(11-).........C.....................G........(5+).........T...........-.......A.
H70  ..--.A.....-.(0S)(54-).G..--G.................................(11-)..............G......G......T..G........(5+).........T.......C...-......C..
H71  ..--.......-.(0S)(54-)....--G.................................(11-)..............G......G......T..G........(5+).........T.......C...-......C..
H72  ..--.......-.(0S)(54-).G..--G.................................(11-)..............G......G......T..G........(5+).........T.......C...-......C..
H73  ..--.......-.(0S)(54-).G..--G.................................(11-)......A.......G......G......T..G........(5-).........T.......C...-......C..
H74  ..--.......-.(0S)(54-).G..--G.................................(11-)......A.......G....G.G......T..G........(5-).........T.......C...-......C..
```

Table 1.18c), Haplotype of *SbeIIb* in *O. rufipogon*

```
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIEEEIIEEIIIIEIIIIIIII IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIEIIIIEEIIIIIIII  IIIIII I IIIIIIII  IIIEEE
        1111111111111111111111111111111111111111111111111111 1111111111111111111111111111111111111111111111122222222222  222222 2 22222222  222222
        1111111111111111111111111111111122222333333444444444 4444444444444444444444444444444444444444446666600000000000  000000 0 00000000  111222
                                         *         *
        TTAGAGATGGGCACGGGG-A-TTT-TCACGACCGGGTCACGTCGTTCGGTG+9GGAATA-GGGAGTCCCGTTGGGTCACGATGAGT+4GCTTCGCCTCGTTTTACCAGT9 CGTCG+2-2GTGTCGT+37CGAGAC
Hap_1   ...........TT...................A........C........A..T....G..C..................................T...........T... ...... . .............C..
Hap_2   ...........TT...................A........C........A. T....G..C.......................... ...................T..  ...... . ........  ...C..
Hap_3   ............T.....................................A. T..GA.T....................A....A.. ...........T..........  ...... . .......-  ...C..
Hap_4   ..........A......A...C.GA....A....T.................  T. ..................T............. ....T.................  A....- . ........  -.....
Hap_5   ...........TT...................A........C........A. T....G..C.......................... .....A................  ...... . ...A....  ...C..
Hap_6   ...........TT...................A.....G...........A. T....G..C.......................... .....A................  ...... . A.......  ...C..
Hap_7   G.................................................A. T......A...................T....... ......................  ...... . ...A...-  ......
Hap_8   ...........TT...................A........C........A. T....G..C.......................... .....A................  ...... . .......-  ...C..
Hap_9   ...........TT...................A........C........A. T....G..C.......................... ..G.............T.....  ...... . ...A..C-  ...C..
Hap_10  ......G....TT.............T....TA.................A. T....G..C.......................... .....A................  ...... . .......-  ...C..
Hap_11  .............................................C...... ................................... ......................8 ...... . ...A....  ...C..
Hap_12  .............................................C...... ................................... ...............CT.....8 ...T.. . ...A....  ....G.
Hap_13  ..G.........T...............T.......-.............A. T..G..T............................ A.........T...........  ...... + .......-  ...C..
Hap_14  ........................A........................... .........C......................... ................T.....  ...... . ...A....  ...C..
Hap_15  ........................A........................... .........C......................... ................TA....  ...... . ...A....  ..TCG.
Hap_16  ...........TT.C.................A.................A. T....G..C.......................... .....A................  ...... . .......-  ...C..
Hap_17  ..............C.................A.................A. T....G..C.......................... .....A................  ...... . .......-  ...C..
Hap_18  ..................................................A. ....................A.............. ..G.............T.....  ...... . ...A..C-  ...C..
Hap_19  ...........TT...................A........C....T...A. T....G..C.......................... .....A................  ...... . ........  ...C..
Hap_20  ...........TT...................A........C.......... .................A................. .....A................  ...... . ........  ...C..
Hap_21  ...A...CA.......T.T-.......................T........ ..-...................A............ ...C..................  ...... . ....T...  ......
Hap_22  .........A.......................................... ........................T.......... .......T........T.....  ...... . ...A....  ...C..
Hap_23  ...............................................A.... .................A................. .........A..G....A....8 .A.... . .G......  ......
Hap_24  ...........TTA..................A.................A. T....G..C.....................A.... .....A........C.......  ...... . ........  ...C..
Hap_25  ........A...............A........................... ................T.................. .T....................  A....- . ........  -.....
Hap_26  ........A...............A........................... ................T..............G... .T....................  A....- . ........  -.....
Hap_27  .A...................................T....T.......A. T..........................A....... ......................  ...... . ........  ...C..
Hap_28  .A...................................T....T.......A. T.................................. ......................  ...... . ........  ...C..
Hap_29  ............................................CC...... ...........A.T..................... .......T..............8 ...... . ........  ...C..
Hap_30  ............................................CC...... ...........A.T..................... .......T........TA....8 ..C... . ........  ......
Hap_31  ............T...............................CC....A- T.........G.......................- ...............C......8 ...T.. . ...A...-  ......
Hap_32  ...A...CA.......T.T-................................ ..-...................A............ .........A..G.........  ...... . ....T...  ......
Hap_33  ...........TT...................A........C........A. T....G..C.......................... .....A................  ...... . ........  ...C..
Hap_34  ....T...............A............................... ................T......C........... ........T........AA.A.10...... . ........  ......
Hap_35  .................................................... ...............................G... .............-........  ...... . ........  ......
Hap_36  .....T.............................................. ...............................G... .............-...A....  ...... . ........  ......
Hap_37  ...........TT...................A........C........A. T....G..C.......................... .....A................  ...... . ........  ...C..
Hap_38  ..................................................A. ....................A.............. ..G.............T.....  ...... . ....T...  ......
Hap_39  ...A...CA.......T.T-................................ ..-...................A............ ...C..................  ...... . ....T...  ......
Hap_40  ..................................................A. T.................................. ................T.....  ...... . ........  ...C..
Hap_41  ............T.....................................A. T.................................. ......................  ...... . ........  ...C..
Hap_42  ..................................................A. T....G............................. .....A................  ...... . ........  ...C..
Hap_43  ..................................................A. ................................... ................T.....  ...... . ........  ...C..
Hap_44  ......................C..A.......................... ............C.T.................... ......................  ...... . ........  .....T
Hap_45  ......................C..A........................A. T....G............................. .....A................  ...... . ........  ...C..
```

```
Hap_46   ...........TT.....................................A. T...........C.T................... .....................  ...... . ........  .....T
Hap_47   ..........A.............A.....G..A.....T............ .T............................G... ....T................8 ...... . ..T.....  -.....
Hap_48   ..........A......A................................A. .T..................T............. ....T................  ....A- . ........  -.....
Hap_49   ..........A.T.....................................A. T....G............................ .....A...............  .....- . ........  -..C..
Hap_50   ..........A......A...C............................A. .T..................T............. ....T................  .....- . ........  -.....
Hap_51   ..........A.............A..............T............ .T................................ ....T................8 ...... . ........  -.....
Hap_52   ..........A......A...........A...................... .T.......................T........ ....T................  .....- . ........  -A....
Hap_53   ..........A......................................... .T.......................T........ ....T................  .....- . ........  -.....
Hap_54   ..........A......A...C..A....A....T................. .T..................T............. ....T................  A....- . ........  -.....
Hap_55   ..........A.............A..............T.........G.. .T................................ ....T................8 ...... . ........  -.....
Hap_56   .................................................... .................C............G... ....T................  A....- . .....A..  -.....
Hap_57   ..........A......A.....G.....A...................... .T..................T............. ....T................  A....- . .....A..  -.....
Hap_58   ..........A......A...........A...................... .T..................T.........G... ....T................  A....- . .....A..  -.....
Hap_59   ..........A..........C.......A...................... .T..................T............. .....................  A....- . .....A..  -.....
Hap_60   ..........A......A...C.......A...................... .T..................T.........G... ....T................  A....- . .....A..  -.....
Hap_61   ...........................T........................ ....................T...........C. .....................  A....- . ........  -.....
Hap_62   .................................................... ....................T...........C. .....................  A....- . ........  -.....
Hap_63   ..........A......A.................................. .T..................T............. ....T................  .....- . ........  -.....
Hap_64   .................................................... .................................. ...............T.....  ...... . ........  ...C..
Hap_65   ............T............................C........A. T....G............................ .....A...............  ...... . ........  ...C..
Hap_66   ......................C..A...............C.......... T...........C.T................... .....................  ...... . ........  .....T
Hap_67   ......................C..A...............C........A. T....G............................ .....A...............  ...... . ........  ...C..
Hap_68   ...........TT......................A.....C.......... T...........C.T................... .....................  ...... . ........  .....T
Hap_69   ...............A........A...............A.......A... ..................C.....C..CC.G... .....................  A....- . ........  -.....
```

Talbe 18d), Haplotypes of *SsIIa* in *O. rufipogon*

```
        I  IIIIII IIIIIIIIIIIIIIIIIIIIIIIIIIIIIEEEEIIIIIIIIIIIIIIEII II   IIII  IIIIIIIIIIIIIIIIIIIIIIIII IEEEEEEEEEEEEEEEE
        3  333333 3333333333333333333333333333344444444444666666677777 77   7777  7777777777777777777777777 78888888888888888
                                                                                                             * *   *
Hap_1   TTCGCGTA-7C--ACGCCGCGCGACGC-TT-GCGACTGTCTGGGACGGTAAACACACCA+8GATAACCTTTTTGCCTG-TCGCATGTGGGTAGGGGCCCCGGGGGCCGAAGGCGG
Hap_2   ........-7.TA....G.....G...-.--.......G......A....T..G.....+8..C...AA---A.....-.....C.........A.--T.A..............
Hap_3   ........-7.TA...A........A.-.--.......G....................+8.................-......A.A......A.--T......A......T..
Hap_4   ........-7.TA....G.....G...-.--.......G....................+8.................-......A.A......A.--T......A......T..
Hap_5   ........-7.TA....G.....G...-.--.......G...........T..G.....+8..C...AA---A.....-.....C.........A.--T.A........G....A
Hap_6   ........-7.TA....G.....GT..-.--............................+8..............T..-...G..AA.......A.--T..........G.....
Hap_7   ........-7.TA............A.-.--............................+8..............T..-...G..AA.......A.--T..........G.....
Hap_8   ........-7.TA....G.T.T.....-.--...G...............T....C...+8.............A...-....C.A........A.--T..A..T....G...AA
Hap_9   ........-7.TA....G.....GT..-..-........TC.A.T.....T.T......+8.................-.......A..A....A.--T..........G.....
Hap_10  ........-7.TA.A..G.....GT.T-.--........TC.A.T..---.........+8..............T..-..AG..A......A.AT--T..........G.....
Hap_11  ........-7.TA....G....AGT..-..T.....G..TC.A.T.....T.......T+8.................-.T.....A..A....A.--T..........G.....
Hap_12  ........-7.TA....G....AGT..-..T........TC.A.T.....T.......T+8.................-.T.....A..A....A.--T..........G.....
Hap_13  ........+7.TA..A.G.....G...-..-............A......T........+8.................-.......A.......A.--T..........G.....
Hap_14  ---.....-7.TA....G.....G...-.--.......G...........T..G.....+8..C...AA---AA....-.....C.........A.--T...........T....
Hap_15  ........-7.TA...A........A.-..-.........................T..+8..............T..-C..A..A......A.AT--T..........G.....
Hap_16  ........-7.TA..............-..-............................+8.................-..A.................................
Hap_17  ......A.-7ATA....G..A..G...-..-.........C..A......T...T....+8................AT.................--T.......T..G.....
Hap_18  ....T...-7.TA....G.....GT..-.--CT.......C...T.....T........+8.................-.......A.AA....A.--T..........G.....
Hap_19  ........-7.TA....G.........-.--.........C...T.....T........+8.................-.......A.AA....A.--T..........G.....
Hap_20  ........-7.TA...AG.......A.-..-...............A............+8.................-......A......A.A.--T..........G.C...
Hap_21  ........-7.TA..A.G.......A.-..-...............A............+8.................-......A......A.A.--T..........G.C...
Hap_22  ........-7.TAT...GA........-..-...................T........+8.....A...........-......A......A.AT--T..........G.....
Hap_23  ........-7.TA....G......T..-..-..TG........................+8.................-......A........A.--T..........G.C...
Hap_24  .....A..-7.TA....G.....GT..-..-......T............T........-8A----............-......A...A..A.A.--T........CCG.....
Hap_25  ........-7.TA....G.........-..-...................T........+8.................-......A......AAA.--TT...............
Hap_26  .......T-7.TA....G.....GT..-..-........TC.A.T.....T........+8.................-.......A..A....A.--T....A.....G.....
Hap_27  ........-7.TA..A.G.....G...-..-............A......T........+8.................-......A......AAA.--TT...............
Hap_28  ........-7.TA....G.....G...-.--..........A........T......T.+8.................-........A......A.--T......A......T..
Hap_29  ........-7.TA....G.T.T.....-.--...G...............T....C...+8.............A...-....C.A........A.--T.....T....G...AA
Hap_30  ........-7.TA....G...T.....-.--....T.....A........T........+8.................-........A......A.--T......A......TAA
Hap_31  ........-7.TA....G...T.....-.--....T.....A........T........+8.............A...-....C.A........A.--T......A......TAA
Hap_32  ........-7.TA....G...T.....-.--....T.....A........T........+8.............A...-....C.A........A.--T......A...G..TA.
Hap_33  ........-7.TA....G...T.....-.--....T.....A........T......T.+8.................-........A......A.--T......A......TAA
Hap_34  ........-7.TA....G.....G...A---..........A........T......T.+8.................-........A......A.--T...A......G....A
Hap_35  ........-7.TA....G...T.....-.--...................T....C...+8.............A...-....C.A........A.--T......A......T..
Hap_36  ........-7.TA....G...T.....-.--...................T........+8.................-........A......A.--T......A......T..
Hap_37  ...-....-7.TA....G...T.....-.--..........A........T........+8.................-........A......A.--T......A......T..
Hap_38  ...-....-7.TA....G...T.....-.--..........A........T....C...+8.............A...-....C.A........A.--T......A......T..
Hap_39  ........-7.TA....G.T.T.....-.--...G...............T....C...+8.............A...-....C.A....C...A.--T.....T....G...AA
Hap_40  ........-7.TA....G.T.T.....-.--...G...............T........+8.............A...-....C.A....C...A.--T.....T....G...AA
Hap_41  ........-7.TA....G...T.....-.--..........A........TG.......+8.............A...-....C.A........A.--T.............T.A
```

```
Hap_42  ........-7.TA....G...T.....-.--..........A........T........+8.................-........A......A.--T...A.........T.A
Hap_43  ........-7.TA....G...T.....-.--..........A........T........+8.............A...-....C.A........A.--T.............T.A
Hap_44  ........-7.TA....G.T.T.....-.--...G...............T........+8.............A...-....C.A........A.--T.....T....G...AA
Hap_45  ........-7.TA....G...T.....-.--...................T......T.+8.................-........A...G..A.--T......A......T.A
Hap_46  ........-7.TA....G...T.....-.--...................T....C...+8.............A...-....C.A........A.--T......A......T.A
Hap_47  ........-7.TA....G.T.T.....-.--...G...............T........+8.............A...-....C.A........A.--T.....T....G....A
Hap_48  ........-7.TA..A.G.....G...-..-C..................T........+8.................-......A........A.--TT...............
Hap_49  ........-7.TA....G.....G...-..-C..................T........+8.................-......AA.AA....A.--T..........G.....
Hap_50  ........-7.TA....G.T.T.....-.--...G......A........T........+8.............A...-....C.A........A.--T.....TA...G....A
Hap_51  ........-7.TA....G.....G...-.--..........A........T......T.+8...............G.-........A......A.--T......A......T..
```

Table 1.18e), Haplotypes of *Sh2* in *O. rufipogon*

```
        EEEIIIIIIIIIIEEIIIIIIIIIIIIIIIEIEIIIEEIIIIEIIIIIIEEIIIIII
                                                  111111111111111
        333333333333344444444444444444556778999999001111122222334
         **          *                      **    *       *
        GGGTAAGGTCCCTGCGCCCTAATAGTACACTACCATCGGCGCTCCTTCCCCCCAACA
Hap_1   A...................................T....................
Hap_2   ....................................T....................
Hap_3   .........................................................
Hap_4   ...............A.................................T.......
Hap_5   .......T...T.......................C.....................
Hap_6   ...................C................T....................
Hap_7   ............C....T.C................T....................
Hap_8   .......T...T...........................G.................
Hap_9   .......T...T............................A................
Hap_10  .......T...T.............................................
Hap_11  .........T...............................................
Hap_12  ...............................-T...........A............
Hap_13  ..........................................C.......A....A.
Hap_14  ..............T..........................................
Hap_15  ...........T..T..........................................
Hap_16  .......T...T..T..........................................
Hap_17  ..A..........A....-...........C............T............T
Hap_18  .......................G.................................
Hap_19  ..A..........A....-..T........C............T............T
Hap_20  ...........T...A.........................................
Hap_21  ...........T...AT...................................G....
Hap_22  ...C............................T........................
Hap_23  ..........................T.G............................
Hap_24  .....G....T....................................G...A..G..
Hap_25  ...............AT........C.......................T.......
Hap_26  ........................................A................
Hap_27  AA.........TC...........................A................
Hap_28  .............A...........................................
Hap_29  .............A.............T...............T............T
Hap_30  ....................T.............T.............A........
Hap_31  .............................................C...........
Hap_32  ...C............................T........T...............
Hap_33  A..........T........................T....................
Hap_34  .............A.......T........C............T.............
Hap_35  ...........T.........T..................A................
Hap_36  ....G........A..........A..T..........T..................
Hap_37  .....G....T......................T...A.........G.....TG..
Hap_38  ......-..T......................T........................
Hap_39  .....G....T....................................G......G..
Hap_40  ...............................................G.........
Hap_41  .......................G.................T...............
Hap_42  .................................................T.......
Hap_43  ...........T.............................................
Hap_44  ..............................................C..........
Hap_45  ........A........................................T.......
Hap_46  ......................CG.................................
Hap_47  ........A................................................
Hap_48  ........A.....................................C..T.......
Hap_49  .......................G......................C..........
Hap_50  .........................................T.....G.........
Hap_51  .....G.........................................G......G..
Hap_52  .......................G.....T...........................
Hap_53  .............................T...........................
```

Table 1.19) The results of MLHKA tests in *tropical japonica*

Model	Description	lnL[a]	Comparison	Likelihood-ratio statistics (d.f)	P[b]	K[c]
A	neutral (all K=1)	-41.5428				
B	selection at *Sh2*	-41.3914	A vs. B	0.1514(1)	>0.05	1.37
C	selection at *Bt2*	-40.7597	A vs. C	0.7831(1)	>0.05	1.94
D	selection at *SsIIa*	-41.6383	A vs. D	-0.0955(1)	>0.05	1.28
E	selection at *SbeIIb*	-41.256	A vs. E	0.2868(1)	>0.05	1.46
F	selection at *Iso1*	-41.4092	A vs. F	0.1336(1)	>0.05	1.66
G	selection at *Wx*	-35.5304	A vs. G	6.0124(1)	<0.05*	0.13

[a] the likelihood value of the model

[b] the possibility of chi-square distribution

[c] the selection parameter of the gene designed as under selection in the model

Figure 1.1), Simplified starch synthesis pathway with all the studied starch genes

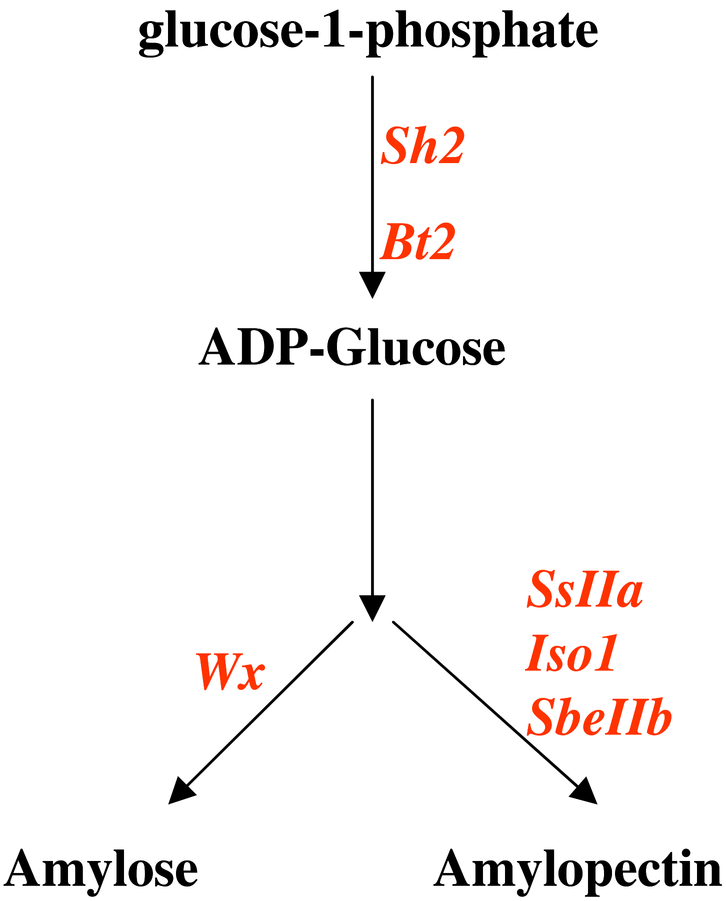

glucose-1-phosphate
Sh2
Bt2
ADP-Glucose
Wx
SsIIa
Iso1
SbeIIb
Amylose
Amylopectin

Figure 1.2), Location and size of the sequenced regions; The boxes represent exons; Shaded boxes correspond to translated regions; The lines connecting the boxes are introns; The gene size is indicated above the boxes; The sequenced regions and size are shown under the boxes.

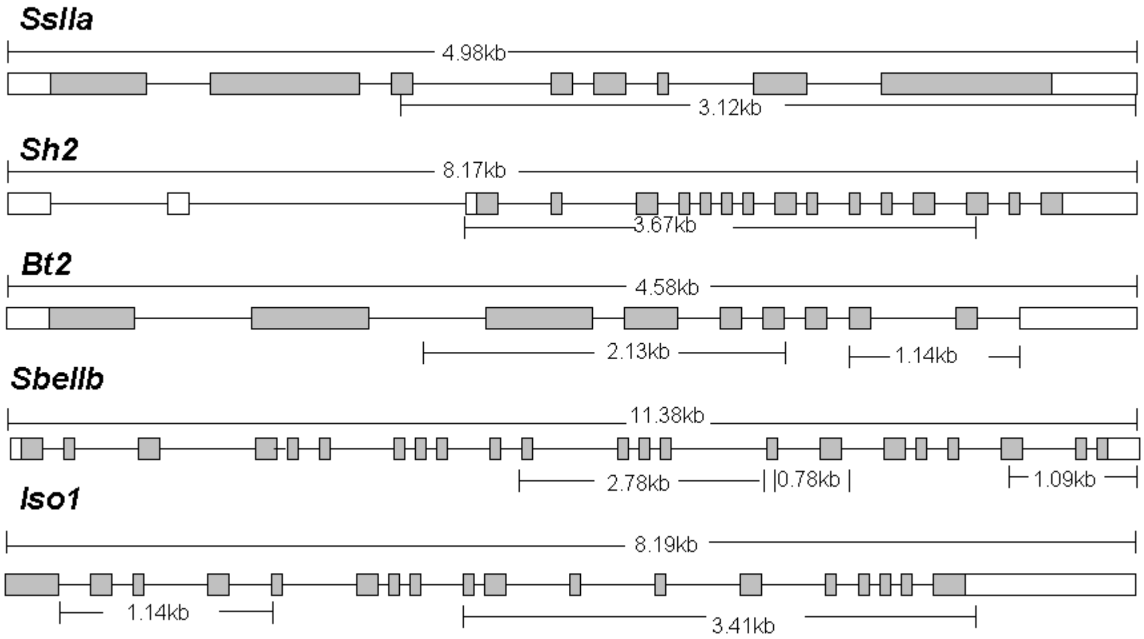
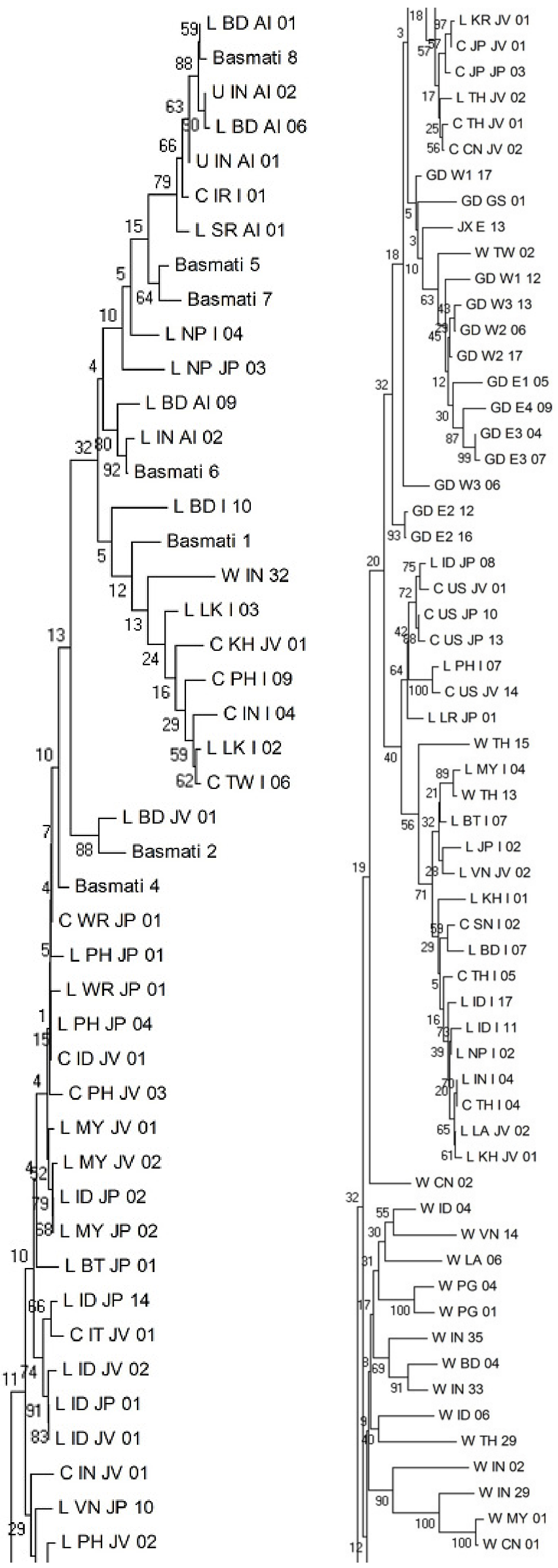

SsIIa
4.98kb
3.12kb
Sh2
8.17kb
3.67kb
Bt2
4.58kb
2.13kb
1.14kb
SbeIIb
11.38kb
2.78kb
0.78kb
1.09kb
Iso1
8.19kb
1.14kb
3.41kb

Figure 1.3), Neighbor joining tree of the studied samples based on five studied genes (no *Wx*). The number on the branches is the bootstrap support. The tree is spliced into three parts because the tree can not fit into one page.


L BD AI 01
Basmati 8
U IN AI 02
L BD AI 06
U IN AI 01
C IR I 01
L SR AI 01
Basmati 5
Basmati 7
L NP I 04
L NP JP 03
L BD AI 09
L IN AI 02
Basmati 6
L BD I 10
Basmati 1
W IN 32
L LK I 03
C KH JV 01
C PH I 09
C IN I 04
L LK I 02
C TW I 06
L BD JV 01
Basmati 2
Basmati 4
C WR JP 01
L PH JP 01
L WR JP 01
L PH JP 04
C ID JV 01
C PH JV 03
L MY JV 01
L MY JV 02
L ID JP 02
L MY JP 02
L BT JP 01
L ID JP 14
C IT JV 01
L ID JV 02
L ID JP 01
L ID JV 01
C IN JV 01
L VN JP 10
L PH JV 02
L KR JV 01
C JP JV 01
C JP JP 03
L TH JV 02
C TH JV 01
C CN JV 02
GD W1 17
GD GS 01
JX E 13
W TW 02
GD W1 12
GD W3 13
GD W2 06
GD W2 17
GD E1 05
GD E4 09
GD E3 04
GD E3 07
GD W3 06
GD E2 12
GD E2 16
L ID JP 08
C US JV 01
C US JP 10
C US JP 13
L PH I 07
C US JV 14
L LR JP 01
W TH 15
L MY I 04
W TH 13
L BT I 07
L JP I 02
L VN JV 02
L KH I 01
C SN I 02
L BD I 07
C TH I 05
L ID I 17
L ID I 11
L NP I 02
L IN I 04
C TH I 04
L LA JV 02
L KH JV 01
W CN 02
W ID 04
W VN 14
W LA 06
W PG 04
W PG 01
W IN 35
W BD 04
W IN 33
W ID 06
W TH 29
W IN 02
W IN 29
W MY 01
W CN 01

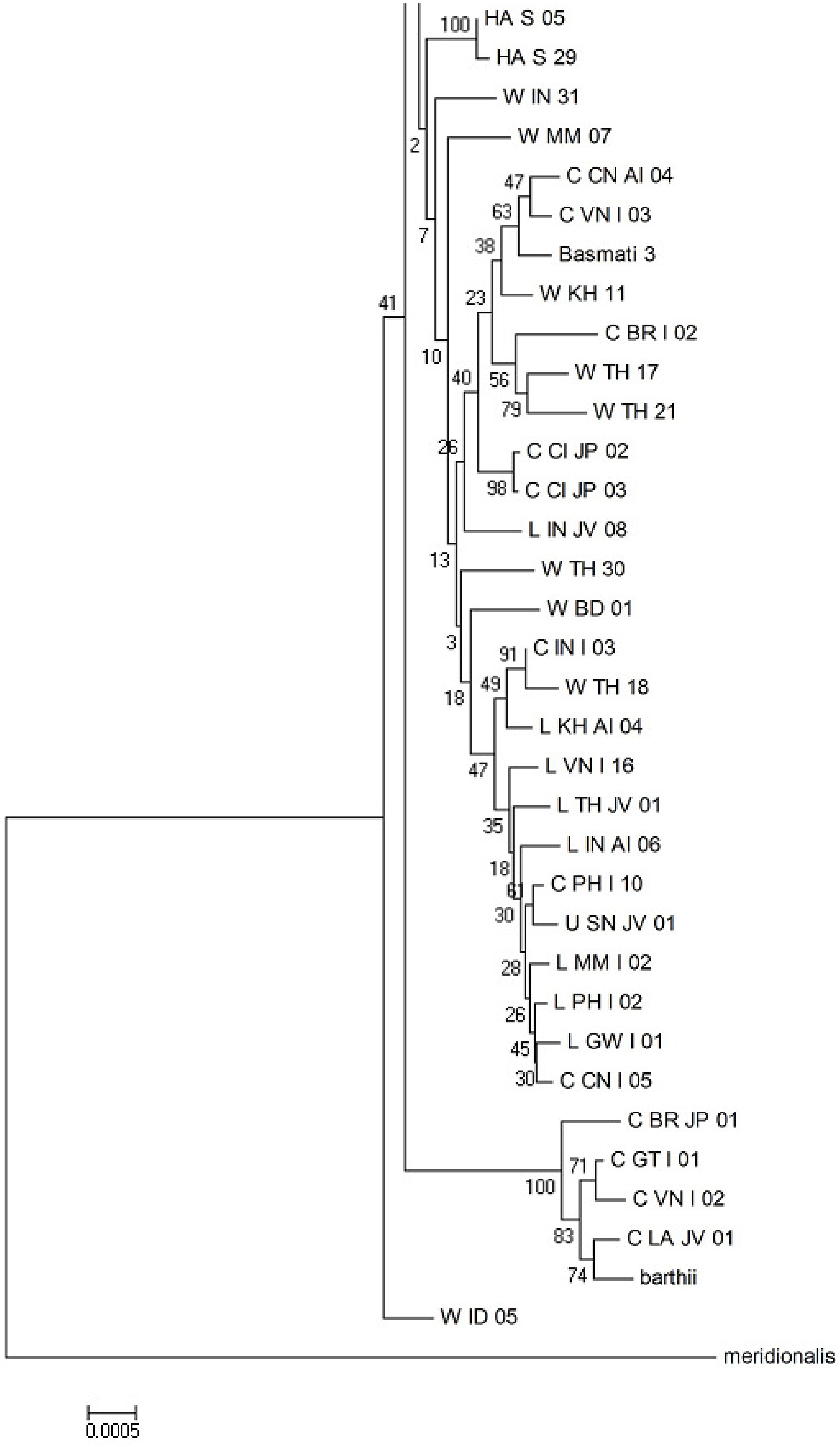
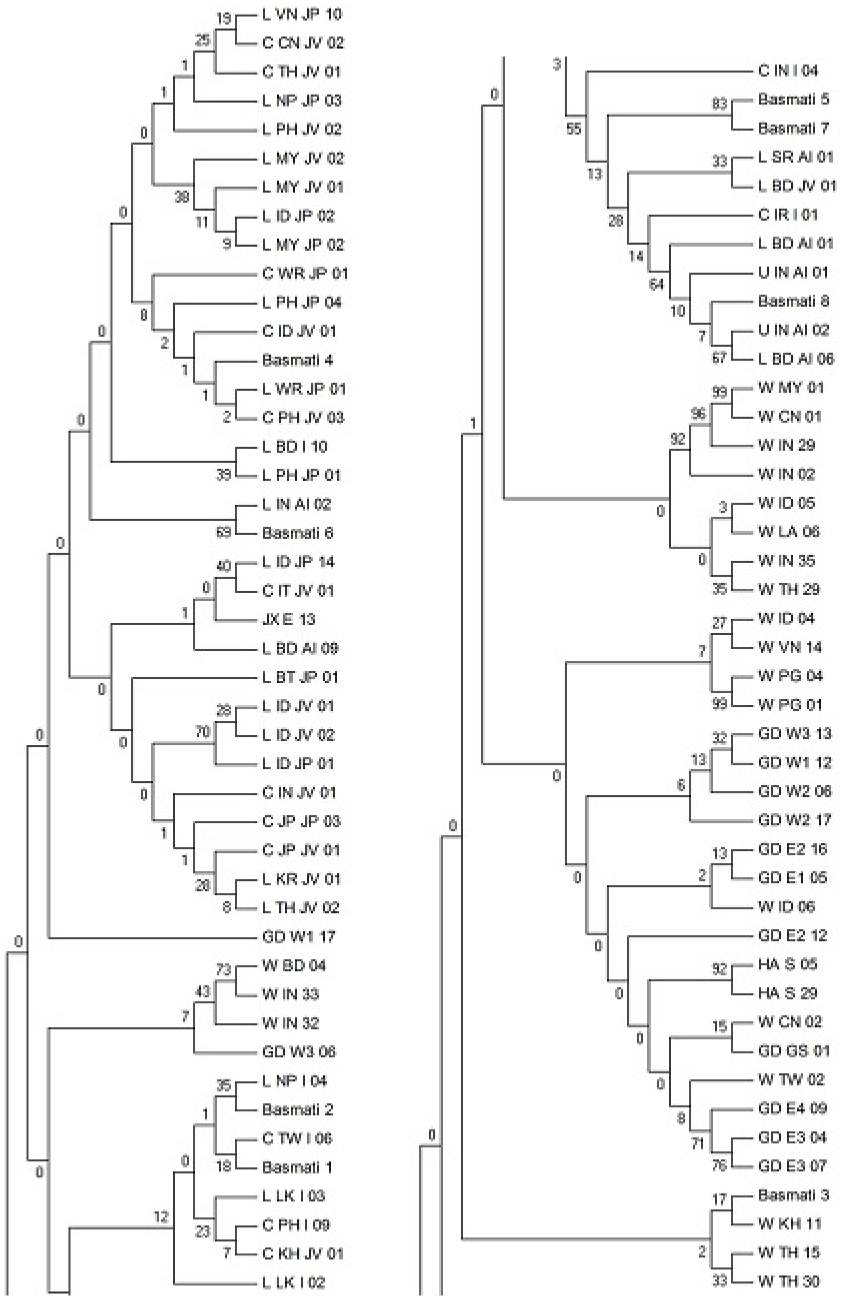

100
HA S 05
HA S 29
W IN 31
2
W MM 07
47
C CN AI 04
63
C VN I 03
38
Basmati 3
7
W KH 11
23
41
C BR I 02
10
56
W TH 17
40
79
W TH 21
26
C CI JP 02
98
C CI JP 03
L IN JV 08
13
W TH 30
W BD 01
3
91
C IN I 03
49
W TH 18
18
L KH AI 04
47
L VN I 16
L TH JV 01
35
L IN AI 06
18
61
C PH I 10
30
U SN JV 01
28
L MM I 02
26
L PH I 02
45
L GW I 01
30
C CN I 05
C BR JP 01
71
C GT I 01
100
C VN I 02
83
C LA JV 01
74
barthii
W ID 05
meridionalis
0.0005

Figure 1.4), Maximum parsimony tree of the studied samples based on five studied genes (no *Wx*). The number on the branches is the bootstrap support. The tree is divided into three parts because the tree can not fit into one page.


L VN JP 10
C CN JV 02
C TH JV 01
L NP JP 03
L PH JV 02
L MY JV 02
L MY JV 01
L ID JP 02
L MY JP 02
C WR JP 01
L PH JP 04
C ID JV 01
Basmati 4
L WR JP 01
C PH JV 03
L BD I 10
L PH JP 01
L IN AI 02
Basmati 6
L ID JP 14
C IT JV 01
JX E 13
L BD AI 09
L BT JP 01
L ID JV 01
L ID JV 02
L ID JP 01
C IN JV 01
C JP JP 03
C JP JV 01
L KR JV 01
L TH JV 02
GD W1 17
W BD 04
W IN 33
W IN 32
GD W3 06
L NP I 04
Basmati 2
C TW I 06
Basmati 1
L LK I 03
C PH I 09
C KH JV 01
L LK I 02
C IN I 04
Basmati 5
Basmati 7
L SR AI 01
L BD JV 01
C IR I 01
L BD AI 01
U IN AI 01
Basmati 8
U IN AI 02
L BD AI 06
W MY 01
W CN 01
W IN 29
W IN 02
W ID 05
W LA 06
W IN 35
W TH 29
W ID 04
W VN 14
W PG 04
W PG 01
GD W3 13
GD W1 12
GD W2 06
GD W2 17
GD E2 16
GD E1 05
W ID 06
GD E2 12
HA S 05
HA S 29
W CN 02
GD GS 01
W TW 02
GD E4 09
GD E3 04
GD E3 07
Basmati 3
W KH 11
W TH 15
W TH 30

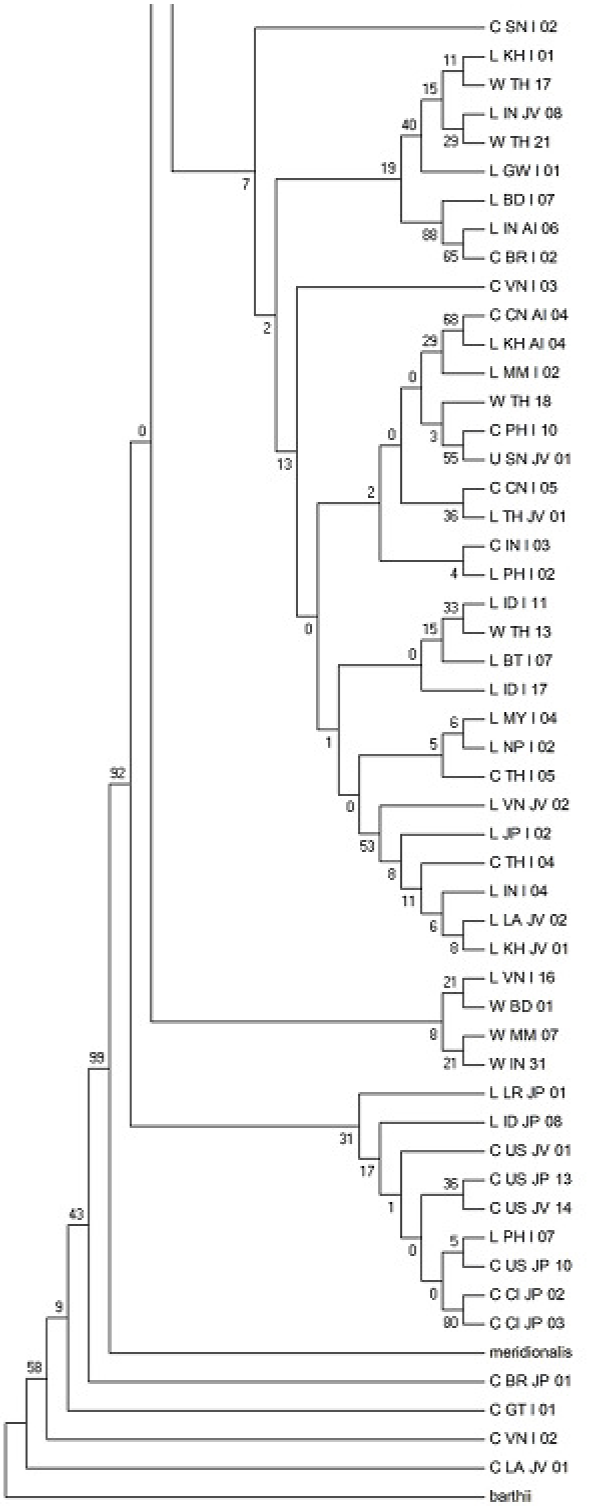
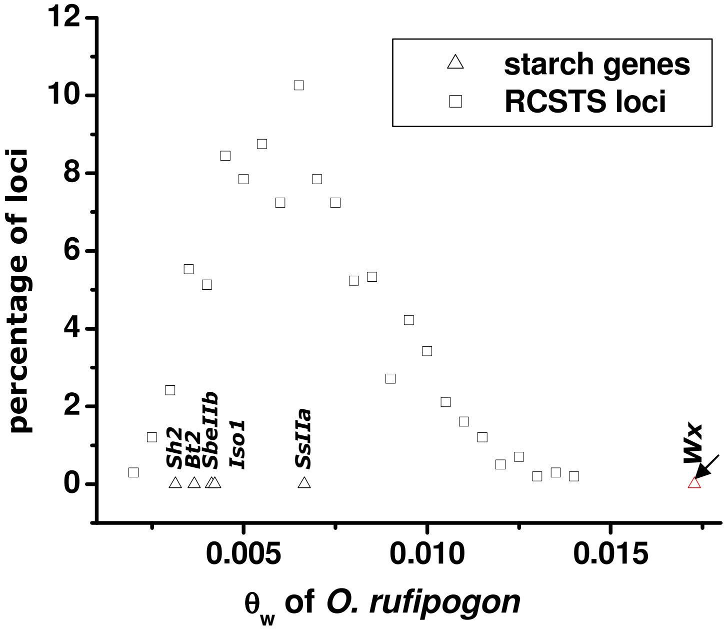
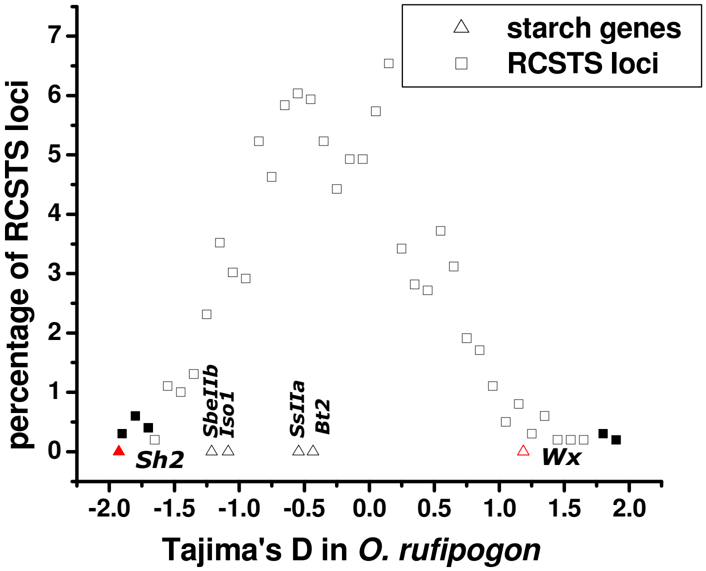
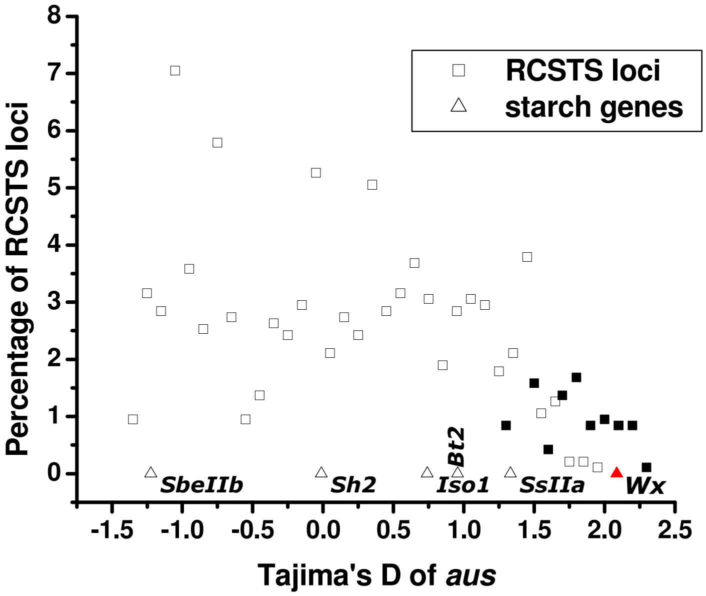
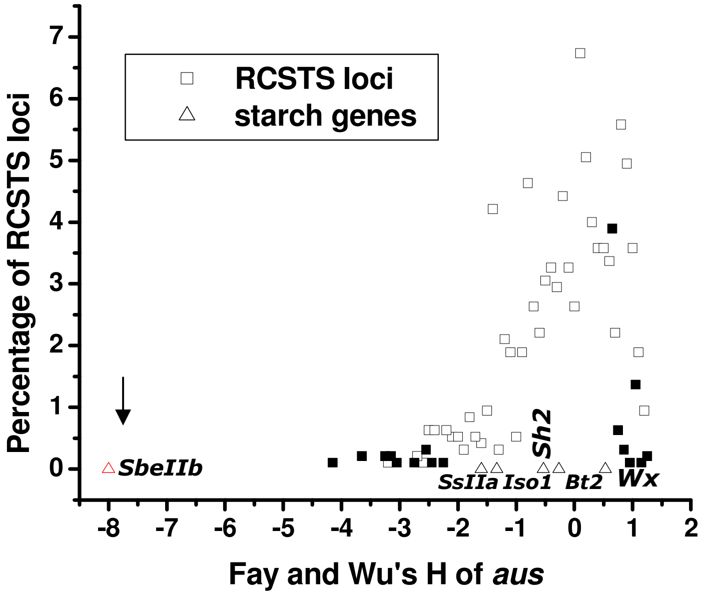

C SN I 02
11
L KH I 01
W TH 17
15
L IN JV 08
40
29
W TH 21
19
L GW I 01
7
L BD I 07
88
L IN AI 06
65
C BR I 02
C VN I 03
68
C CN AI 04
29
L KH AI 04
2
L MM I 02
0
W TH 18
0
C PH I 10
3
0
55
U SN JV 01
13
C CN I 05
2
36
L TH JV 01
C IN I 03
4
L PH I 02
33
L ID I 11
0
W TH 13
15
L BT I 07
0
L ID I 17
6
L MY I 04
1
5
L NP I 02
92
C TH I 05
0
L VN JV 02
53
L JP I 02
8
C TH I 04
11
L IN I 04
6
L LA JV 02
8
L KH JV 01
21
L VN I 16
W BD 01
8
W MM 07
21
W IN 31
99
L LR JP 01
L ID JP 08
31
C US JV 01
17
36
C US JP 13
1
C US JV 14
43
5
L PH I 07
0
C US JP 10
0
C CI JP 02
80
C CI JP 03
9
meridionalis
58
C BR JP 01
C GT I 01
C VN I 02
C LA JV 01
barthii

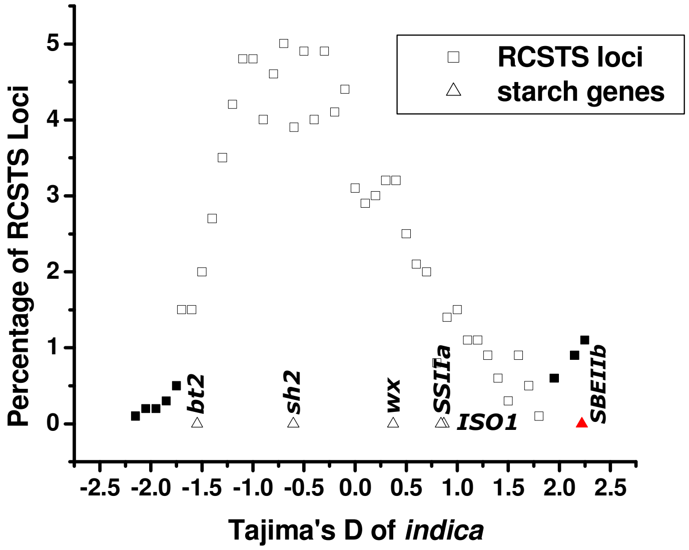
Figure 1.5), The results of comparison between studied starch genes and RCSTS loci. Triangle represents studied starch genes. Square represents RCSTS loci. The filled triangle or square represents deviation from standard neutral model. Red triangle means that the gene significantly deviates from genome-wide RCSTS loci by 5% significantly level. The gene that pointed by an arrow is the gene that is significantly deviated from genome-wide RCSTS loci even after Bonferroni correction.

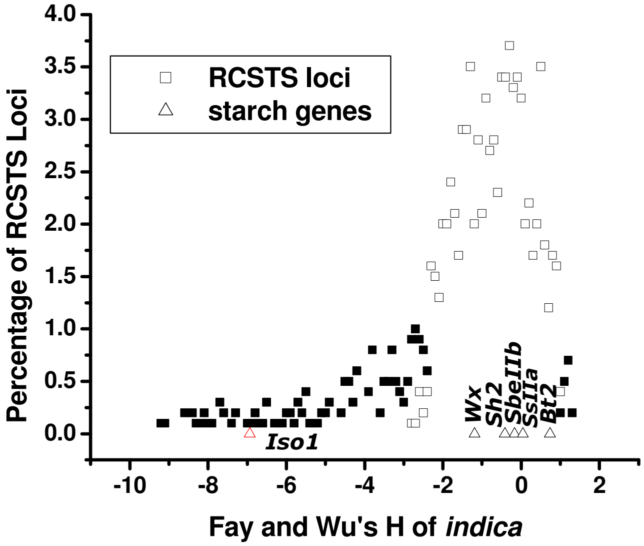

starch genes
RCSTS loci
percentage of loci
0
2
4
6
8
10
12
0.005
0.010
0.015
θw of O. rufipogon
Sh2
Bt2
SbeIIb
Iso1
SsIIa
Wx

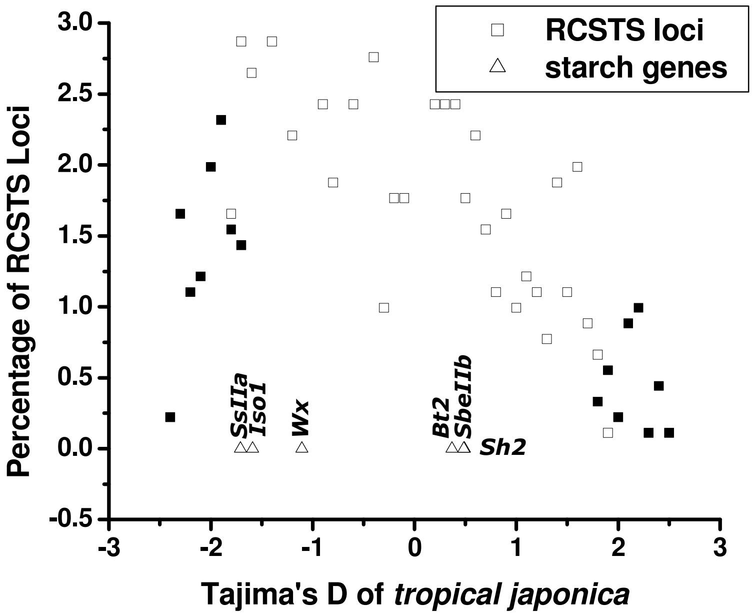

starch genes
RCSTS loci
percentage of RCSTS loci
0
1
2
3
4
5
6
7
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Tajima's D in O. rufipogon
Sh2
SbeIIb
Iso1
SsIIa
Bt2
Wx

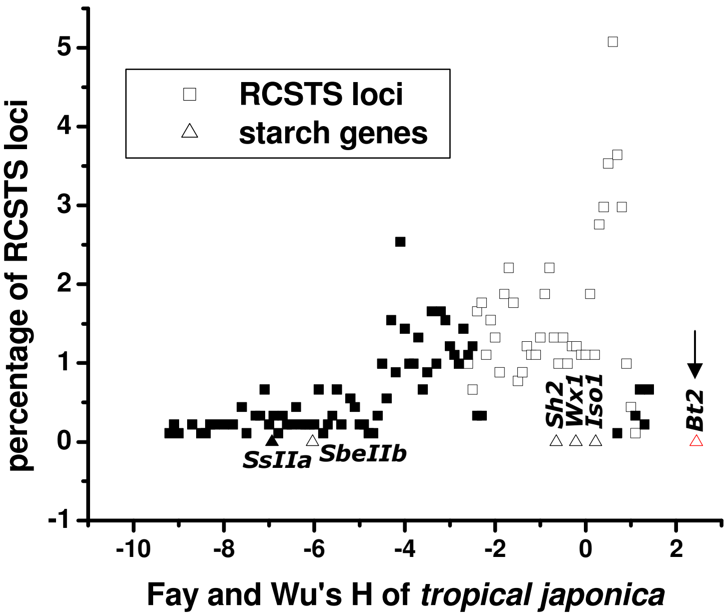

RCSTS loci
starch genes
Percentage of RCSTS loci
Tajima's D of *aus*
SbeIIb
Sh2
Iso1
Bt2
SsIIa
Wx
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
5
6
7
8

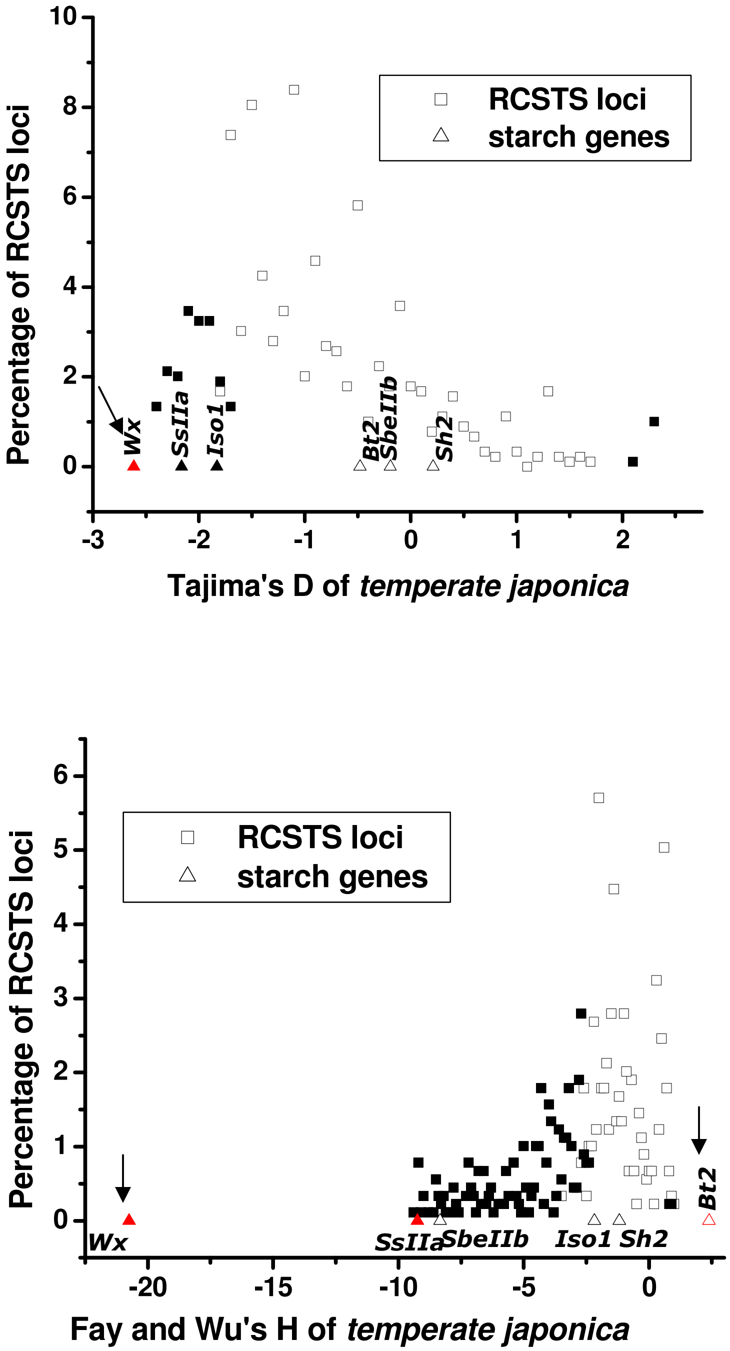

RCSTS loci
starch genes
Percentage of RCSTS loci
Fay and Wu's H of *aus*
SbeIIb
SsIIa
Iso1
Sh2
Bt2
Wx
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
0
1
2
3
4
5
6
7

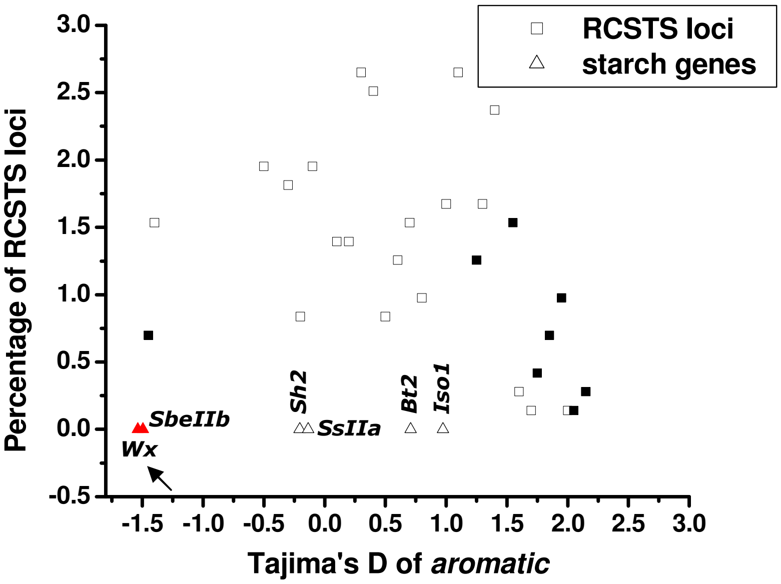

Percentage of RCSTS Loci
0
1
2
3
4
5
RCSTS loci
starch genes
bt2
sh2
wx
SSIIa
ISO1
SBEIIb
-2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5
Tajima's D of *indica*


Percentage of RCSTS Loci
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
RCSTS loci
starch genes
Wx
Sh2
SbeIIb
SsIIa
Bt2
Iso1
-10
-8
-6
-4
-2
0
2
Fay and Wu's H of *indica*


RCSTS loci
starch genes
Percentage of RCSTS Loci
Tajima's D of *tropical japonica*
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-3
-2
-1
0
1
2
3
SsIIa
Iso1
Wx
Bt2
SbeIIb
Sh2


RCSTS loci
starch genes
percentage of RCSTS loci
Fay and Wu's H of *tropical japonica*
5
4
3
2
1
0
-1
-10
-8
-6
-4
-2
0
2
SsIIa
SbeIIb
Sh2
Wx1
Iso1
Bt2


Percentage of RCSTS loci
10
8
6
4
2
0
-3
-2
-1
0
1
2
Tajima's D of *temperate japonica*
□ RCSTS loci
△ starch genes
Wx
SsIIa
Iso1
Bt2
SbeIIb
Sh2
Percentage of RCSTS loci
6
5
4
3
2
1
0
-20
-15
-10
-5
0
Fay and Wu's H of *temperate japonica*
□ RCSTS loci
△ starch genes
Wx
SsIIa
SbeIIb
Iso1
Sh2
Bt2


RCSTS loci
starch genes
Percentage of RCSTS loci
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
SbeIIb
Wx
Sh2
SsIIa
Bt2
Iso1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Tajima's D of *aromatic*

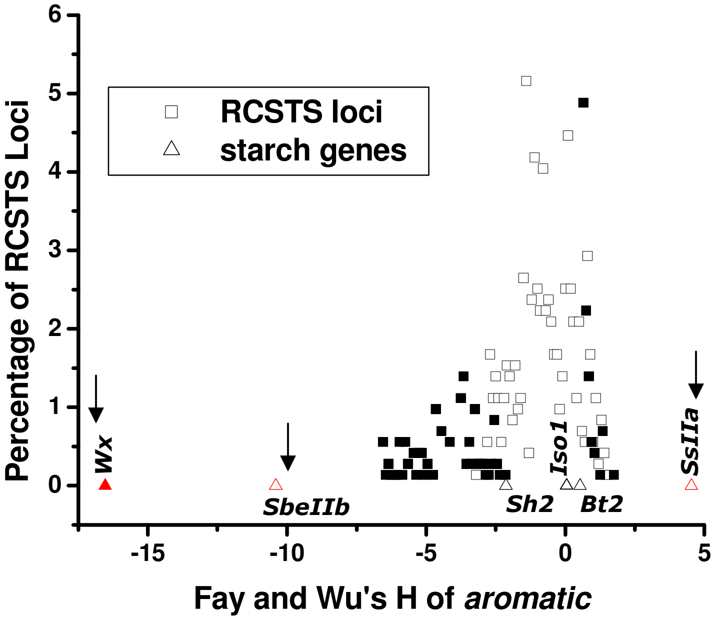

RCSTS loci
starch genes
Percentage of RCSTS Loci
6
5
4
3
2
1
0
Wx
SbeIIb
Iso1
Sh2
Bt2
SsIIa
-15
-10
-5
0
5
Fay and Wu's H of *aromatic*

Figure 1.6), Level of diversity (estimated by θ_w) in *O. rufipogon* and rice variety groups. The landrace is used for *tropical japonica*, *temperate japonica* and *indica* in order to know the change of variation level during initial domestication by comparing them with *O. rufipogon*. The red arrow shows the extreme reduction of diversity in *tropical japonica.*

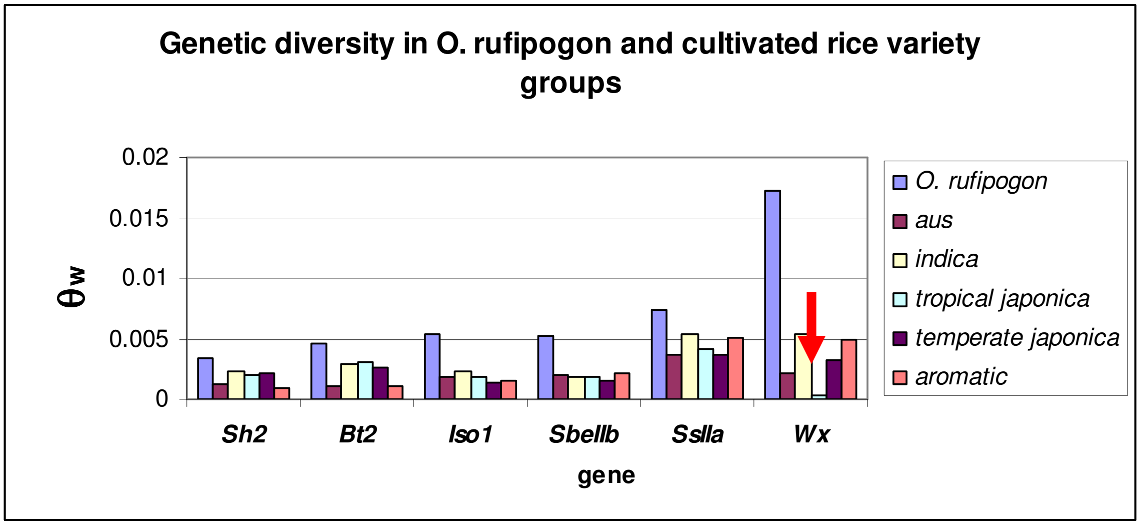
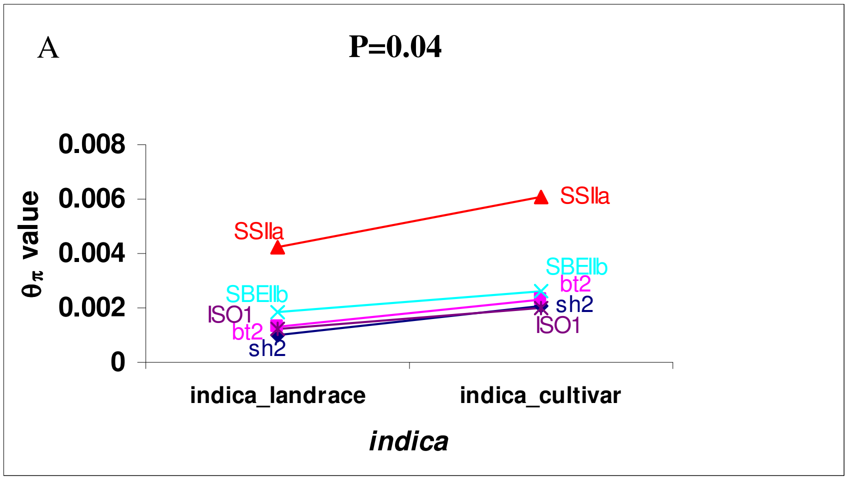
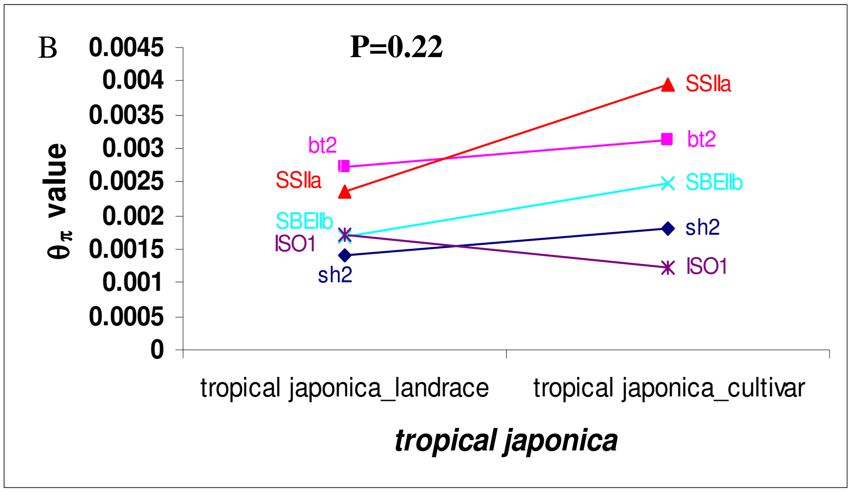
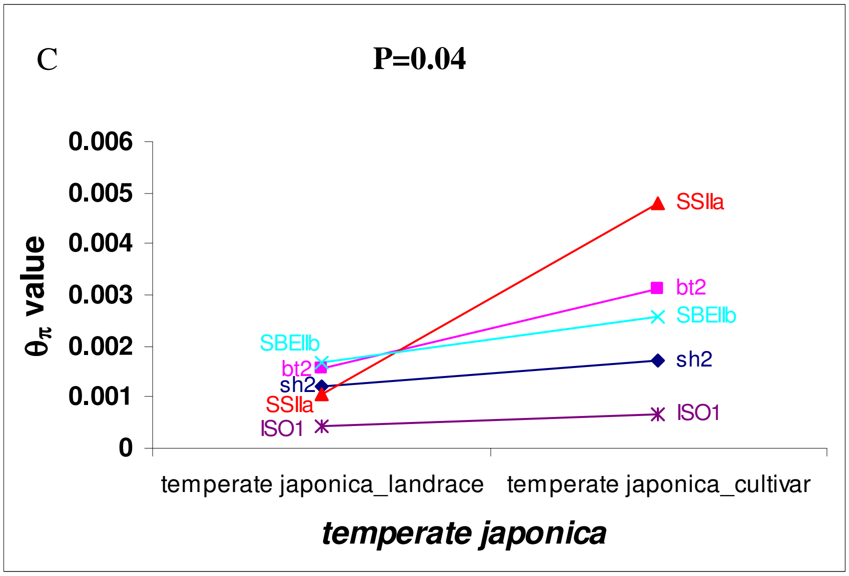

Genetic diversity in O. rufipogon and cultivated rice variety groups
θw
0.02
0.015
0.01
0.005
0
Sh2
Bt2
Iso1
SbeIIb
SsIIa
Wx
gene
O. rufipogon
aus
indica
tropical japonica
temperate japonica
aromatic

Figure 1.7), Comparison of levels of diversity between landraces and modern cultivars. Comparison by θ_π is shown in fig A, B, C; Comparison by θ_w is shown in fig D, E, F;


A
P=0.04
θπ value
0.008
0.006
0.004
0.002
0
SSIIa
SBEIIb
ISO1
bt2
sh2
SSIIa
SBEIIb
bt2
sh2
ISO1
indica_landrace
indica_cultivar
indica


B
P=0.22
θπ value
0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
bt2
SSIIa
SBEIIb
ISO1
sh2
SSIIa
bt2
SBEIIb
sh2
ISO1
tropical japonica_landrace
tropical japonica_cultivar
tropical japonica


C
P=0.04
θπ value
0.006
0.005
0.004
0.003
0.002
0.001
0
SBEIIb
bt2
sh2
SSIIa
ISO1
SSIIa
bt2
SBEIIb
sh2
ISO1
temperate japonica_landrace
temperate japonica_cultivar
temperate japonica

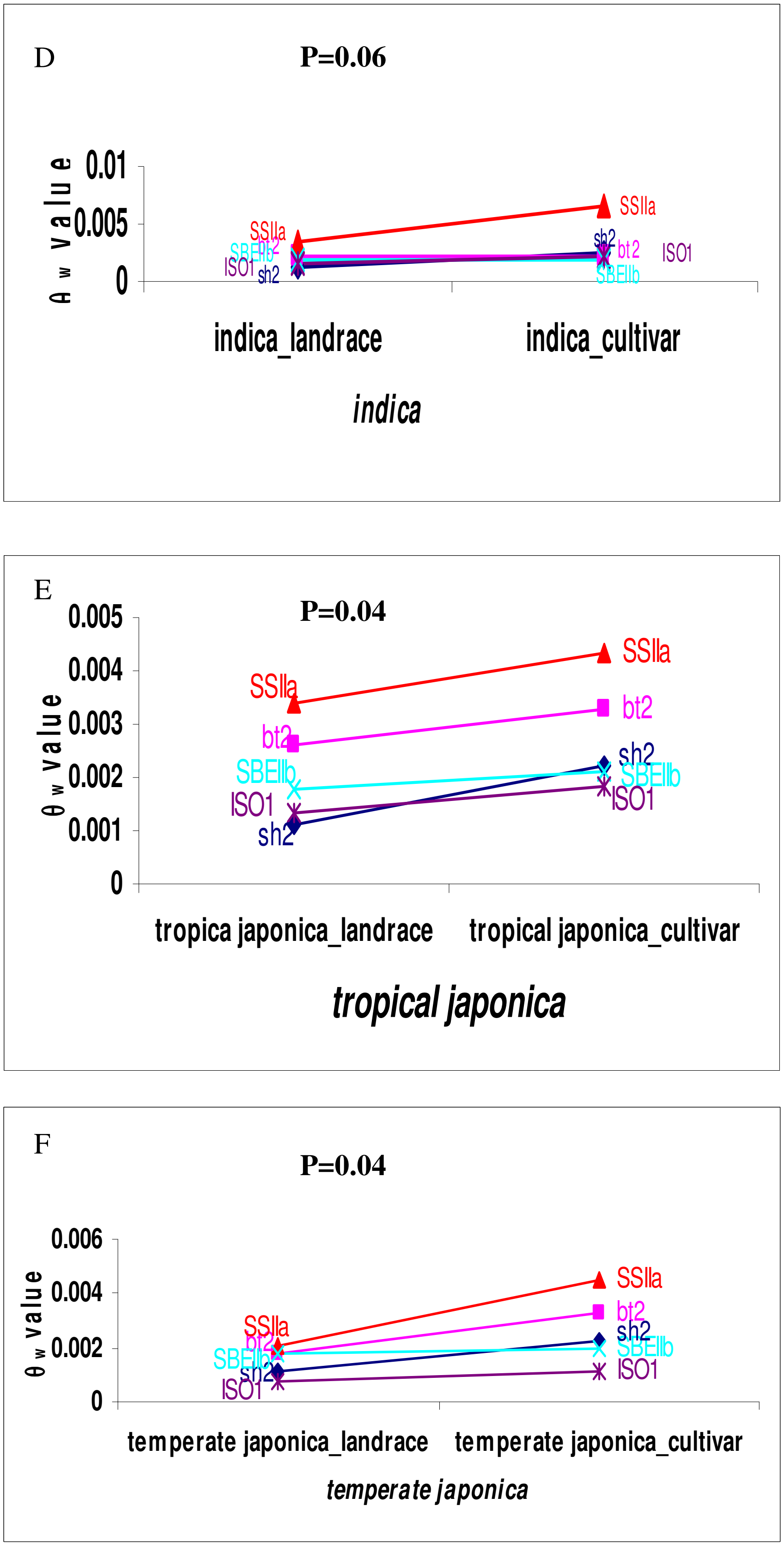
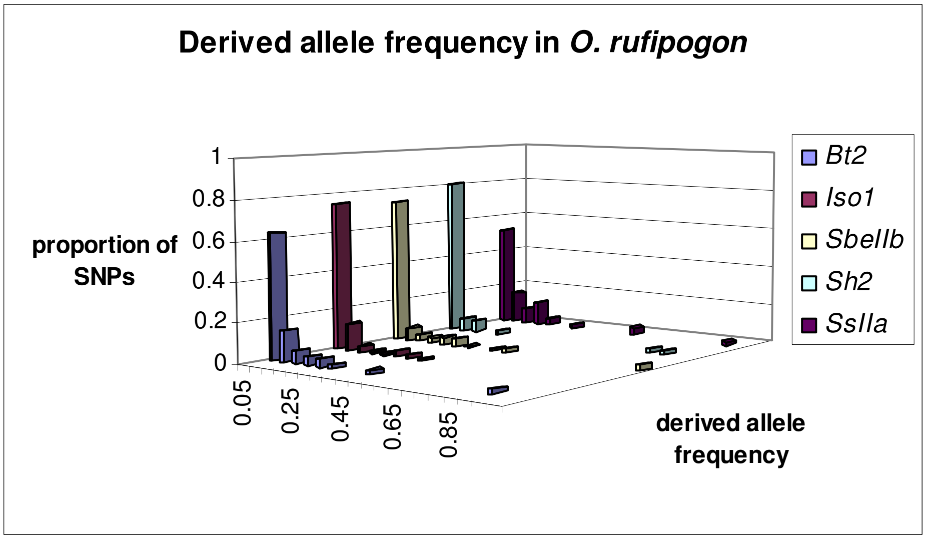
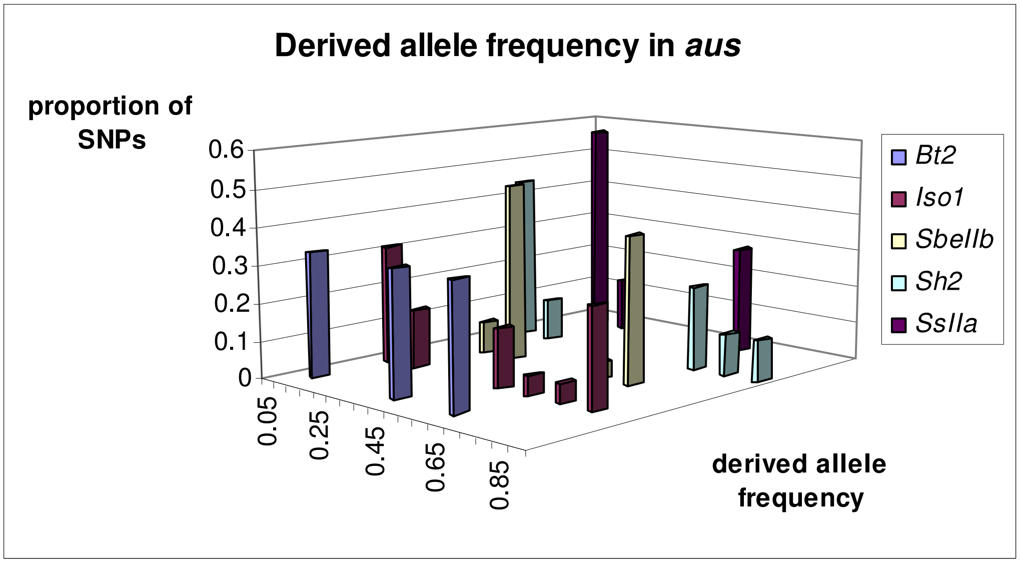
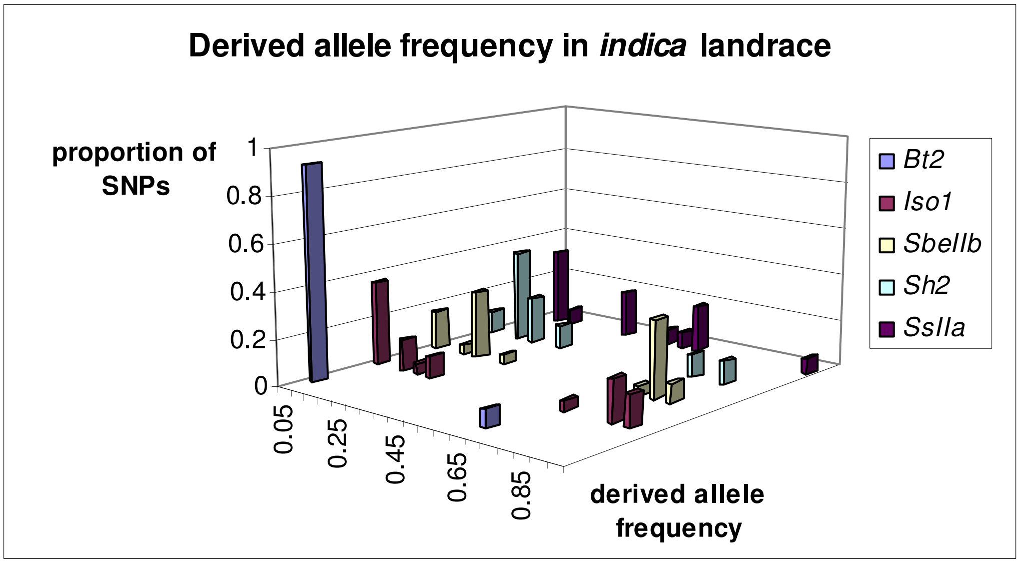

D
P=0.06
θ w value
0.01
0.005
0
SSIIa
bt2
SBEIIb
ISO1
sh2
indica_landrace
indica_cultivar
indica
E
P=0.04
θ w value
0.005
0.004
0.003
0.002
0.001
0
SSIIa
bt2
SBEIIb
ISO1
sh2
tropica japonica_landrace
tropical japonica_cultivar
tropical japonica
F
P=0.04
θ w value
0.006
0.004
0.002
0
SSIIa
bt2
SBEIIb
sh2
ISO1
temperate japonica_landrace
temperate japonica_cultivar
temperate japonica

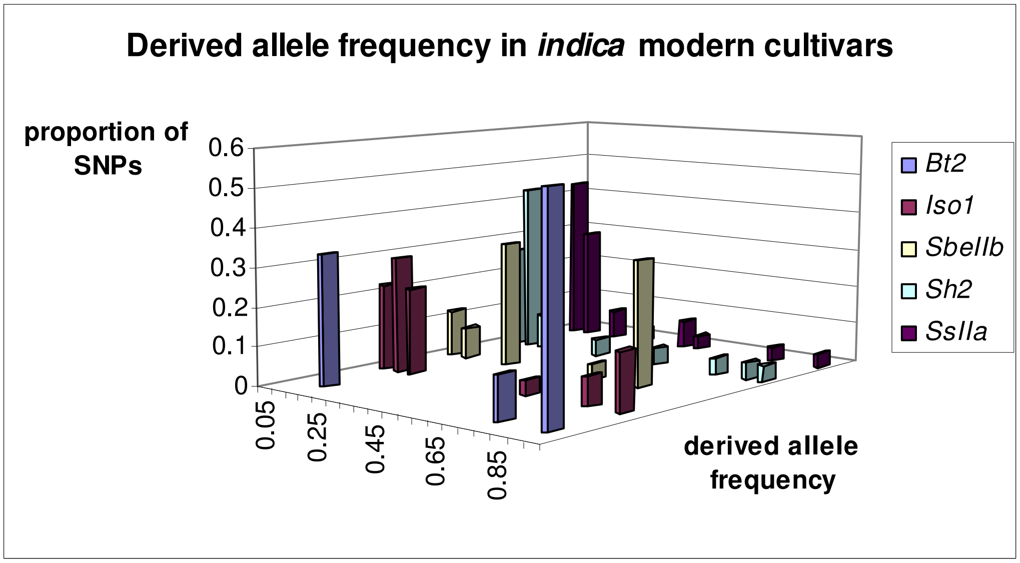
Figure 1.8), Derived allele frequency distributions


Derived allele frequency in *O. rufipogon*
proportion of SNPs
1
0.8
0.6
0.4
0.2
0
0.05
0.25
0.45
0.65
0.85
derived allele frequency
Bt2
Iso1
SbeIIb
Sh2
SsIIa

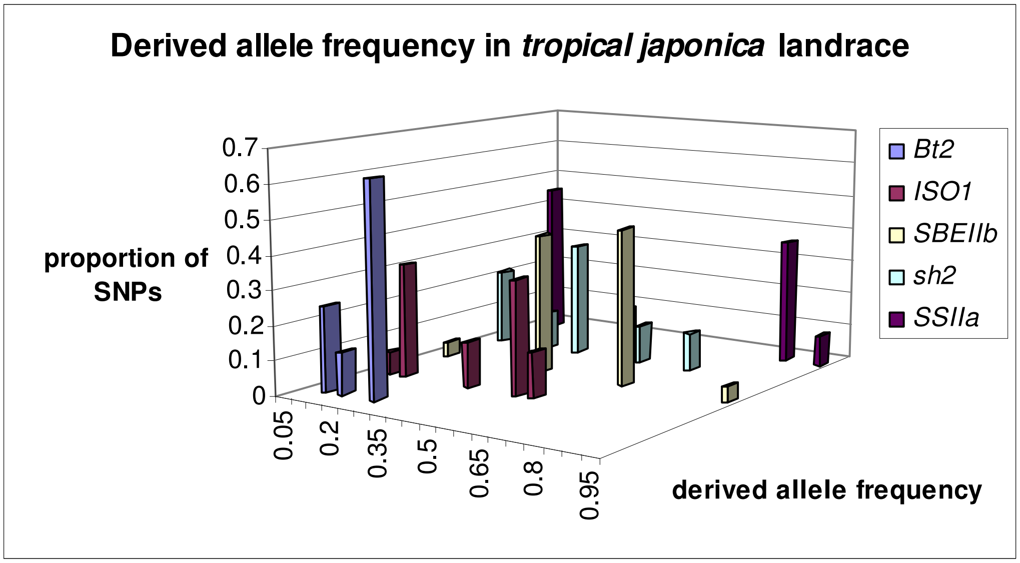

Derived allele frequency in *aus*
proportion of SNPs
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05
0.25
0.45
0.65
0.85
derived allele frequency
Bt2
Iso1
SbeIIb
Sh2
SsIIa



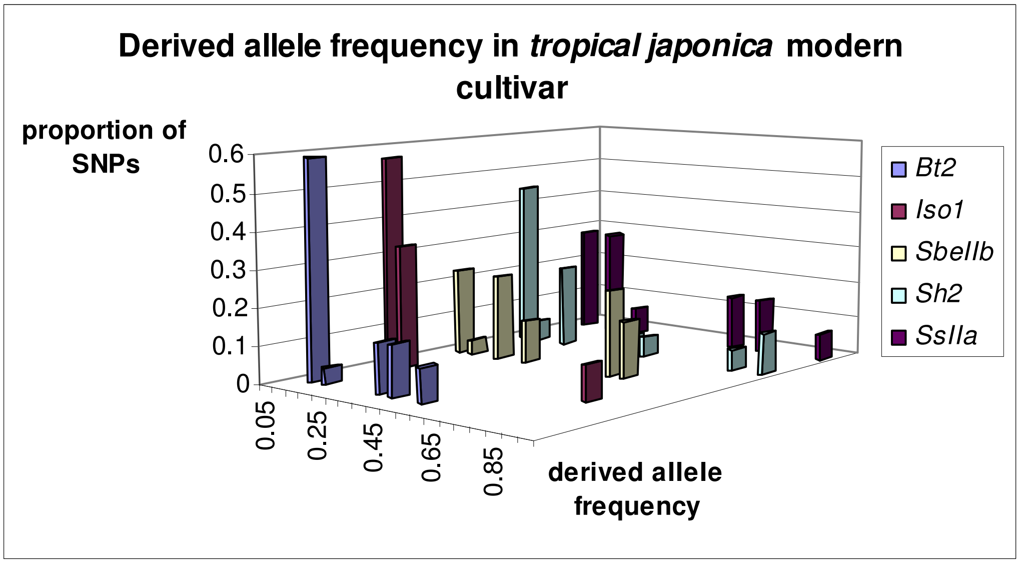
Derived allele frequency in *indica* landrace



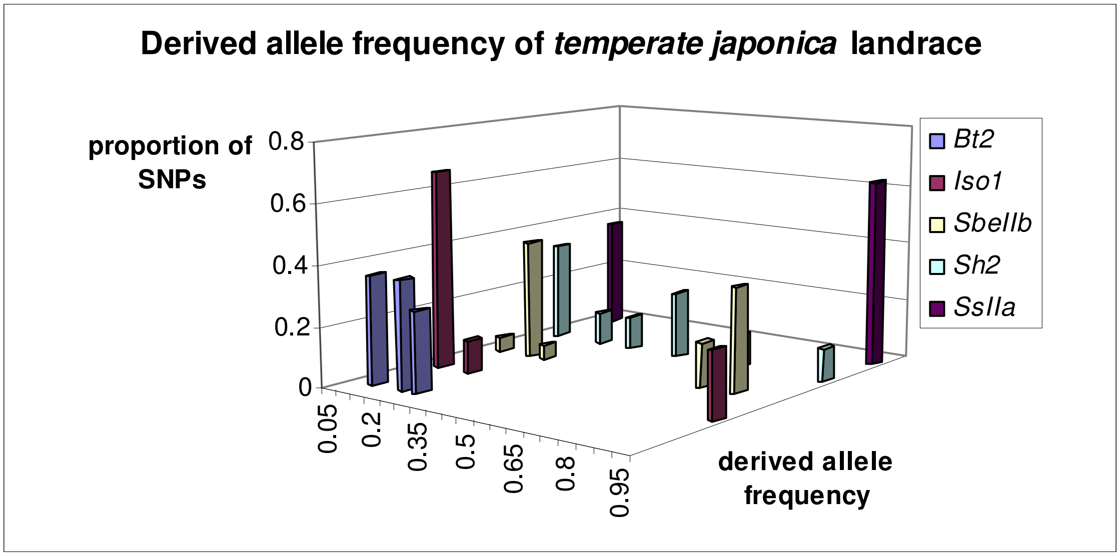
Derived allele frequency in *indica* modern cultivars

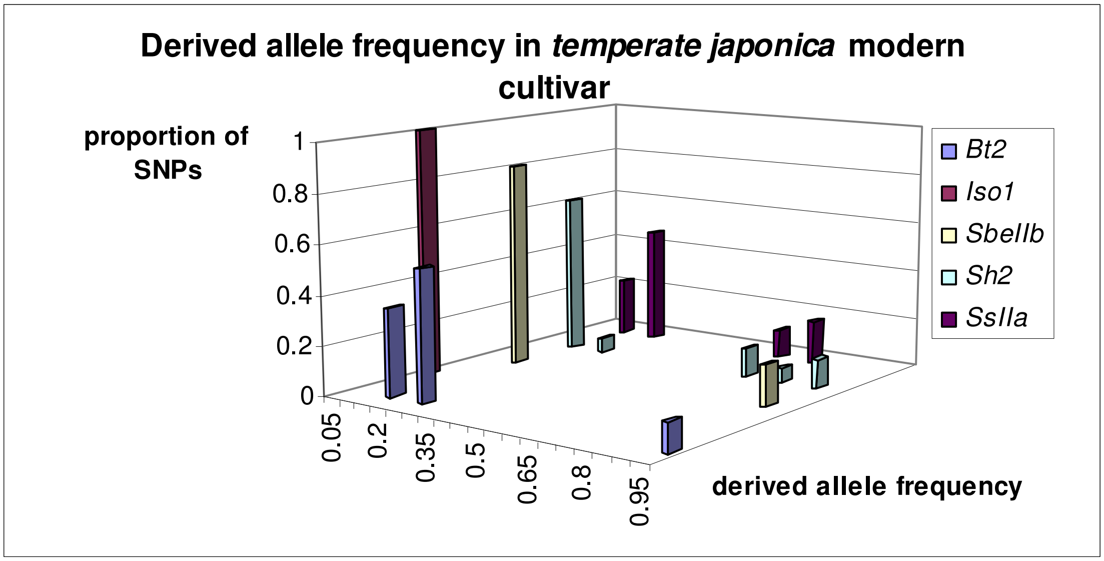


Derived allele frequency in *tropical japonica* landrace



Derived allele frequency in *tropical japonica* modern cultivar


Derived allele frequency of *temperate japonica* landrace
proportion of SNPs
0.8
0.6
0.4
0.2
0
0.05
0.2
0.35
0.5
0.65
0.8
0.95
derived allele frequency
Bt2
Iso1
SbeIIb
Sh2
SsIIa


Derived allele frequency in *temperate japonica* modern cultivar
proportion of SNPs
1
0.8
0.6
0.4
0.2
0
0.05
0.2
0.35
0.5
0.65
0.8
0.95
derived allele frequency
Bt2
Iso1
SbeIIb
Sh2
SsIIa

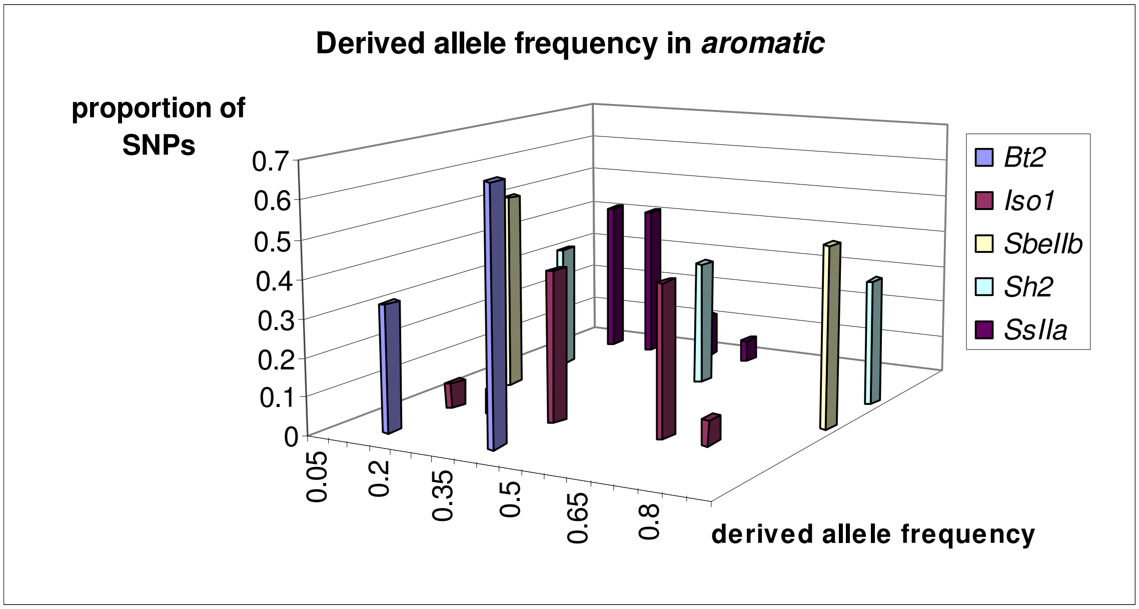
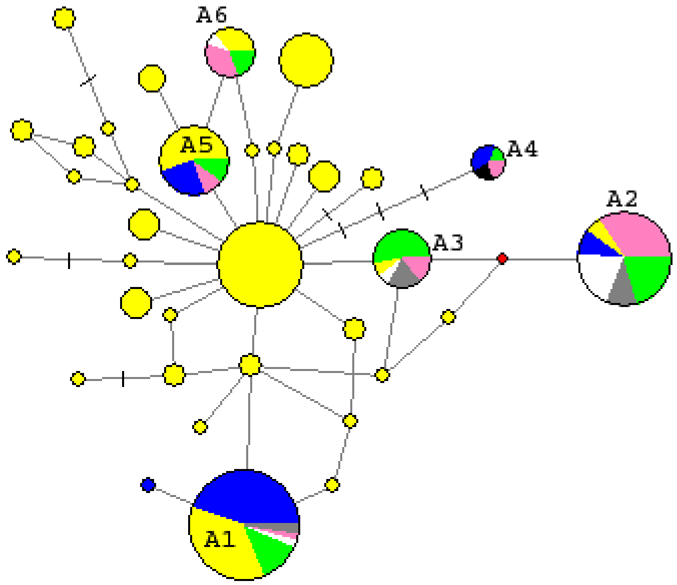
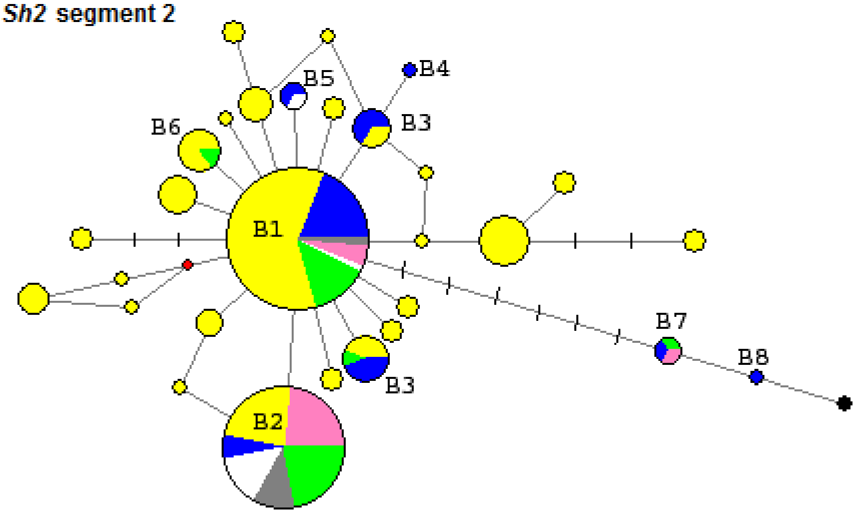

Derived allele frequency in *aromatic*
proportion of SNPs
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05
0.2
0.35
0.5
0.65
0.8
derived allele frequency
Bt2
Iso1
SbeIIb
Sh2
SsIIa

Figure 1.9), Haplotype networks of *O. rufipogon* and *O. sativa* Variety groups and taxons are indicated as follows: *aus* (□), *indica* (🟦), *tropical japonica* (🟩), *temperate japonica* (🟪), *aromatic* (▨), *O. barthii* (■), *O. rufipogon* (🟨). The circles represent the nodes. The size the circle is proportional to the number of individuals. The line connecting the nodes is roughly proportional to the number of mutations.

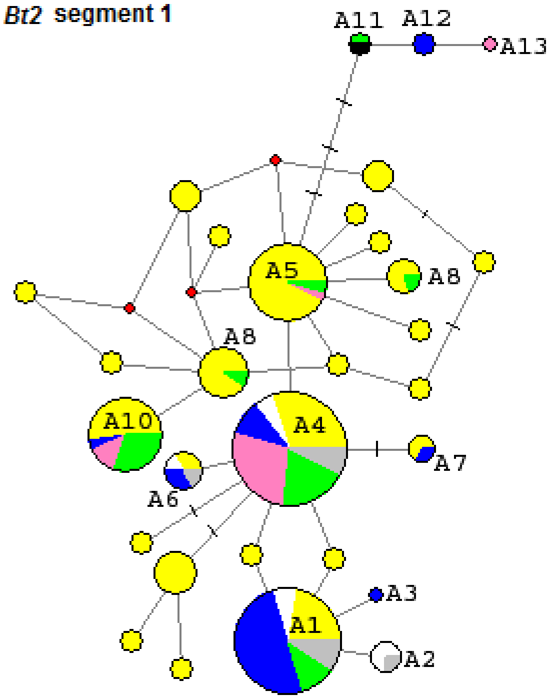
Sh2 segment 1



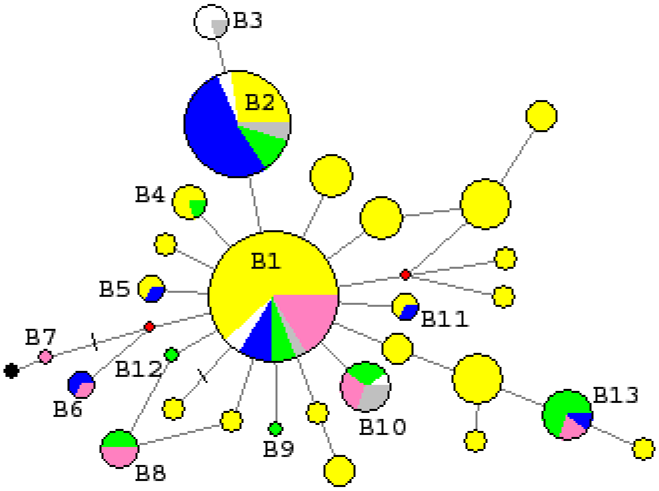
Sh2 segment 2



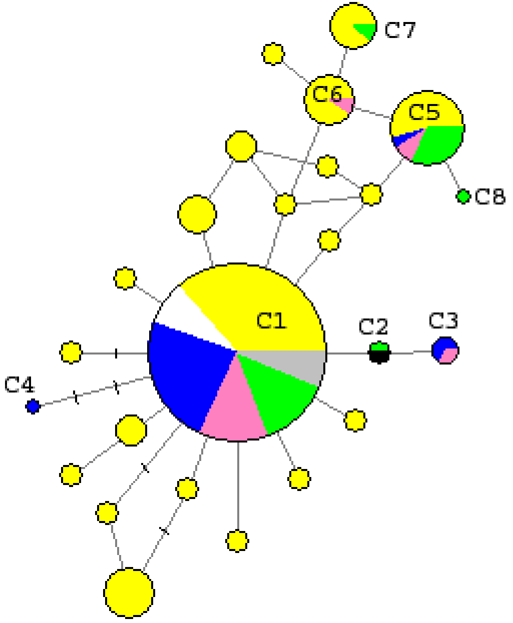
Bt2 segment 1



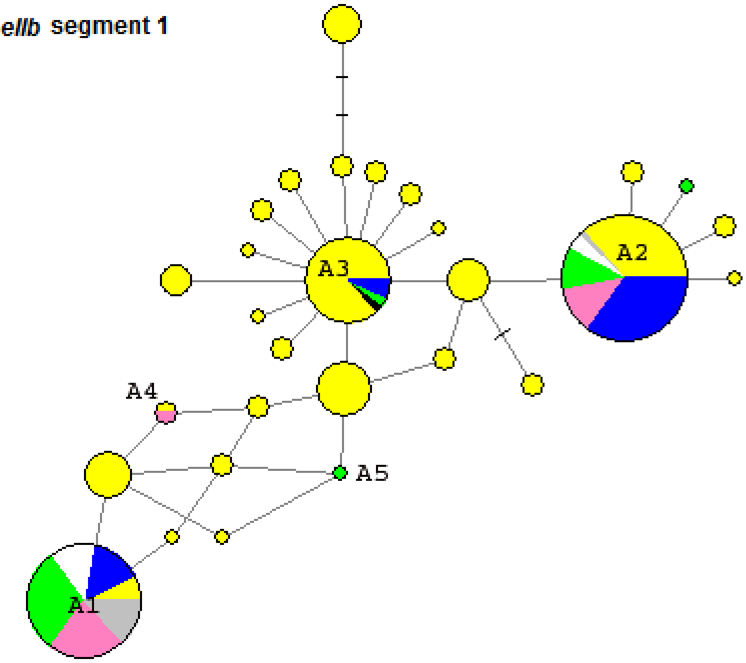
Bt2 segment 2

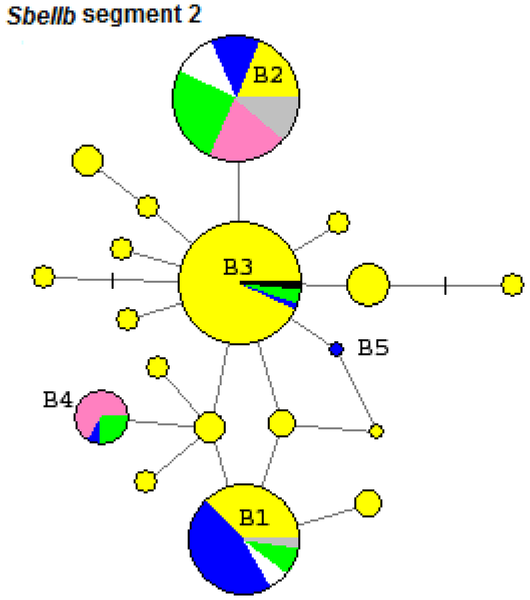


Bt2 segment 3

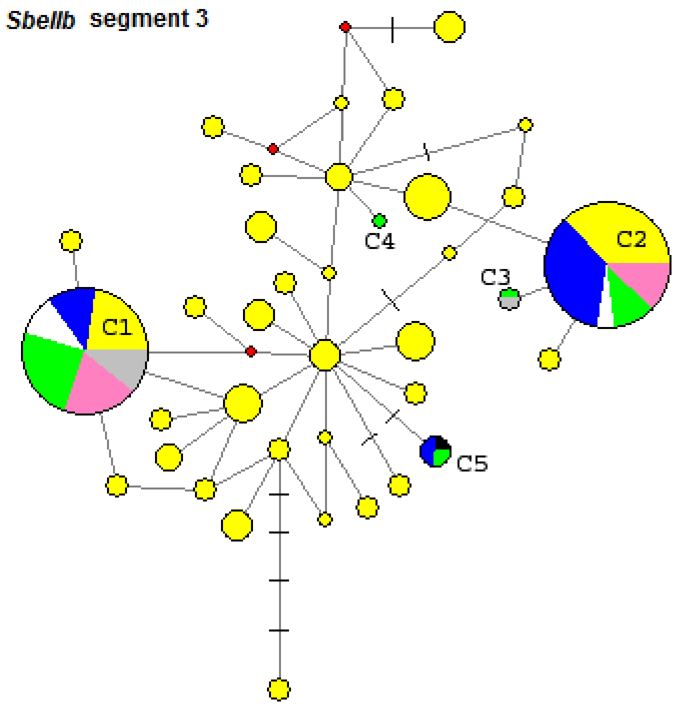


Sbellb segment 1



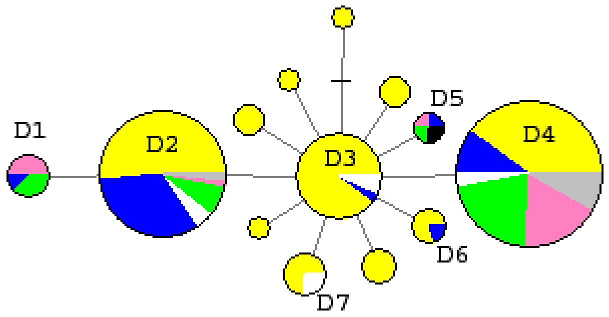


Sbellb segment 2

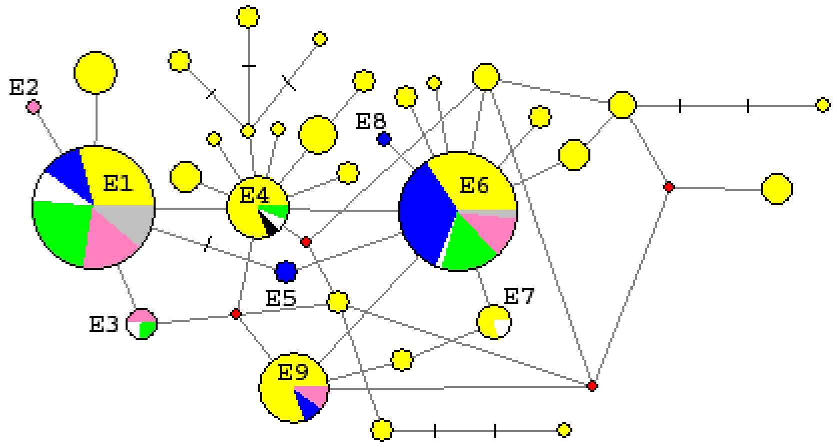


Sbellb segment 3

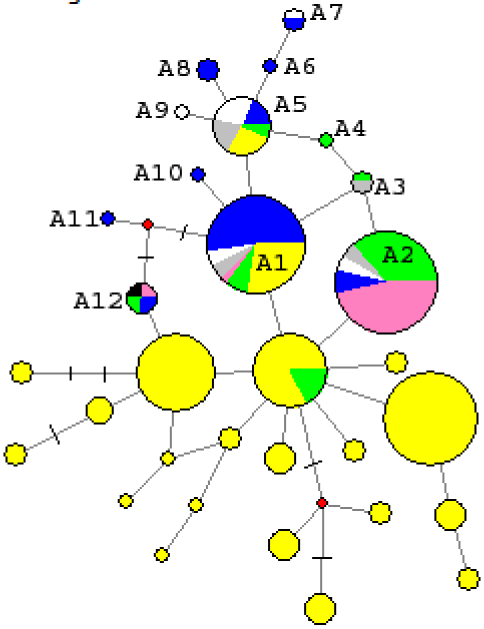
***SbeIIb* segment 4**



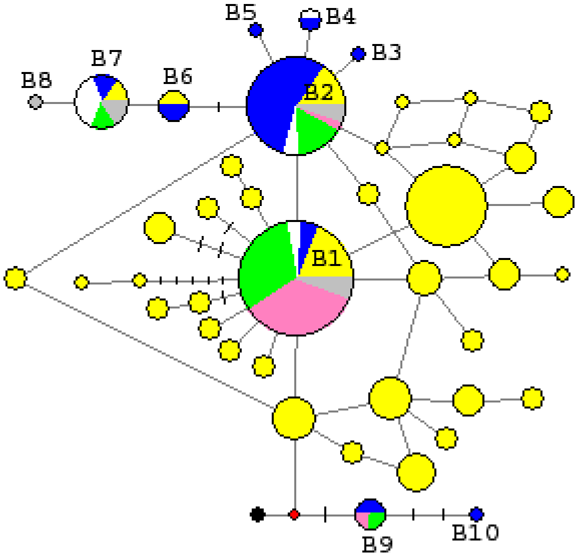
***SbeIIb* segment 5**

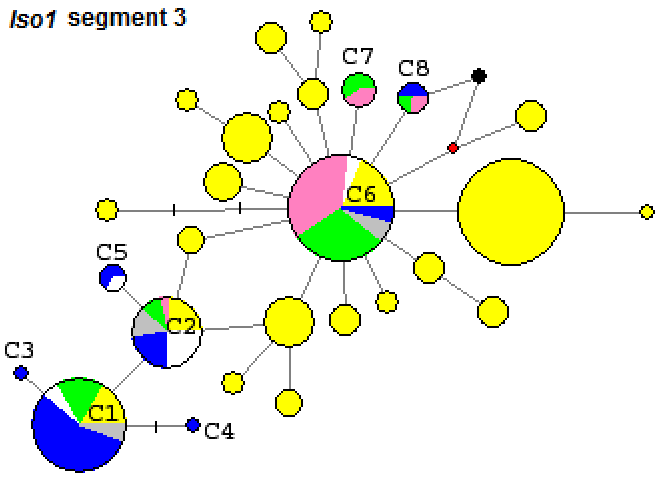


Iso1 segment 1

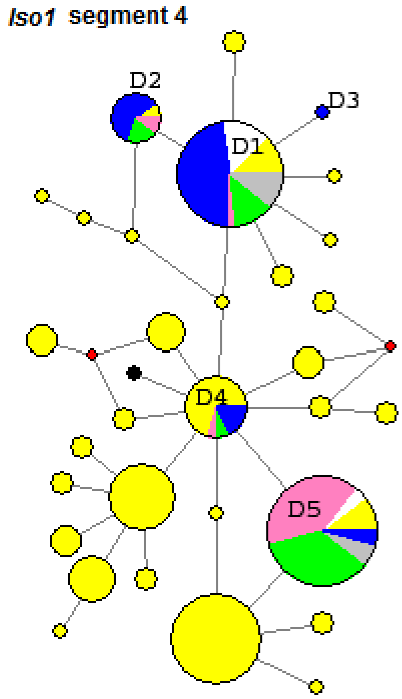


Iso1 segment 2



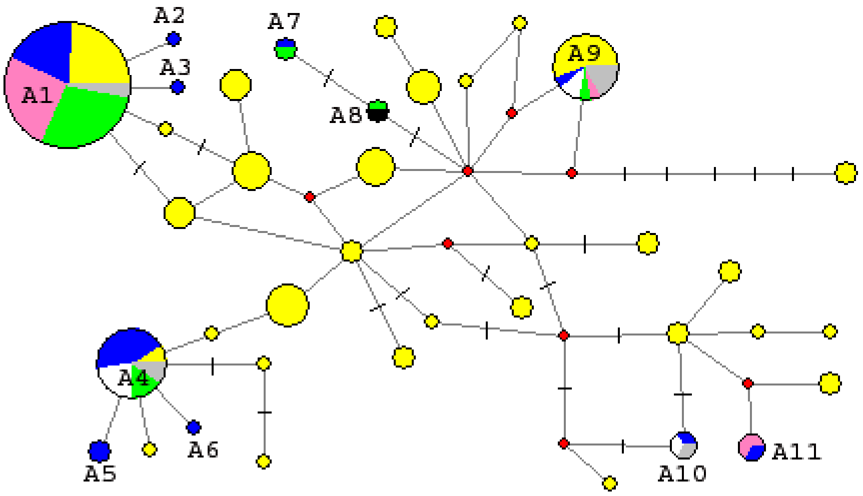


Iso1 segment 3

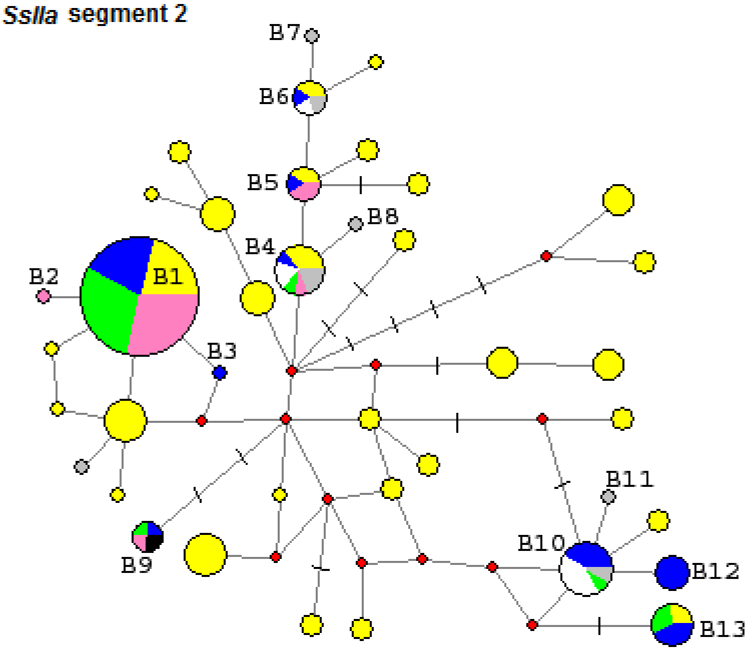


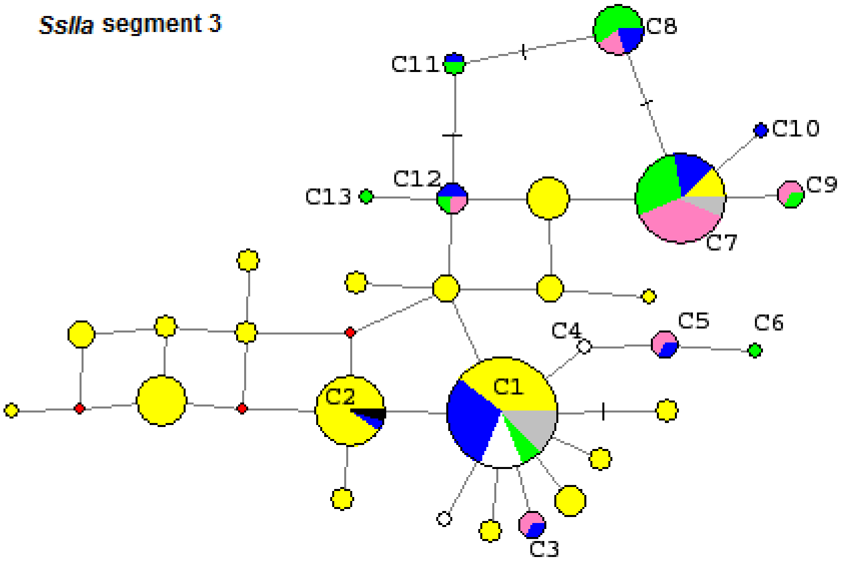
Iso1 segment 4

SsIIa segment 1



SsIIa segment 2

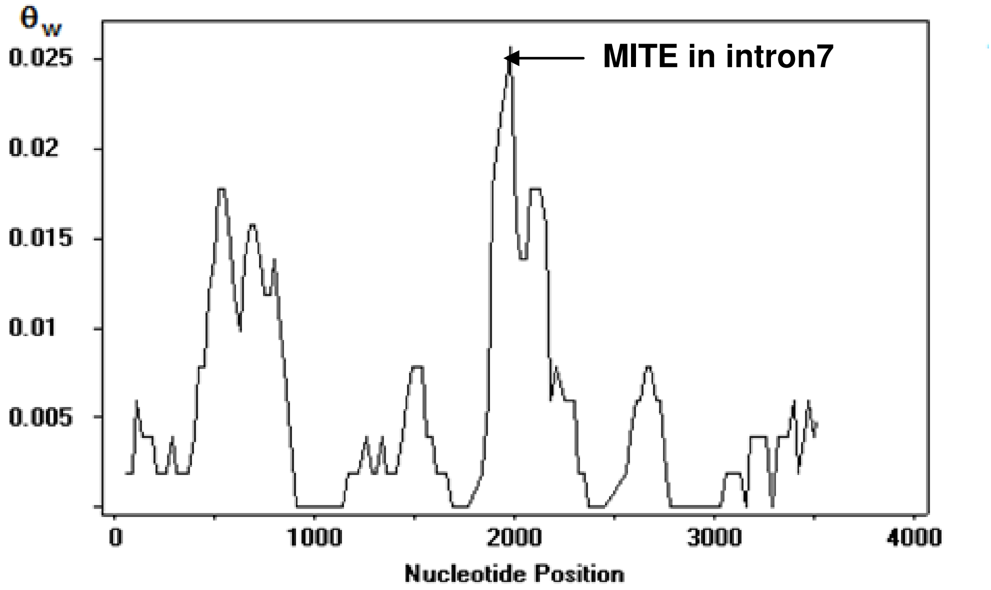




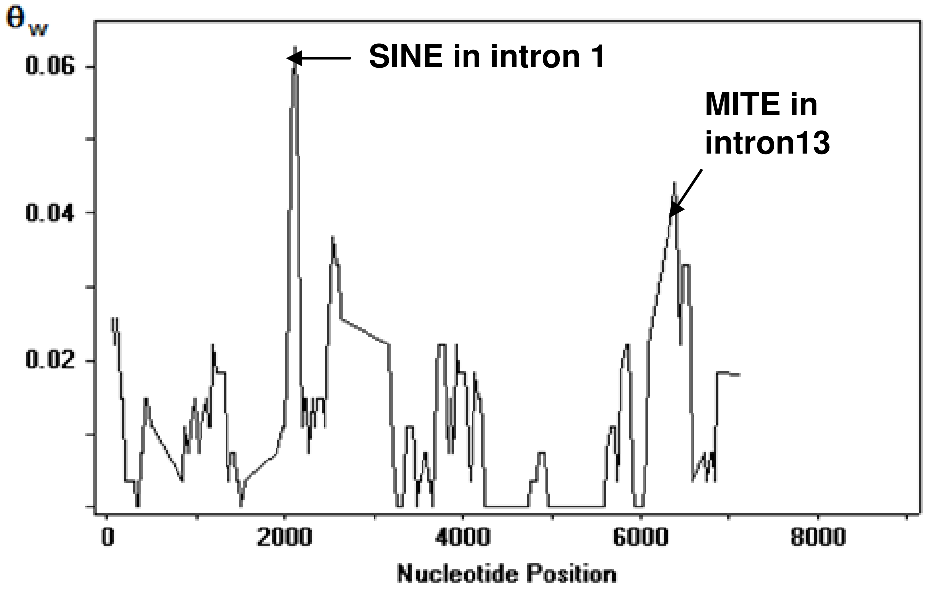
SsIIa segment 3

Figure 1.10), Nucleotide polymorphism (θ_w) across *SsIIa* and *Wx* in *O. rufipogon.* The locations of transposable elements are indicated as arrows.

Nucleotide polymorphism across *SsIIa* in *O. rufipogon*



Nucleotide polymorphism across *Wx1* in *O. rufipogon*



Chapter 2

Association between Nonsysnonymous Mutations of Starch Synthase IIa and Starch Quality in *Oryza sativa* and its Wild Ancestor, *Oryza rufipogon*

Introduction

One of the fundamental goals in the study of evolution is to uncover the genetic basis of phenotypic diversity within species and to understand the origin of phenotypic divergence between species (O'CONNOR and MUNDY 2009). Several approaches have been applied for this. One approach is genome-wide association (GWA) study (AMUNDADOTTIR *et al.* 2009), which examines the genome-wide variation across a genome to search for gene regions or single nucleotide polymorphisms (SNPs) associated with observable traits. Another approach is quantitative trait locus study (QTL). QTL methods typically makes crosses between two or more lines that differ genetically with regard to a trait of interest (LYNCH and WALSH 1998). The crosses are then genotyped using SNPs (single nucleotide polymorphism) or other markers across the whole genome, and statistical associations of the linkage disequilibrium between genotype and phenotype are identified. QTL analysis usually identifies one or several genomic regions with dozens of genes and requires further investigation to find the specific genes associated with a particular phenotype. Another approach, candidate gene association study, examines genetic variation across candidate genes and seeks to identify the genes and/or SNPs associated with a particular phenotypes (NACHMAN *et al.* 2003). This approach does not require genome-wide variation across a genome. It only requires information about the candidate genes for a particular phenotype. Moreover, with increasing knowledge of the underlying physiology and biochemistry of specific traits, an increasing number of candidate genes will become available.

In domesticated crops, more candidate genes for important agronomic traits, especially “domesticated” traits favored by early farmers (e.g. reduction of seed shattering and dormancy, increased yield) are being identified (BENTSINK *et al.* 2006; HARLAN 1992; KONISHI *et al.* 2006; LI and SANG 2006; LIN *et al.* 1998). Several emerging approaches have been applied to study these candidate genes in crop species and their wild ancestors. These approaches include examining the variation pattern of these candidate genes for signals of selection, searching for an association between the candidate gene variation and phenotypic variation. For example, recent studies of six starch synthesis pathway genes in *Zea mays* and its wild relative *Z. mays* ssp. *parviglumis* identified three targeted genes of artificial selection during maize domestication: *brittle2* (*Bt2*) *, sugary1* (*Su1*), and *amylose extender1* (*ae1*) (WHITT *et al.* 2002). And it was also found that genes *Bt2*, *shrunken1* (*Sh1)*, and *shrunken2* (*Sh2)* showed significant associations for kernel composition traits, *ae1* and *sh2* showed significant associations for starch pasting properties, *ae1* and *sh1* associated with amylose levels (WILSON *et al.* 2004). In addition, phylogeographic analysis of candidate genes has been used to identify the origin of an allele associated with a specific phenotype. Recently studies of *Wx* (*Waxy*) and betaine aldehyde dehydrogenase gene (*BADH2)* successfully identified a single origin for glutinous rice allele and the fragrance allele (*badh2.1)* respectively (KOVACH *et al.* 2009; OLSEN *et al.* 2006).

Here, I present an association analysis of nonsynonymous variation at the starch synthase IIa (*SsIIa*) gene exon 8 region with starch disintegration level in alkali in Asian rice, *Oryza sativa* and its ancestor species, *Oryza rufipogon*. Evolution of starch quality (determined by starch disintegration level in alkali) during rice domestication is also analyzed by comparing it among rice variety groups and their wild ancestor, *O. rufipogon*. The evolutionary relationship of the nonsynonymous mutations at *SsIIa* exon 8 is also examined.

Study system

Asian rice, *Oryza sativa*, is one of the most important food sources and feeds about half of the world's population (KHUSH 2005). It is also one of the oldest domesticated species, domesticated in Asia at least 10,000 years ago (DIAMOND 2002). Rice is highly variable in phenotype with an estimated 120,000 varieties (KHUSH 1997a). Most varieties of rice can be placed into two subspecies or races, *Oryza sativa ssp.indica* and *Oryza sativa ssp. japonica*, based on their morphological, physiological and ecological differences (KHUSH 1997a). Recent studies by molecular markers identified five rice variety groups including *aus, indica, tropical japonica, temperate japonica* and *aromatic* rice (GARRIS *et al.* 2005). Among these five variety groups, *indica*, *tropical japonica* and *temperate japonica* are the major groups, which are widely grown in Asia (MATHER *et al.* 2007). Due to their ecological difference, *indica* and *tropical japonica* varieties are mainly grown in the tropical or subtropical regions such as India, Southeast Asia; *temperate japonica* varieties are common in temperate region such as Northeastern Asia. *Aus* varieties are known as early maturing and drought tolerant upland rice, with a restricted distribution in Bangladesh and West Bengal state of India. *Aromatic* rice predominates in the Indian subcontinent (KHUSH 1997a). Rice was domesticated from the wild species, *O. rufipogon* (KHUSH 1997a). Recent studies have indicated that there were at least two domestication centers: one in south China for *japonica* rice, another in south and southwest of the Himalayan mountain range for *indica* rice. *Aus* rice may be a third domestication event (GARRIS *et al.* 2005; LONDO *et al.* 2006). Within *japonica* rice, *temperate japonica* is believed to be derived from *tropical japonica* during the spread of rice cultivation towards the north in Asia (GARRIS *et al.* 2005; KHUSH 1997a).

Starch, the major component of cereal grains, is a major determinant of both yield and quality in cereal crops. Starch is composed of amylose and amylopectin. Amylose is a linear molecule of (1→4) linked α-D-glucopyranosyl units. Amylopectin is the highly branched component of starch. It is formed through chains of α-D-glucopyranosyl residues linked together by 1→4 linkages but with 1→6 bonds at the branch points (BULÉON *et al.* 1998). Amylopectin molecules vary in fine structure by the length of branches and are classified into two types: L-type and S-type. The L-type amylopectin differs from the S-type amylopectin in that the former has a dramatically lower proportion of short amylopectin chains with a degree of polymerization (DP) <=10 (NAKAMURA *et al.* 2006).

Two types of rice endosperm starch*,* *japonica* type or *indica* type has been identified based on starch disintegration levels in alkali (OKA and MORISHIMA 1997). The *indica* type starch tends to have discrete, nocohesive grains when cooked and has a high amylose and L-type amylopectin level (MORISHIMA *et al.* 1992). The *japonica* type forms cohesive grains when cooked, has a low amylose and S-type amylopectin level (JULIANO and VILLAREAL 1993). Variety difference in starch disintegration in alkali (1.5% KOH) solution were first reported by Warth and Darabsett (WARTH and DARABSETT 1914) and standardized later by Little et al., into a numerical scale (numerical scales 1–7), which is called starch alkali spreading score (SASS) in this study (LITTLE *et al.* 1958).

Previous genetic studies revealed a gene which controls the starch alkali disintegration difference between *indica* type and *japonica* type starch (KUDO 1968). This gene was designated *alkali* (*alk*) and mapped on chomosome 6 (KUDO 1968). Recently the *alk* was identified as the gene starch synthase IIa (*SsIIa)*, which encodes an enzyme of the starch synthase and is involved in the synthesis of amylopectin in rice endosperm (GAO *et al.* 2003; UMEMOTO and AOKI 2005; UMEMOTO *et al.* 2004; UMEMOTO *et al.* 2002). *SsIIa* plays the distinct role of elongating short chains (DP<=10) of amylopectin cluster. Extremely low *SsIIa* enzyme activity (as in varieties with *japonica* type starch) will result in S-type amylopectin, which have enriched short chains (DP, 6-10) and few long chains (DP, 12-22), wheareas high *SsIIa* enzyme activity will cause L-type amylopectin (as in varieties with *indica*-type starch) (UMEMOTO *et al.* 2004).

Four nonsysnonymous SNPs have been observed at *SsIIa* within cultivated rice (UMEMOTO *et al.* 2004). One is located at the exon 1 region (Fig 2.1, designated SNP0), and the other three are located in the exon 8 region (Fig 2.1, designated SNP1, 2 and 3). In order to determine the effect of these four nonsynonymous SNPs on enzyme activity of *SsIIa*, previous studies expressed the genes with all possible combination of these four nonsysnonymous SNPs in *Escherichia coli*. SNP0 has no effect on *SsIIa* enzyme activity in *E. coli* (see Fig 2.1) (NAKAMURA *et al.* 2005; UMEMOTO and AOKI 2005). However, SNP 2 or 3 have a marked effect on *SsIIa* enzyme activity by replacing either of two amino acids (See Fig 2.1) at *SsIIa* exon 8. SNP 1 will only have slight effect on *SsIIa* enzyme activity only when the nucleotides at SNP 2 and 3 are G and T. This study showed the relationship between variation of *SsIIa* enzyme activity in *E. loci* and the nonsynonymous SNPs at *SsIIa* exon 8 (Fig 2.1, SNP1, 2, and 3) (NAKAMURA *et al.* 2005).

This relationship found in *E. coli* between nonsynonymous SNPs at *SsIIa* exon 8 region and *SsIIa* enzyme activity can be examined in rice. Due to the significant association between starch alkali spreading score (SASS) and amount of *SsIIa* protein associated with starch granule (this reflects *SsIIa* enzyme activity in rice) in rice (Fig 2.2), SASS can be used to test this relationship in rice. The relationship between nonsynonymous SNPs at *SsIIa* and starch phenotypes (including SASS) in rice has been surveyed in *O. sativa*. However no statistical power was provided in those studies due to low sample size and the fact that population structure within samples was not considered (UMEMOTO *et al.* 2004; WATERS *et al.* 2006). It has been demonstrated that population structure within samples from association studies can cause spurious results (MARCHINI *et al.* 2004). Without knowing the population structure, it is difficult to distinguish the real association between genotype and phenotype from any false associations, which might result from different populations with different phenotypes. To avoid this potential problem, two approaches are available. The first includes information about population structure as covariate in association analyses. However, this approach requires genome-wide markers to calculate the relative kinship matrix (which reflects population structure) of the sampled materials (BRADBURY *et al.* 2007). The other approach is to perform an association analysis within each subpopulation when the population structure of the study system is known (MARCHINI *et al.* 2004). In this current study, I sampled both *O. rufipogon* and five *O. sativa* variety groups and performed association analysis in each rice subpopulation to exclude the effect of population structure on association analysis.

Here, I sampled 289 *O. sativa* and 57 *O. rufipogon* accessions. My objectives are to: 1), determine the starch quality difference among *O. rufipogon* and five rice variety groups by SASS; 2), analyze the evolutionary relationship of three nonsynonymous mutations at *SsIIa* exon 8; 3), determine the association between haplotypes at *SsIIa* exon 8 and starch alkali spreading score in *O. rufipogon* and each rice variety group.

Materials and Methods

Plant materials

Both *O. sativa* and *O. rufipogon* were sampled. Two other *Oryza* species, *O. barthii* and *O. meridionalis,* were included by a single accession to serve as outgroups. The collections are listed in Table 2.1. *O. rufipogon* sample collections cover the entire range except Australia. Most of the collections are from centers of its diversity: Thailand, India and China. The *O. sativa* collection includes five variety groups recently recognized by both SSR and chloroplast markers: *aus*, *indica*, *aromatic*, *tropical japonica*, and *temperate japonica*. Most of the collections are from three major variety groups (*indica, tropical japonica, temperate japonica*), which are grown widely throughout the world. Due to availability of DNA and seed materials, different numbers of collections were used for sequencing and phenotypic data because of the lack of *O. rufipogon* seeds or the difficulty of growing for seeds in greenhouse. Overlapping samples were used for the genotype-phenotype association analysis. A summary of collections is given in Table 2.2.

DNA extraction, Polymerase chain reaction (PCR) and Sequencing

DNA was extracted from dried leaves by a CTAB method with minor modifications (DOYLE and DOYLE 1990). Except for the *O. rufipogon* leaf materials collected from China, all the other *O. rufipogon* samples and all the *O. sativa* samples were obtained from International Rice Research Institute (IRRI) and grown for leaf materials in the greenhouse at Washington University in St. Louis. The samples of *O. rufipogon* from IRRI were self-fertilized in the greenhouse for two generations to decrease the degree of heterozygosity.

Primers were designed by the software Primer3 (http://frodo.wi.mit.edu/primer3/) from the Nipponbare genomic sequence available from Gramene (http://www.gramene.org/). PCRs were conducted in a Thermal Cycler TX2 or PTC-100. The PCR solutions include 1X Taq buffer, 2mM dNTP, 1 μM primers, 1 unit / 20 μl Taq polymerase, 2.5 mM $MgCl_2$, 1 μg template DNA and sterile deionized water, added to a volume of 20 μl. The following condition was used for PCRs: 95°C for 5 minutes; 30 cycles of 95°C for 50 seconds,53 or 58 °C for 1 minute (The annealing temperature for PCRs differs by primers.), and 72 °C for 2.5 minutes; 10 minutes of extension at 72 °C. Two pairs of primers (pair one: GCACTCCTGCCTGTTTATCTG, CGAGGCCACGGTGTAGTTG; pair two: CGGGAGAACGACTGGAAGATGAAC, CAGACACGAGAGCTAATGAAG) were designed and used for PCRs. The annealing temperature is 53°C for pair one, 58°C for pair two. The PCR products were cleaned using Exo1-

SAP commercial kits, then cycle-sequenced using BigDye Terminator chemistry (Applied Biosystems) and analyzed on an ABI 3130 capillary sequencer (Applied Biosystems).

Genetic diversity analysis

Sequences were aligned and manually adjusted with the software Biolign version 4.0.6.2 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). The sequences of Nipponbare were downloaded from Genbank and included in the analyses for *temperate japonica* (http://www.ncbi.nlm.nih.gov/Genbank/).

Statistics for the levels of variation (number of polymorphic synonymous, nonsynonymous and silent sites; pairwise nucleotide diversity, θ_{π}; average number of segregating sites, θ_W) (TAJIMA 1983; WATTERSON 1975) were performed in DnaSP version 5.0 (LIBRADO and ROZAS 2009). Silent sites include synonymous sites in coding region and all noncoding sites. Only silent sites were used for estimation of θ_{π} and θ_W. Neutrality tests, Tajima's D and Fay and Wu's H, were also performed in DNAsp version 5.0 (FAY and WU 2000; ROZAS *et al.* 2003; TAJIMA 1989). *Oryza meridionalis* served as an outgroup for Fay and Wu's H because the evolutionary relationship between *O. barthii* and *O. rufipogon* is too close to serve as an outgroup (see Fig 2.3). Coalescent simulations with 10,000 replications were conducted to determine statistical significance for Tajima's D and Fay and Wu's H.

Constructing haplotype network

Haplotype networks were constructed for all polymorphic sites and three nonsynonymous SNPs at *SsIIa* exon 8 respectively. Haplotype networks were constructed by medium joining method in software NETWORK version 4.5.1.0 (BANDELT *et al.* 1999). The program was run under default parameters. Haplotype frequencies were calculated for each rice group for all of the haplotypes found in the haplotype network for three nonsynonymous SNPs (SNP1, 2 and 3).

The difference of frequency distribution among rice groups were analyzed by nonparametric Friedman tests due to violation of the normality or equality of variance within the data, required for Analysis of Variance (ANOVA). Friedman test is a nonparamatric test, used to detect differences in three or more matched groups. Significant level of 5% was used.

Estimation of disintegration of rice endosperm starch in alkali solution

In order to exclude the effect of the environment on starch phenotype, all the accessions were grown to the seed stage in the greenhouse for seeds at Washington University in St. Louis. The seeds were harvested and dried in an incubator (Equatherm, Cruthin Matheson Scientific Inc.). The dried seeds were dehusked (Kett, model TR120) and polished (Kett pearlest grain polisher). Six polished seeds of each accession were placed in 1.5% KOH solution for 23 hours at room temperature. The degree of disintegration was quantified by a numerical scale of 1-7 as previously has been suggested in the study by Little *et al* (LITTLE *et al.* 1958). Sample pictures of starch phenotype in alkali are shown in Figure 2.4. Due to the availability of seeds, less than 6 seeds were used for some rice accessions. The average starch spreading score was calculated for each accession for subsequent analyses since 273 out of 325 rice accessions showed no variance within accessions (Fig 2.5).

In order to determine the difference of SASS among *O. rufipogon* and the rice varieties groups, both nonparametric Kruskal-Wallis tests and parametric one way analyses of variance (ANOVA) were performed. The difference of SASS among different haplotype groups (haploytpe is only determined by SNP1, 2 and 3) was also determined by an ANOVA and a Kruskal-Wallis test. The pairwise SASS difference among rice groups and haplotype groups were analyzed by student t tests. Significant level of 5% was used.

There are significant associations between SASS and gelatinization temperature of milled rice (JULIANO *et al.* 1964). Gelatinization temperature (GT) is an important parameter of rice cooking quality. It is the critical temperature at which the starch granules start to lose crystallinity by changing the starch surface from a polarized to a soluble state (KHUSH *et al.* 1979). The range of GT in rice has been classified into three groups: high (>74°C), medium (70-74°C), and low (<70°C) (BHATTACHARYA 1979). The range of SASSs were also classified into groups as its correspondence to GT. SASSs corresponding to GT are as follows: 1–2, low (GT: 74–80°C); 3–5, medium (GT: 70–74°C), and 6–7, high (GT < 70°C) (HE *et al.* 2006). In the above analysis, the SASS was considered as a quantitative trait with a numerical scale from 1 to 7. I did this classification to also analyze SASS as a qualitive trait. After the classification of SASS, the frequency of each SASS group was calculated for *O. rufipogon* and each varieties group respectively. Then these frequency distributions were compared among *O. rufipogon* and rice variety groups by Friedman tests. Significant level of 5% was used.

Genotype-phenotype association analysis

In order to search for nonsynonymous polymorphism sites that are responsible for starch disintegration diversity in rice, a general linear model and nested clade analysis were performed. Both analyses were performed in each of three major cultivated rice groups and *O. rufipogon* to exclude the confounding effect of population structure on association analysis. The *aus* and *aromatic* varieties were not included for analyses due to their low sample size.

The general linear model analysis generates a linear regression by allowing the linear model to be related to the variables via a function. In this study, it generates the linear model for nonsynonymous polymorphism and SASS. The general linear model was conducted in the program TASSEL (BRADBURY *et al.* 2007). Statistical significance was determined by 5% significant level with a Bonferroni correction.

The nested clade analysis was developed by Templeton et al. (TEMPLETON *et al.* 1987). This method requires a haplotype network to define a hierarchy of evolutionary clades. It starts from the tips of the network by nesting 'zero step clades' (the tip haplotypes) within 'one-step clades' (it is separated from tip haplotypes by one mutational change), and proceeds step by step until the final level of nesting includes the entire network. Here, we only need to nest ‘zero step clades’ into ‘one-step clades’ to include all the clades of the entire network for nonsynonymous polymorphic sites. Instead of haplotype network for all polymorphic sites in *SsIIa* exon 8, haplotype network for nonsynonymous polymorphic sites was used for nested clade analyses. The reason is that the nonsynonymous polymorphisms are the most possible candidate polymorphism for starch disintegration diversity in rice, and haplotype network for nonsynonymous polymorphic sites includes a decent number of samples for each clade for statistical analysis (KHUSH 2005). Zero step clades and their one-step clades were compared to determine if the nonsynonymous mutation between them caused phenotypic change. These comparisons were conducted by a non-parametric Kruskal-Wallis test with significance at the 5% level with a Bonferroni correction.

Results

Nucleotide variation at SsIIa exon 8 in rice

We examined the nucleotide variation in a 1.01 kilobase (FULTON *et al.*) region which covers almost the whole *SsIIa* exon 8 region (Fig 2.1) in domesticated Asian rice, *O. sativa* and its ancestor species, *O. rufipogon*. Since population structure has been observed in *O. sativa*, the nucleotide variation in *O. sativa* was examined respectively in five rice variety groups: *indica*, *tropical japonica*, *temperate japonica*, *aus* and *aromatic*.

The results of nucleotide variation in *O. sativa* and *O. rufipogon* are presented in Table 2.3. Both numbers of synonymous and silent polymorphisms are highest in *O. rufipogon* (nonsynonymous, 5; silent, 11) and lowest in *aus* (nonsynonymous or silent, 0). The estimation of silent diversity by θ_{π} and θ_{W} is the highest in *O. rufipogon* (θ_{π}= 0.00508 and θ_{W} = 0.00486), followed by *indica* (θ_{π} = 0.00421 and θ_{W} = 0.00362), *temperate japonica*, *tropical japonica*, *aromatic*, and is the lowest in *aus* (θ_{π} or θ_{W} = 0).

Five nonsynonymous SNPs were observed in *O. rufipogon* samples, three in *indica*, *tropical japonica* and *temperate japonica*, and one in *aus* and *aromatic* respectively. The three nonsynonymous SNPs observed in *tropical japonica* and *temperate japonica* were SNP 1, 2 and 3 (Fig 2.1). The nonsynonymous SNP observed in *aus* and *aromatic* is SNP 1. The three nonsysnonymous SNPs observed in *indica* include SNP 1, 3 and a singleton which only exists in one *indica* individual out of given samples. The five nonsynonymous SNPs observed in *O. rufipogon* include SNP 1 and 3. The other three nonsynonymous SNPs, which were only found in *O. rufipogon,* were in very low frequency. Among these three low frequency nonsynonymous SNPs, two were singletons, and found as a homozygote in one individual of *O. rufipogon* (The surrounding 20 bp of the nonsynonymous SNPs are CTGGAGGTGC***R***CGACGACGTG, CAAGTACAAG***S***AGAGCTGGAG). The other one is found as heterozygotes in two individuals of our *O. rufipogon* samples (The surrounding 20 bp of the nonsynonymous SNP is TGGACGTTCG*R*CCGCGCCGAG).

Among all rice groups studied here including *O. rufipogon*, Tajima's D value ranges from -1.05482 to 0.61507, and Fay and Wu's H value ranges from -9.42828 to 1.40351. No

Tajima's D value deviates from neutral expectation. Fay and Wu's H value deviates from neutral expectations only in *tropical japonica*.

Haplotype networks

The haplotype network based on all the mutations at *SsIIa* exon 8 was constructed to show the evolutionary relationships between rice variety groups and *O. rufipogon* (Fig 2.6). Haplotypes found in *O. sativa* are primarily a subset of haplotypes in *O. rufipogon*. However, some 'unique' haplotypes (haplotype D to M) were found in *O. sativa*. Moreover, haplotypes B and C have a high frequency in *O. sativa* but are in a low frequency in *O. rufipogon*.

The haplotype network based on SNP 1, 2 and 3 was constructed for nested clade analyses (Fig 2.7). The nonsynonymous SNPs which are only found in *indica* or *O. rufipogon* were not included in the network due to their extreme low frequency.

Five haplotypes in total were observed based on SNP 1, 2 and 3. Haplotype frequencies for *O. rufipogon* and each rice variety group are given in Figure 2.7. Haplotype H1 and H3 show high frequency in our samples. Haplotype H4 only exists in three *temperate japonica* and one *tropical japonica* individuals. Haplotype H5 corresponds to haplotype J in the haplotype network for all mutations in *SsIIa* exon 8, and only exists in one *tropical japonica* and one *indica* individual. Different rice groups showed different haplotype frequency distributions. *O. rufipogon*, *indica, aromatic* and *aus* have haplotype H1 in the highest frequency, while *tropical japonica* and *temperate japonica* have H3 and H2 with the highest frequency respectively, with the highest frequency. However, the results of Friedman tests showed no significant haplotype frequency differences among rice groups including *O. rufipogon* (Friedman chi square=4.058, degree of freedom = 5, P=0.54103).

Haplotype networks constructed by using all the mutations at *SsIIa* exon 8 or by SNP 1, 2 and 3 were shown in Figure 2.6 and 2.7 respectively. Loops were observed within the haplotype networks, which indicate ambiguous connections.

Starch quality difference among rice groups and among genotype groups

The frequency distribution of SASS in each rice group including *O. rufipogon* is shown in Figure 2.3. The SASS differences among rice groups were tested by both a Kruskal-Wallis test and an ANOVA test. The results of the Kruskal-Wallis test (degree of freedom=5, H =9.377, P=0.0949) were not significant. However, the results of the ANOVA tests (degree of freedom=5, sum of squares =39.747, mean square = 7.949, F = 2.6173, P = 0.02448) were significant. Pairwise comparisons by student's t test were conducted between rice groups including *O.rufipogon* (Fig 2.8). Based on a 5% significant level, significant differences were observed between *tropical japonica* and *aus*, and between *tropical japonica* and *O. rufipogon*. The mean value, standard error and standard deviation of SASS for each rice group is shown in Figure 2.8.

The frequency distribution of SASS categories in each rice group including *O. rufipogon* is shown in Figure 2.3. All rice groups have the highest frequency at the medium SASS category. The low SASS category showed similar frequency (18.92-25%) in all rice groups expect in *aromatic* (7.14%). The Medium SASS category showed the highest frequency in *aromatic* (92.86%), and the lowest frequency in *tropical japonica* (42.11%). The high SASS category showed zero frequency in *aromatic* and *aus*, and a very low frequency in *O. rufipogon* (5.41%), medium frequency in *indica* and *temperate japonica* (19.35%, 19.61%), the highest frequency in *tropical japonica* (38.6%). Although there are frequency differences at low and high SASS categories among the rice groups, the results of Friedman tests (degree of freedom = 5; Friedman chi square= 0.7692; P = 0.97895) are not significant.

The frequency distributions of different SASS categories in each haplotype group are shown in Figure 2.3. The mean value, standard error and standard deviation of SASS in each rice group is shown in Figure 2.8. The starch quality differences among haplotype groups were analyzed by a Kruskal-Wallis test and an ANVOA test. The results of Kruskal-Wallis test (degree of freedom=4, H =105.2615, P=0.0001) and ANOVA (degree of freedom=4, sum of squares =334.5960, mean square = 83.6490, F = 91.9706, P = 0.0001) are significant. The P values of student t tests for the pairwise comparisons among rice groups are listed in Figure 2.8. SASS is not significantly different between H1 and H2 haplotype groups, among H3, H4 and H5 groups respectively, but significantly different between haplotypes H1, H2 and haplotypes H3, H4, H5. Haplotype H3, H4 and H5 groups have high SASS while haplotype H1 and H2 groups have low and medium SASS. However, not all the individuals with haplotype H1 or H2 have low or medium SASS. One individual with haplotype H1 and one individual with haplotype H2 has high SASS. Not all the individuals with haplotype H3 have high SASS. One individual with H3 has medium SASS.

Association between phenotype and genotype

The result of an association analysis between starch phenotype and nonsynonymous mutations at *SsIIa* exon 8 by a general linear model is shown in Table 2.4. Based on a 5% significant level without a Bonferroni correction, there is significant association in *indica* and *temperate japonica* at SNP 1, in *temperate japonica* at SNP 2, in *indica, tropical japonica* and *temperate japonica* at SNP 3. However, only SNP 3 continues to be significantly associated with SASS in *indica*, *tropical japonica* and *temperate japonica* after a Bonferroni correction. The association analysis at SNP2 and 3 cannot be performed in *O. rufipogon* because of the lack of *O. rufipogon* samples with SNP2 and SNP3.

The results of an association analysis between phenotype and haplotype by a nested clade analysis are presented in Table 2.5. Haplotype H5 is not included in the nested clade analysis since it is most likely the result of recombination between H1 and H3 (see discussion). According to Templeton's nesting rule (TEMPLETON *et al.* 1987), haplotypes H1, H3 and H4 are 'zero step clades', and haplotype H2 is a 'first step clade'. All the zero step clades were nested within the first step clades, haplotype H2. Phenotypic comparisons between zero step clades and first step clades were performed to determine if the mutations between these zero and first steps are associated with phenotypic change. Only the phenotypic comparison between H2 and H3 is significant in *indica*, *tropical* and *temperate japonica* based on a 5% significant level without Bonferroni correction. However, this association continues to be significant only in *indica* and *tropical japonica* after Bonferroni correction. A comparison between H2 and H4 cannot be performed in *indica* due to the lack of *indica* samples with haplotype H4, Comparison between H2 and H3, H2 and H4 cannot be performed in *O. rufipogon* due to the lack of *O. rufipogon* samples with haplotype H3 and H4.

Discussion

Pattern of Nucleotide diversity at SsIIa exon 8

Previous studies have observed three nonsynonymous SNPs at *SsIIa* exon 8 in *O. sativa*, which were named as SNP 1, 2 and 3 here (see Fig 2.1) (UMEMOTO and AOKI 2005). Our study also detects these three SNPs. Besides these three nonsynonymous SNPs, one "novel" nonsynonymous SNP was found in *indica*. This suggests that more nonsynonymous SNPs at *SsIIa* exon 8 can be observed in *O. sativa* if more *O. sativa* individuals are sampled. However, since this novel nonsynonymous SNP is a singleton in our samples, it could be the result of PCR error. It was reported that PCR error rate is about 1bp error out of 10000bp (CARIELLO *et al.* 1991).

Our study also showed lower diversity (estimated by θ_{π} and θ_{W}), and lower number of synonymous and nonsynonymous SNPs in cultivated rice variety groups relative to that in *O. rufipogon*. This pattern is consistent with previous studies and has been attributed to bottleneck events during rice domestication (CAICEDO *et al.* 2007; ZHU *et al.* 2007). Most domestication events include population bottlenecks, which can drastically reduce diversity and result in a reduction of diversity in domesticated species relative to their wild ancestors (EYRE-WALKER *et al.* 1998). For example, the SNP diversity in maize is ~80% of that in its wild ancestor (ZHANG *et al.* 2002) an the SNP diversity in *O. sativa* is ~50% of that in it is wild ancestor, *O. rufipogon* (CAICEDO *et al.* 2007). Selection during rice domestication could also result lower diversity in cultivated rice than in *O. rufipgon*. However, no evidence of selection was found by previous study at *SsIIa* in any cultivated rice variety group (See Chapter One).

Starch quality is one of the most important agronomic traits. The region of gene *SsIIa* exon 8 has been suggested as a region that contributes to the starch quality differences between *indica* and *japonica* varieties (UMEMOTO *et al.* 2002). Therefore selection at *SsIIa* exon 8 in *indica* or *japonica* rice variety groups during domestication might be expected. However, no strong evidence of selection was found at *SsIIa* exon 8 in any cultivated rice group based on both Tajima's D or Fay and Wu's H tests. The significant value of Fay and Wu's H in *tropical japonica* suggests an excess of derived variants, which might suggest the misidentification of derived alleles. Evidence of genetic introgression of cultivated rice to/from the wild relatives has been observed in previous studies (SWEENEY and MCCOUCH 2007; SWEENEY *et al.* 2007). This process will cause the wild species carrying the cultivated allele or the cultivated species carrying the wild allele, and therefore result in the misidentification of derived alleles in cultivated rice variety groups.

Although no strong evidence of positive selection at *SsIIa* was found using most available tests, it does not necessarily mean that selection had not occurred during domestication. *SsIIa* may have been under selection in the past but becomes undetectable by the available tests of selection. The likelihood of detecting positive selection depends critically on the strength of selection, the time since fixation of the beneficial mutation, and the amount of recombination between the selected and neutral sites (BRAVERMAN *et al.* 1995; PRZEWORSKI 2002; PRZEWORSKI 2003). In addition to these factors, several other demographic factors including bottleneck events and population expansion also complicate tests for selection in domesticated species (WRIGHT and GAUT 2005). Although we use the genome-wide sequence information as a control for demographic history, the lack of a most likely demographic model, which reflects the most likely demographic scenarios of rice, still makes detection of selection difficult (RAMOS-ONSINS *et al.* 2008; WRIGHT and GAUT 2005).

Haplotype network of SsIIa exon 8

Loops, which are observed within the haplotype networks, indicate ambiguous connections. Such ambiguity in a haplotype network may be due to recombination or recurrent mutations. Haplotypes J and H5 only exist in one *indica* and one *tropical japonica* individual among our samples while the other haplotypes in the loop exist in *O. rufipogon* or had higher frequency than Haplotypes J and H5. Therefore, haplotype J and H5 are more likely the result of recombination between SNP 1 and 3 or the recurrent mutation of SNP 1 and 3 in *O. sativa*. Again, given the short period of time since rice domestication, it is unlikely the result of recurrent mutations in *O. sativa* (KHUSH 1997a).

Haplotype A is considered the ancestor to haplotypes B, C and E for the following reasons. First, haplotype A has closer evolutionary relationship with outgroup species than haplotype B, C and E. Second, haplotype A has higher frequency in wild ancestor species, *O. rufipogon* (Fig 2.8). Haplotypes B, C and E have lower frequency in *O. rufipogon*. Third, internal nodes are evolutionary older than tip nodes. Haplotypes A and B are the internal node while haplotype C and E are tip nodes since haplotype J is possibly a recombinant. Haplotypes H1, H2, H3 and H4 in the haplotype network by the three nonsynonymous SNPs (SNP1, 2 and 3) correspond to the haplotype A, B, C and E in the haplotype network by all SNPs respectively. Therefore, Haplotype H1 is the ancestor of haplotypes H2, H3, and H4.

Different rice groups including *O. rufipogon* show different haplotype frequency distributions (the haplotypes based on nonsynonymous SNP1, 2 and 3) although differences are not statistically significant based on Friedman tests (Fig 2.9). Within *O. rufipogon*, ancestor haplotype H1 has higher frequency than haplotypes H2, H3. This can be simply due to the ancestor haplotype having more time to increase its frequency than the derived haplotypes H2 and H3. *Aus*, *indica* and *aromatic* also have the same haplotype distribution pattern as that of *O. rufipogon*. Previous studies suggest that *aus* and *indica* rice might be domesticated from *O. rufipogon* independently (LONDO *et al.* 2006). The origin of *aromatic* remains unclear. The frequency distribution pattern in *aus*, *indica* and *aromatic* rice might be derived from their ancestor, *O. rufipogon*. *Tropical* and *temperate japonica* rice have higher frequencies of the derived haplotypes than the ancestor of haplotypes. This suggests that haplotype distribution were altered during domestication for *japonica* rice or that these derived haplotypes were selected for in *japonica* rice. However no evidence of selection on the gene *SsIIa* has been discovered in *tropical* or *temperate japonica* (Chapter 1). Therefore, the different frequency distribution pattern among rice variety groups might be simply due to the results of domestication events.

Evolution of starch quality during rice domestication

Previous studies have indicated that the quality of starch between *indica* and *japonica* is distinguishable on the basis of the disintegration of starch in alkali (WARTH and DARABSETT 1914). Starches of *japonica* varieties tend to degraded easily in 1.5% KOH solution (high SASS) while starches of *indica* varieties are resistant (low SASS) (UMEMOTO *et al.* 2004). Previous studies considered *aus* and *indica* as *indica*, *tropical* and *temperate japonica* as *japonica* (KHUSH 1997b; WARTH and DARABSETT 1914). Our study also indicated SASS difference between *indica* and *tropical japonica*, *aus* and *tropical japonica*. *Tropical japonica* has more individuals with high SASS category than *indica* and *aus* (Fig 2.3, 38.6% vs 19.35% and 0). However, the difference is not statistically significant. Also, no SASS difference between *indica* and *temperate japonica* was observed. This might be due to our limited sampling in *temperate japonica*. *Temperate japonica* is mainly grown in temperate region of China (KHUSH 1997b). However none of our *temperate japonica* samples were from China.

The starch quality difference between *indica* and *japonica* rice has been shown not only by SASS. More quantitative methods for rice endosperm starch quality are available (UMEMOTO *et al.* 2004). For example, gelatinisation temperature, which is the critical temperature at which the starch granules start to lose crystallinity by changing the starch surface from a plorized to a soluble state, is also an important parameter of rice cooking quality. There is also direct estimates of starch quality such as amylose content, amylopectin structure (NAKAMURA *et al.* 1997). Here, I do not show statistical difference of starch quality between *indica* and *japonic* rice as previously reported (KHUSH 1997b). Furthur research with more estimates of starch quality are required to study the starch quality differeces among rice groups including *O. rufipogon*.

Association between nonsynonymous SNPs and SASS in rice

Previous studies in *E. coli* suggest that mutations 2 and 3 alter *SsIIa* enzyme activity from high to low while mutation 1 does not (NAKAMURA *et al.* 2005). Previous studies have also suggested that mutations 2 and 3 cause starch quality (such as GT, SASS) changes in rice while mutation 1 does not. However, this conclusion is not statistically supported due to low sample size and did not consider population structure (NAKAMURA *et al.* 2005; UMEMOTO and AOKI 2005; WATERS *et al.* 2006). Both a general linear model and nested clade analyses in my study statistically support that mutation 3 alters the SASS from low to high, which is consistent with previous studies. As previous studies have suggested, mutation 2 has an effect and in our study appears to cause a change from low to high. However, this is not statistically supported due to the lack of samples with H4 haplotype. Halotype H4 is frequently observed in *temperate japonica* which is undersampled here. Therefore, in order to know the effect of mutation 2 on starch quality in rice, further studies with more *temperate japonica* samples from China and Japan are required (see Study System).

Why does mutation 2 or 3 at *SsIIa* exon 8 region cause higher SASS in rice? The *SsIIa* exon 8 region encodes for the C terminal of *SsIIa*. The C terminal residue of *SsIIa* enzyme has been identified to be critical for substrate binding and catalysis in maize (GAO *et al.* 2004; NICHOLS *et al.* 2000). It was also suggested that mutation 2 or 3, which result in amino acid change at the 737 or 781 of *SsIIa* enzyme respectively, most likely alter the *SsIIa* enzyme both in terms of activity and starch granule association (see Fig 2.1) (UMEMOTO and AOKI 2005). The low *SsIIa* enzyme activity or lower starch granule association will result in S-type amylopectin in rice, therefore high SASS (NAKAMURA *et al.* 2005).

My study suggests that the nonsynonymous SNP 3 in *SsIIa* exon 8 is the major SNP contributing to SASS variation in rice. However there is no 100% association between haplotypes and SASS. This suggests that SNPs in other region of *SsIIa* or other genes involved in starch synthesis pathway are interacting with SNP 3 or 2 and also play a role in SASS variation in rice. Those SNPs may be rare because only one out of 64 individuals with H1 and one out of 45 individuals with H2 do not have low or medium SASS, and only two out of 51 individuals with H3 do not have the high SASS. Two individuals with H1 or H2 showed high SASS, suggesting that rare mutation in other genes or other region of *SsIIa* could also result in S-type amylopectin with its corresponding high SASS. The two individuals with H3 showed medium SASS. This suggests that these two individuals have other mutations, which could cause high content of amylose or L-type amylopectin. Currently, over 20 genes involved in the starch synthesis pathway have been identified (MYERS *et al.* 2000). For example, *Waxy* is the major gene involved in amylose production in rice endosperm (see Chapter One). A mutation at the intron 1 region of *Waxy* causes alternative splicing of *Waxy* and results undetectable level of starch synthase (BLIGH *et al.* 1998; WANG *et al.* 1995). To date, no other mutations which could also cause S-type or high content of amylose or L-type amylopectin in rice have been identified. Further studies with more rice accessions and more starch candidate genes will be necessary to fully understand variation in rice endosperm starch quality.

Conclusions

I have shown the pattern of diversity at *SsIIa* exon 8 region in *O. rufipogon* and five cultivated rice variety groups. In addition to three previously identified nonsynonymous mutations at *SsIIa* exon 8, I found three additional nonsynonymous mutations in the wild ancestor of rice, *O. rufipogon.* These newly discovered alleles were in very low frequency in *O. rufipogon.* Previous studies in *E. coli* suggest that mutation 1 will not change SsIIa enzyme activity while mutation 2 and 3 will affect enzyme activities (please see Fig 2.1 for the location of mutation 1, 2 and 3). The haplotype network of *SsIIa* exon 8 alleles suggests that mutation 1 is evolutionary older than mutations 2 and 3. SASS comparison among rice groups showed that *tropical japonica* is different from all other rice groups with more high SASS category individuals. However, no statistical difference of SASS among rice groups is found. Genotype and phenotype association by both a general linear model and nested clade analyses indicate that mutation 3 contributes to SASS diversity in rice and mutation 1 does not, which is statistically supported. I also observe that mutation 2 causes SASS from low to high in my samples. However, this result is tentative since statistical support is lacing due to few accessions which contain mutation 2. While there a significant association between genotype and phenotype. The relationship is not absolute. My study suggests that more samples and more candidate genes are required in order to understand the genetic basis of SASS and hence starch quality diversity in rice.

References

AMUNDADOTTIR, L., P. KRAFT, R. Z. STOLZENBERG-SOLOMON, C. S. FUCHS, G. M. PETERSEN *et al.*, 2009 Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nature Genetics **41:** 986-U947.

BANDELT, H. J., P. FORSTER and A. ROHL, 1999 Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution **16:** 37-48.

BENTSINK, L., J. JOWETT, C. J. HANHART and M. KOORNNEEF, 2006 Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **103:** 17042-17047.

BHATTACHARYA, K. R., 1979 *Gelatinization temperature rice starch and its determination*. IRRI, Laguna, Los Banos, Philippines.

BLIGH, H. F. J., P. D. LARKIN, P. S. ROACH, C. A. JONES, H. Y. FU *et al.*, 1998 Use of alternate splice sites in granule-bound starch synthase mRNA from low-amylose rice varieties. Plant Molecular Biology **38:** 407-415.

BRADBURY, P. J., Z. ZHANG, D. E. KROON, T. M. CASSTEVENS, Y. RAMDOSS *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The Hitchhiking Effect on the Site Frequency-Spectrum of DNA Polymorphisms. Genetics **140:** 783-796.

BULÉON, A., P. COLONNA, V. PLANCHOT and S. BALL, 1998 Starch granules: structure and biosynthesis. International Journal of Biological Macromolecules **23:** 85-112.

CAICEDO, A., S. WILLIAMSON, R. D. HERNANDEZ, A. BOYKO, A. FLEDEL-ALON *et al.*, 2007 Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLoS Genetics **3:** e163.

CARIELLO, N. F., J. A. SWENBERG and T. R. SKOPEK, 1991 Fidelity of Thermococcus-Litoralis DNA-Polymerase (Vent) in Pcr Determined by Denaturing Gradient Gel-Electrophoresis. Nucleic Acids Research **19:** 4193-4198.

DIAMOND, J., 2002 Evolution, consequences and future of plant and animal domestication. Nature **418:** 700-707.

DOYLE, J. J., and J. L. DOYLE, 1990 Isolation of plant DNA from fresh tissue. Focus **12:** 13-15.

EYRE-WALKER, A., R. L. GAUT, H. HILTON, D. L. FELDMAN and B. S. GAUT, 1998 Investigation of the bottleneck leading to the domestication of maize. Proceedings of the National Academy of Sciences of the United States of America **95:** 4441-4446.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

FULTON, T. M., T. BECKBUNN, D. EMMATTY, Y. ESHED, J. LOPEZ *et al.*, 1997 QTL analysis of an advanced backcross of Lycopersicon peruvianum to the cultivated tomato and comparisons with QTLs found in other wild species. Theoretical and Applied Genetics **95:** 881-894.

GAO, Z., P. KEELING, R. SHIBLES and H. P. GUAN, 2004 Involvement of lysine-193 of the conserved "K-T-G-G" motif in the catalysis of maize starch synthase IIa. Archives of Biochemistry and Biophysics **427:** 1-7.

GAO, Z. Y., D. L. ZENG, X. CUI, Y. H. ZHOU, M. YAN *et al.*, 2003 Map-based cloning of the ALK gene, which controls the gelatinization temperature of rice. Science in China Series C-Life Sciences **46:** 661-668.

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.

HARLAN, J. R., 1992 *Crops and Man*. American Society of Agronomy, Madison, WI.

HE, Y., Y. HAN, L. JIANG, C. XU, J. LU *et al.*, 2006 Functional analysis of starch-synthesis genes in determining rice eating and cooking qualities. Molecular Breeding **18:** 277-290.

JULIANO, B. O., G. M. BAUTISTA, J. C. LUGAY and A. C. REYES, 1964 Studies on the physicochemical properties of rice. Journal of agricultural and food chemistry **12:** 131-138.

JULIANO, B. O., and C. P. VILLAREAL, 1993 *Grain quality evaluation of world rices*. International Rice Research Institute, Los Baños, Laguna, Philippines.

KHUSH, G. S., 1997a Origin, dispersal, cultivation and variation of rice. Plant Molecular Biology **V35:** 25-34.

KHUSH, G. S., 1997b Origin, dispersal, cultivation and variation of rice. Plant Molecular Biology **35:** 25-34.

KHUSH, G. S., 2005 What it will take to Feed 5.0 Billion Rice consumers in 2030. Plant Molecular Biology **59:** 1-6.

KHUSH, G. S., C. M. PAULE and N. M. DE LA CRUZ, 1979 *Rice grain quality evaluation and improvement at IRRI*. IRRI, Laguna, Los Banos, Philippines.

KONISHI, S., T. IZAWA, S. Y. LIN, K. EBANA, Y. FUKUTA *et al.*, 2006 An SNP caused loss of seed shattering during rice domestication. Science **312:** 1392-1396.

KOVACH, M. J., M. N. CALINGACION, M. A. FITZGERALD and S. R. MCCOUCH, 2009 The origin and evolution of fragrance in rice (Oryza sativa L.). Proceedings of the National Academy of Sciences of the United States of America **106:** 14444-14449.

KUDO, M., 1968 Genetical and thremmatological studies of characters, physiological or ecological, in the hybrids between ecological rice groups. Bulletin of the National Institute of Agricultural Sciences **Series D:** 1-84.

LI, C., and T. SANG, 2006 Rice Domestication by Reducing Shattering. Science **311:** 1936-1939.

LIBRADO, P., and J. ROZAS, 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25:** 1451-1452.

LIN, S. Y., T. SASAKI and M. YANO, 1998 Mapping quantitative trait loci controlling seed dormancy and heading date in rice, Oryza sativa L., using backcross inbred lines. Theoretical and Applied Genetics **96:** 997-1003.

LITTLE, R., G. HILDER and E. DAWSON, 1958 Differential effect of dilute alkali on 25 varieties of milled white rice. Cereal Chemistry **35:** 111-126.

LONDO, J. P., Y. C. CHIANG, K. H. HUNG, T. Y. CHIANG and B. A. SCHAAL, 2006 Phylogeography of Asian wild rice, Oryza rufipogon, reveals multiple independent domestications of cultivated rice, Oryza sativa. Proceedings of the National Academy of Sciences of the United States of America **103:** 9578-9583.

LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Sunderland, MA.

MARCHINI, J., L. R. CARDON, M. S. PHILLIPS and P. DONNELLY, 2004 The effects of human population structure on large genetic association studies. Nature Genetics **36:** 512-517.

MATHER, K. A., A. L. CAICEDO, N. R. POLATO, K. M. OLSEN, S. MCCOUCH *et al.*, 2007 The extent of linkage disequilibrium in rice (Oryza sativa L.). Genetics **177:** 2223-2232.

MORISHIMA, H., Y. SANO and H. I. OKA, 1992 Evolutionary studies in cultivated rice and its wild relatives. Oxford Surveys in Evolutionary Biology **8:** 135-184.

MYERS, A. M., M. K. MORELL, M. G. JAMES and S. G. BALL, 2000 Recent progress toward understanding biosynthesis of the amylopectin crystal. Plant Physiology **122:** 989-997.

NACHMAN, M. W., H. E. HOEKSTRA and S. L. D'AGOSTINO, 2003 The genetic basis of adaptive melanism in pocket mice. Proceedings of the National Academy of Sciences of the United States of America **100:** 5268-5273.

NAKAMURA, Y., P. B. FRANCISCO, Y. HOSAKA, A. SATO, T. SAWADA *et al.*, 2005 Essential amino acids of starch synthase IIa differentiate amylopectin structure and starch quality between japonica and indica rice varieties. Plant Molecular Biology **58:** 213-227.

NAKAMURA, Y., A. KUBO, T. SHIMAMUNE, T. MATSUDA, K. HARADA *et al.*, 1997 Correlation between activities of starch debranching enzyme and α-polyglucan structure in endosperms of sugary-1 mutants of rice. The plant journal **12:** 143-153.

NAKAMURA, Y., A. SATO and B. O. JULIANO, 2006 Short-chain-length distribution in debranched rice starches differing in gelatinization temperature or cooked rice hardness. Starch-Starke **58:** 155-160.

NICHOLS, D. J., P. L. KEELING, M. SPALDING and H. P. GUAN, 2000 Involvement of conserved aspartate and glutamate residues in the catalysis and substrate binding of maize starch synthase. Biochemistry **39:** 7820-7825.

O'CONNOR, T. D., and N. I. MUNDY, 2009 Genotype-phenotype associations: substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. Bioinformatics **25:** I94-I100.

OKA, H., and H. MORISHIMA, 1997 *Wild and cultivated rice*. Genetics, Nobunkyo, Tokyo.

OLSEN, K. M., A. L. CAICEDO, N. POLATO, A. MCCLUNG, S. MCCOUCH *et al.*, 2006 Selection under domestication: Evidence for a sweep in the rice Waxy genomic region. Genetics **173:** 975-983.

PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179-1189.

PRZEWORSKI, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164:** 1667-1676.

RAMOS-ONSINS, S. E., E. PUERMA, D. BALANA-ALCAIDE, D. SALGUERO and M. AGUADE, 2008 Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in Arabidopsis thaliana. Molecular Ecology **17:** 1211-1223.

ROZAS, J., J. C. SÀNCHEZ-DELBARRIO, X. MESSEQUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496-2497.

SWEENEY, M., and S. MCCOUCH, 2007 The Complex History of the Domestication of Rice. Annals of Botany **100:** 951-957.

SWEENEY, M. T., M. J. THOMSON, Y. G. CHO, Y. J. PARK, S. H. WILLIAMSON *et al.*, 2007 Global Dissemination of a Single Mutation Conferring White Pericarp in Rice. PLoS Genetics **3:** e133.

TAJIMA, F., 1983 Evolutionary Relationship of DNA-Sequences in Finite Populations. Genetics **105:** 437-460.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

TEMPLETON, A. R., E. BOERWINKLE and C. F. SING, 1987 A Cladistic-Analysis of Phenotypic Associations with Haplotypes Inferred from Restriction Endonuclease Mapping .1. Basic Theory and an Analysis of Alcohol-Dehydrogenase Activity in Drosophila. Genetics **117:** 343-351.

UMEMOTO, T., and N. AOKI, 2005 Single-nucleotide polymorphisms in rice starch synthase IIa that alter starch gelatinisation and starch association of the enzyme. Functional Plant Biology **32:** 763-768.

UMEMOTO, T., N. AOKI, H. X. LIN, Y. NAKAMURA, N. INOUCHI *et al.*, 2004 Natural variation in rice starch synthase IIa affects enzyme and starch properties. Functional Plant Biology **31:** 671-684.

UMEMOTO, T., M. YANO, H. SATOH, A. SHOMURA and Y. NAKAMURA, 2002 Mapping of a gene responsible for the difference in amylopectin structure between japonica-type and indica-type rice varieties. Theoretical and Applied Genetics **104:** 1-8.

WANG, Z. Y., F. Q. ZHENG, G. Z. SHEN, J. P. GAO, D. P. SNUSTAD *et al.*, 1995 The Amylose Content in Rice Endosperm Is Related to the Posttranscriptional Regulation of the Waxy Gene. Plant Journal **7:** 613-622.

WARTH, F., and D. DARABSETT, 1914 disintegration of rice grains by means of alkali. Bulletin of Agricultural Research Institute **38:** 1-9.

WATERS, D. L. E., R. J. HENRY, R. F. REINKE and M. A. FITZGERALD, 2006 Gelatinization temperature of rice explained by polymorphisms in starch synthase. Plant Biotechnology Journal **4:** 115-122.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biololgy **7:** 256-276.

WHITT, S. R., L. M. WILSON, M. I. TENAILLON, B. S. GAUT and E. S. BUCKLER, 2002 Genetic diversity and selection in the maize starch pathway. Proceedings of the National Academy of Sciences of the United States of America **99:** 12959-12962.

WILSON, L. M., S. R. WHITT, A. M. IBANEZ, T. R. ROCHEFORD, M. M. GOODMAN *et al.*, 2004 Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell **16:** 2719-2733.

WRIGHT, S. I., and B. S. GAUT, 2005 Molecular Population Genetics and the Search for Adaptive Evolution in Plants. Molecular Biology and Evolution **22:** 506-519.

ZHANG, L. Q., A. S. PEEK, D. DUNAMS and B. S. GAUT, 2002 Population genetics of duplicated disease-defense genes, hm1 and hm2, in maize (Zea mays ssp mays L.) and its wild ancestor (Zea mays ssp parviglumis). Genetics **162:** 851-860.

ZHU, Q., X. ZHENG, J. LUO, B. S. GAUT and S. GE, 2007 Multilocus Analysis of Nucleotide Variation of Oryza sativa and Its Wild Relatives: Severe Bottleneck during Domestication of Rice. Molecular Biology and Evolution **24:** 875-888.

Table 2.1), Collections of *O. rufipogon* and *O. sativa* from IRRI and field used in the study

IRRI #[a]	Cultivar Name[b]	Species[c]	Race[d]	origin[e]	DNA Label[f]	phenotype[g]	genotype[h]
29016	Aus 196	*sativa*	aus	Bangledesh	U_BD_AI_01	yes	
64773	Dharia	*sativa*	aus	Bangledesh	L_BD_AI_03	yes	
	ARC 7046	*sativa*	aus	India	L_IN_AI_01	yes	
29046	Aus 254	*sativa*	aus	Bangledesh	U_BD_AI_02	yes	
64772	Chondoni	*sativa*	aus	Bangledesh	L_BD_AI_02	yes	
29241	Aus 463	*sativa*	aus	Bangledesh	U_BD_AI_03	yes	
20461	ARC 7046	*sativa*	aus	India	U_IN_AI_01	yes	
21289	ARC 11287	*sativa*	aus	India	U_IN_AI_02	yes	yes
22739	Bei Khe	*sativa*	aus	Cambodia	L_KH_AI_04	yes	yes
64771	Chikon Shoni	*sativa*	aus	Bangledesh	L_BD_AI_01	yes	yes
66765	Asha	*sativa*	aus	Bangledesh	L_BD_AI_06	yes	yes
66828	Tepi Borua	*sativa*	aus	Bangledesh	L_BD_AI_09	yes	yes
67700	Bijri	*sativa*	aus	India	L_IN_AI_06	yes	yes
CIor 5987	Ramgarh	*sativa*	aus	India, Bihar	L_IN_AI_02	yes	yes
CIor 12374	P166	*sativa*	aus	China, Sichuan	C_CN_AI_04	yes	yes
3397	Hashikalmi	*sativa*	aus	Suriname	L_SR_AI_01	yes	
9466	Amarelao	*sativa*	indica	Brazil	C_BR_I_02	yes	yes
PI 584594	Che Eu Hung	*sativa*	indica	China	C_CN_I_03	yes	
	Cica 4	*sativa*	indica	Colombia	C_CO_I_01	yes	
CIor 5256	Arroz en Granza	*sativa*	indica	Guatemala	C_GT_I_01	yes	
	CO 39	*sativa*	indica	India, Orissa	C_IN_I_03	yes	yes
6663		*sativa*	Indica	India	C_IN_I_04	yes	yes
PI 584560	Gerdeh	*sativa*	indica	Iran	C_IR_I_01	yes	yes
	IR45	*sativa*	indica	Philippines	C_PH_I_01	yes	
	M1-48	*sativa*	Indica	Philippines	C_PH_I_02	yes	
	IR8	*sativa*	Indica	Philippines	C_PH_I_04	yes	
	UPL RI-5	*sativa*	indica	Philippines	C_PH_I_06	yes	yes
	IR20	*sativa*	indica	Philippines, Luzon	C_PH_I_09	yes	yes
	IR36	*sativa*	indica	Philippines, Luzon	C_PH_I_10	yes	yes
	Balislus	*sativa*	indica	Senegal	C_SN_I_02	yes	
	Daw Pao	*sativa*	indica	Thailand	C_TH_I_02	yes	
15058	KU188	*sativa*	indica	Thailand	C_TH_I_04	yes	yes
8240		*sativa*	Indica	Taiwan	C_TW_I_06	yes	yes
	Tunsart	*sativa*	indica	Vietnam	C_VN_I_01	yes	yes
PI 5845548	Chiem Chanh	*sativa*	indica	Vietnam	C_VN_I_02	yes	yes
5868	Doc Phung Lun	*sativa*	indica	Vietnam	C_VN_I_03	yes	yes
66770	Bamura	*sativa*	indica	Bangledesh	L_BD_I_07	yes	yes
27513		*sativa*	Indica	Bangladesh	L_BD_I_10	yes	yes
75782	Noumoufiedougou	*sativa*	indica	Burkina Faso	L_BF_I_01	yes	yes

67859	Zakha	*sativa*	indica	Bhutan	L_BT_I_07	yes	yes
PI 503152	Du Jiang Yu	*sativa*	indica	China, Beijing	L_CN_I_01	yes	
1000	Gin Goo Sun	*sativa*	indica	China	L_CN_I_02	yes	
1051	Hsiang Tao	*sativa*	indica	China	L_CN_I_03	yes	
PI 608412	Xu Xu Zhan	*sativa*	indica	China, Sichuan	L_CN_I_05	yes	
66463	Younoussa	*sativa*	indica	Guinea	L_GW_I_01	yes	yes
18804	Siam Kuning	*sativa*	indica	Indonesia	L_ID_I_04	yes	
66535	Alur Kuning	*sativa*	indica	Indonesia	L_ID_I_05	yes	
66626	Rom Putih	*sativa*	indica	Indonesia	L_ID_I_07	yes	yes
67459	Caka Putih	*sativa*	indica	Indonesia	L_ID_I_09	yes	
67460	Gembira Kuning	*sativa*	indica	Indonesia	L_ID_I_10	yes	yes
74625	Gembira Kuning	*sativa*	indica	Indonesia	L_ID_I_11	yes	yes
74642	Padi Hanyut	*sativa*	indica	Indonesia	L_ID_I_12	yes	
77622	Pare Pulu Siam	*sativa*	indica	Indonesia	L_ID_I_15	yes	yes
43545		*sativa*	Indica	Indonesia (E. Kalimantan)	L_ID_I_17	yes	yes
PI 376255	ARC 11956	*sativa*	indica	India, Arunachal Pradesh	L_IN_I_04	yes	yes
67708	Dudh Malai	*sativa*	indica	India	L_IN_I_07	yes	
CIor 6742	Tosa Bozu	*sativa*	indica	Japan, Kyoto	L_JP_I_01	yes	
PI 389291	Aikoku	*sativa*	indica	Japan, Okinawa	L_JP_I_02	yes	yes
12110	Phcar Tien	*sativa*	indica	Cambodia	L_KH_I_01	yes	
22716	Ang Kang KP	*sativa*	Indica	Cambodia	L_KH_I_02	yes	yes
22722	Angroeus	*sativa*	indica	Cambodia	L_KH_I_03	yes	
81421	Muoy Doy Day	*sativa*	indica	Cambodia	L_KH_I_05	yes	yes
8948	Pokkali	*sativa*	indica	Sri Lanka	L_LK_I_02	yes	yes
78911	Eimayaebaw	*sativa*	indica	Myanmar	L_MM_I_01	yes	yes
78916	Let Yone Gyi	*sativa*	indica	Myanmar	L_MM_I_02	yes	yes
82100	Kaukpwa	*sativa*	indica	Myanmar	L_MM_I_07	yes	
71508	Batu	*sativa*	indica	Malaysia	L_MY_I_04	yes	yes
71540	Kiodung	*sativa*	indica	Malaysia	L_MY_I_05	yes	yes
11431	Karma	*sativa*	indica	Nepal	L_NP_I_01	yes	yes
11443	Ramdulari	*sativa*	indica	Nepal	L_NP_I_02	yes	yes
58930		*sativa*	Indica	Nepal	L_NP_I_04	yes	yes
CIor 4637	Lupa	*sativa*	indica	Philippines	L_PH_I_02	yes	yes
CIor 4957	Macunting	*sativa*	indica	Philippines	L_PH_I_03	yes	
78092	IR24632-34-2	*sativa*	indica	Philippines	L_PH_I_06	yes	yes
26872		*sativa*	Indica	Philippines	L_PH_I_07	yes	yes
PI 392174	Torh	*sativa*	indica	Pakistan, Sind	L_PK_I_01	yes	
78244	Jao'Mali	*sativa*	indica	Thailand	L_TH_I_05	yes	
78250	Khao Gu Lahb	*sativa*	indica	Thailand	L_TH_I_07	yes	
IRGC 201	Doc Phung	*sativa*	indica	Vietnam	L_VN_I_01	yes	
209	Soc Nau	*sativa*	indica	Vietnam	L_VN_I_02	yes	yes

7074	Cadung Gocong	*sativa*	indica	Vietnam	L_VN_I_03	yes	
10242	Lua Noi	*sativa*	indica	Vietnam	L_VN_I_04	yes	yes
47473	Canh Nong Nghe An	*sativa*	indica	Vietnam	L_VN_I_06	yes	yes
73201	Khau Danh	*sativa*	indica	Vietnam	L_VN_I_07	yes	yes
78360	Nep Cham Ca Hoa Binh	*sativa*	indica	Vietnam	L_VN_I_09	yes	
79070	Chet Cut	*sativa*	Indica	Vietnam	L_VN_I_12	yes	yes
79078	Nep Dai Loan	*sativa*	indica	Vietnam	L_VN_I_15	yes	yes
56036		*sativa*	Indica	Vietnam	L_VN_I_16	yes	yes
15174	Bokolon	*sativa*	indica	Ivory Coast(Cote D'Ivoire)	U_CI_I_02	yes	
57627	Americain	*sativa*	indica	Cote D'Ivoire	U_CI_I_05	yes	yes
56724	BS017	*sativa*	indica	Guinea-Bissau	U_GN_I_01	yes	yes
16159	Dudhe	*sativa*	indica	Nepal	U_NP_I_01	yes	
56221	Ebandioulaye	*sativa*	indica	Senegal	U_SN_I_03	yes	
56264	Adiallo	*sativa*	indica	Senegal	U_SN_I_04	yes	
27748		*sativa*	Indica	Thailand	U_TH_JV_N06	yes	
56160	ES002	*sativa*	indica	Tanzania	U_TZ_I_01	yes	
43369		*sativa*	Indica	Indonesia (S. Sumatra)		yes	
66970*		*sativa*	Indica	Philippines		yes	
8952		*sativa*	Indica	Sri Lanka		yes	
PI 279131		*sativa*	Indica	Taiwan		yes	
12425		*sativa*	Indica	India		yes	
58892		*sativa*	Indica	Nepal		yes	
51400		*sativa*	Indica	China		yes	
45011		*sativa*	Indica	India/Bangladesh		yes	
46202		*sativa*	Indica	India		yes	
51300		*sativa*	Indica	China		yes	
PI 280681		*sativa*	Indica	Philippines		yes	
12995		*sativa*	Indica	Loas		yes	
51250		*sativa*	Indica	China		yes	
	IAC 25	*sativa*	indica	Brazil, Rio Grande do Sul	C_BR_I_03		yes
	SP-1	*sativa*	indica	Thailand	C_TH_I_05		yes
PI 584575	Canella De Ferro	*sativa*	temp. Jap.	Brazil	C_BR_JP_01	yes	yes
	OS6	*sativa*	temp. Jap.	Zaire	C_CD_JP_01	yes	
	OS4	*sativa*	temp. Jap.	Zaire	C_CD_JP_02	yes	
	IRAT 104	*sativa*	temp. Jap.	Cote D'Ivoire	C_CI_JP_01	yes	
	63-83	*sativa*	temp. Jap.	Cote D'Ivoire	C_CI_JP_02	yes	yes
	IRAT 13	*sativa*	temp. Jap.	Cote D'Ivoire	C_CI_JP_03	yes	yes
16856	Monolaya	*sativa*	temp. Jap.	Colombia	C_CO_JP_02	yes	
	Koshi Hikari	*sativa*	temp. Jap.	Japan, Fukui	C_JP_JP_02	yes	

	Koshihikari	*sativa*	temp. Jap.	Japan, Fukui	C_JP_JP_03	yes	yes
	Elliott	*sativa*	temp. Jap.	Liberia	C_LR_JP_01	yes	yes
6457	Elliott	*sativa*	temp. Jap.	Liberia	C_LR_JP_02	yes	
	OS4	*sativa*	temp. Jap.	Nigeria	C_NG_JP_01	yes	
	Palawan	*sativa*	temp. Jap.	Philippines	C_PH_JP_05	yes	yes
328	Azucena	*sativa*	temp. Jap.	Philippines	C_PH_JP_07	yes	
	E425	*sativa*	temp. Jap.	Senegal	C_SN_JP_01	yes	
CIor 1496	Wanica	*sativa*	temp. Jap.	Suriname	C_SR_JP_01	yes	yes
	MayBelle	*sativa*	temp. Jap.	USA	C_US_JP_02	yes	
	Alan	*sativa*	temp. Jap.	USA	C_US_JP_03	yes	
	Bengal	*sativa*	temp. Jap.	USA	C_US_JP_04	yes	yes
	Cocodrie	*sativa*	temp. Jap.	USA	C_US_JP_05	yes	
	Cypress	*sativa*	temp. Jap.	USA	C_US_JP_06	yes	
	KayBonnet	*sativa*	temp. Jap.	USA	C_US_JP_08	yes	
	LaGrue	*sativa*	temp. Jap.	USA	C_US_JP_09	yes	
	Adair	*sativa*	temp. Jap.	USA	C_US_JP_10	yes	yes
	Drew	*sativa*	temp. Jap.	USA	C_US_JP_11	yes	
	Bluebonnet 50	*sativa*	temp. Jap.	USA	C_US_JP_12	yes	yes
CIor 9277	Smooth Zenith	*sativa*	temp. Jap.	USA, Texas	C_US_JP_13	yes	yes
64878	Bjanam	*sativa*	temp. Jap.	Bhutan	L_BT_JP_01	yes	yes
64886	Chonam	*sativa*	temp. Jap.	Bhutan	L_BT_JP_02	yes	
64929	Silekachum	*sativa*	temp. Jap.	Bhutan	L_BT_JP_04	yes	
67826	Bhotay Dhan	*sativa*	temp. Jap.	Bhutan	L_BT_JP_05	yes	
67852	Takmaru	*sativa*	temp. Jap.	Bhutan	L_BT_JP_06	yes	yes
PI 564580	Pare Riri	*sativa*	temp. Jap.	Indonesia, Celebes	L_ID_JP_01	yes	yes
PI 419449	Silewah	*sativa*	temp. Jap.	Indonesia, Sumatra	L_ID_JP_02	yes	yes
66647	Si Gepai	*sativa*	temp. Jap.	Indonesia	L_ID_JP_08	yes	yes
	Pare Pulu Lotong	*sativa*	temp. Jap.	Indonesia	L_ID_JP_14	yes	yes
14945	Kolleh	*sativa*	temp. Jap.	Liberia	L_LR_JP_01	yes	yes
14383	Padi Babas	*sativa*	temp. Jap.	Malaysia	L_MY_JP_02	yes	yes
	Alubis	*sativa*	temp. Jap.	Malaysia	L_MY_JP_03	yes	
	Purak Siriba	*sativa*	temp. Jap.	Malaysia	L_MY_JP_07	yes	
	Wangkod	*sativa*	temp. Jap.	Malaysia	L_MY_JP_08	yes	yes
13375	Jumula 2	*sativa*	temp. Jap.	Nepal	L_NP_JP_03	yes	yes
CIor 4602	Kinabugan	*sativa*	temp. Jap.	Philippines	L_PH_JP_01	yes	yes
CIor 12145	Kinabugan selection	*sativa*	temp. Jap.	Philippines, Palawan	L_PH_JP_04	yes	yes
23362	Pinidwa Qan Qipugo Walay	*sativa*	temp. Jap.	Philippines	L_PH_JP_05	yes	yes
	Nep Lao Hoa Binh	*sativa*	temp. Jap.	Vietnam	L_VN_JP_10	yes	yes
9669	Yacca	*sativa*	temp. Jap.	West Africa	L_WR_JP_01	yes	yes
	nipponbare	*sativa*	temp. Jap.		nipponbare	yes	yes

15127	Marupi	*sativa*	temp. Jap.	Ivory Coast(Cote D'Ivoire)	U_CI_JP_01	yes	
56767	Fossa	*sativa*	temp. Jap.	Cote D'Ivoire	U_CI_JP_03	yes	
23319	Hinglu	*sativa*	temp. Jap.	Philippines	U_PH_JP_01	yes	yes
9669	Yacca	*sativa*	temp. Jap.	West Africa	L_WR_JP_01	yes	yes
CIor 5548	Tung Ting Yellow	*sativa*	trop. Jap.	China, Jiangsu	C_CN_JV_02	yes	yes
PI 584570	Arias	*sativa*	trop. Jap.	Indonesia, Java	C_ID_JV_01	yes	yes
PI 597344	Baber	*sativa*	trop. Jap.	India, Kashmir	C_IN_JV_01	yes	
	Silla	*sativa*	trop. Jap.	Italy	C_IT_JV_01	yes	yes
2545		*sativa*	trop. Jap.	Japan	C_JP_JV_01	yes	yes
PI 596815	376	*sativa*	trop. Jap.	Cambodia	C_KH_JV_01	yes	yes
PI 373703	Deng Mak Tek	*sativa*	trop. Jap.	Laos	C_LA_JV_01	yes	yes
	Khao Luang	*sativa*	trop. Jap.	Laos	C_LA_JV_02	yes	
	Azucena	*sativa*	trop. Jap.	Philippines	C_PH_JV_03	yes	yes
38698		*sativa*	trop. Jap.	Pakistan	C_PK_JV_01	yes	
15993	Opa	*sativa*	trop. Jap.	Senegal	C_SN_JV_03	yes	
	RD1	*sativa*	trop. Jap.	Thailand	C_TH_JV_01	yes	
66756		*sativa*	trop. Jap.	Texas	C_US_JV_01	yes	yes
CIor 9672	Earlirose	*sativa*	trop. Jap.	USA, California	C_US_JV_14	yes	yes
25901		*sativa*	trop. Jap.	Bangladesh	L_BD_JV_01	yes	yes
10354	Tung Ching Chuen	*sativa*	trop. Jap.	China	L_CN_JV_04	yes	yes
16428		*sativa*	trop. Jap.		L_ID_JV_01	yes	yes
43372		*sativa*	trop. Jap.	Indonesia (Bali)	L_ID_JV_02	yes	yes
PI 251346	W 129	*sativa*	trop. Jap.	India, Karnataka	L_IN_JV_03	yes	yes
67759	Sathiya	*sativa*	trop. Jap.	India	L_IN_JV_08	yes	
22796		*sativa*	trop. Jap.	Cambodia	L_KH_JV_01	yes	yes
77638	Aeguk	*sativa*	trop. Jap.	Korea	L_KR_JV_01	yes	yes
77673	Udigo	*sativa*	trop. Jap.	Korea	L_KR_JV_02	yes	yes
11624	Khao Hom	*sativa*	trop. Jap.	Laos	L_LA_JV_01	yes	
12922		*sativa*	trop. Jap.		L_LA_JV_02	yes	yes
14371	Padi Siam	*sativa*	trop. Jap.	Malaysia	L_MY_JV_01	yes	yes
8244		*sativa*	trop. Jap.	Philippines	L_PH_JV_02	yes	yes
24225		*sativa*	trop. Jap.		L_TH_JV_01	yes	yes
15046		*sativa*	trop. Jap.	Thailand	L_TH_JV_02	yes	yes
12104		*sativa*	trop. Jap.		L_VN_JV_02	yes	yes
49204		*sativa*	trop. Jap.	Bangladesh	U_BD_JV_N01	yes	yes
17881		*sativa*	trop. Jap.	Indonesia	U_ID_JV_N01	yes	yes
51976		*sativa*	trop. Jap.	India	U_IN_JV_N01	yes	yes
53972		*sativa*	trop. Jap.	India	U_IN_JV_N02	yes	yes
96955		*sativa*	trop. Jap.	Cambodia	U_KH_JV_N01	yes	yes
83932		*sativa*	trop. Jap.	Cambodia	U_KH_JV_N02	yes	yes

89247		*sativa*	trop. Jap.	Laos	U_LA_JV_N01	yes	yes
30177		*sativa*	trop. Jap.	Laos	U_LA_JV_N02	yes	yes
29465		*sativa*	trop. Jap.	Laos	U_LA_JV_N03	yes	yes
29586		*sativa*	trop. Jap.	Laos	U_LA_JV_N04	yes	yes
29546		*sativa*	trop. Jap.	Laos	U_LA_JV_N05	yes	yes
12988		*sativa*	trop. Jap.	Laos	U_LA_JV_N06	yes	yes
95821		*sativa*	trop. Jap.	Myanmar	U_MM_JV_N01	yes	yes
90586		*sativa*	trop. Jap.	Vietnam	U_MM_JV_N02	yes	yes
95860		*sativa*	trop. Jap.	Myanmar	U_MM_JV_N03	yes	yes
95822		*sativa*	trop. Jap.	Myanmar	U_MM_JV_N04	yes	yes
96080		*sativa*	trop. Jap.	Myanmar	U_MM_JV_N05	yes	yes
96023		*sativa*	trop. Jap.	Myanmar	U_MM_JV_N06	yes	yes
95899		*sativa*	trop. Jap.	Myanmar	U_MM_JV_N07	yes	yes
96062		*sativa*	trop. Jap.	Myanmar	U_MM_JV_N08	yes	yes
95858		*sativa*	trop. Jap.	Myanmar	U_MM_JV_N09	yes	yes
64111		*sativa*	trop. Jap.	Nepal	U_NP_JV_N01	yes	yes
15814	Kalor	*sativa*	trop. Jap.	Senegal	U_SN_JV_01	yes	yes
64285		*sativa*	trop. Jap.	Thailand	U_TH_JV_N01	yes	yes
64303		*sativa*	trop. Jap.	Thailand	U_TH_JV_N02	yes	yes
65563		*sativa*	trop. Jap.	Thailand	U_TH_JV_N03	yes	yes
48028		*sativa*	trop. Jap.	Thailand	U_TH_JV_N04	yes	yes
47873		*sativa*	trop. Jap.	Thailand	U_TH_JV_N05	yes	yes
90572		*sativa*	trop. Jap.	Vietnam	U_VN_JV_N01	yes	yes
90619		*sativa*	trop. Jap.	Vietnam	U_VN_JV_N02	yes	yes
25499		*sativa*	trop. Jap.	Indonesia		yes	
12954		*sativa*	trop. Jap.	Laos		yes	
96012		*sativa*	trop. Jap.	Myanmar		yes	
67441		*sativa*	trop. Jap.	Philippines		yes	
51048		*sativa*	trop. Jap.	Sri Lanka		yes	
12491		*sativa*	trop. Jap.	India		yes	
59101		*sativa*	trop. Jap.	Nepal		yes	
52197		*sativa*	trop. Jap.	India		yes	
13149		*sativa*	trop. Jap.	Maylasia		yes	
43328		*sativa*	trop. Jap.	Indonesia		yes	
64062		*sativa*	trop. Jap.	Indonesia		yes	
95771		*sativa*	trop. Jap.	Myanmar		yes	
32145		*sativa*	trop. Jap.	Vietnam		yes	
53533		*sativa*	trop. Jap.	Bangladesh		yes	
52377		*sativa*	trop. Jap.	India		yes	
12442		*sativa*	trop. Jap.	India		yes	

54192		*sativa*	trop. Jap.	Indonesia		yes	
25732		*sativa*	trop. Jap.	Indonesia		yes	
87660		*sativa*	trop. Jap.	Cambodia		yes	
52469		*sativa*	trop. Jap.	India		yes	
47574		*sativa*	trop. Jap.	Maylasia		yes	
25892		*sativa*	trop. Jap.	Bangladesh		yes	
71553		*sativa*	trop. Jap.	Maylasia		yes	
47225		*sativa*	trop. Jap.	Philippines		yes	
76969		*sativa*	trop. Jap.	Indonesia		yes	
14381		*sativa*	trop. Jap.	Maylasia		yes	
44122		*sativa*	trop. Jap.	Maylasia		yes	
59055		*sativa*	trop. Jap.	Nepal		yes	
53042		*sativa*	trop. Jap.	Philippines		yes	
61349		*sativa*	trop. Jap.	Thailand		yes	
49116		*sativa*	trop. Jap.	Bangladesh		yes	
53725		*sativa*	trop. Jap.	India		yes	
27502		*sativa*	trop. Jap.	Indonesia		yes	
77603		*sativa*	trop. Jap.	Indonesia		yes	
14362		*sativa*	trop. Jap.	Maylasia		yes	
61359		*sativa*	trop. Jap.	Thailand		yes	
16974		*sativa*	trop. Jap.	Vietnam		yes	
56082		*sativa*	trop. Jap.	Vietnam		yes	
23318		*sativa*	trop. Jap.	Philippines		yes	
66522		*sativa*	trop. Jap.	Sri Lanka		yes	
81292		*sativa*	trop. Jap.	Cambodia		yes	
40673		*sativa*	trop. Jap.	Thailand		yes	
48824		*sativa*	trop. Jap.	Indonesia		yes	
29539		*sativa*	trop. Jap.	Laos		yes	
12941		*sativa*	trop. Jap.	Laos		yes	
22796		*sativa*	trop. Jap.	Cambodia		yes	
95864		*sativa*	trop. Jap.	Myanmar		yes	
8261		*sativa*	trop. Jap.	Indonesia		yes	
17757*		*sativa*	trop. Jap.	Indonesia		yes	
43675		*sativa*	trop. Jap.	Indonesia (East Java)		yes	
43325		*sativa*	trop. Jap.	Indonesia (West Java)		yes	
12922		*sativa*	trop. Jap.	Laos		yes	
3967		*sativa*	trop. Jap.	Philippines		yes	
3764/6949		*sativa*	trop. Jap.	Philippines (introduced)		yes	
PI 65323	Basmati	*sativa*	aromatic	India, Assam		yes	
PI 393146	Basmati 5854	*sativa*	aromatic	Pakistan		yes	

PI 385421	Basmati	*sativa*	aromatic	Pakistan, Punjab		yes	
PI 402762	Basmati 37	*sativa*	aromatic	India		yes	
PI 233103	Aus Basmati	*sativa*	aromatic	India, Orissa		yes	
PI 385443	Basmati	*sativa*	aromatic	Pakistan, Punjab		yes	
CIor 12524	Basmati	*sativa*	aromatic	India, Punjab	Basmati_3	yes	yes
PI 159992	Basmati 3	*sativa*	aromatic	India, Delhi	Basmati_2	yes	yes
PI 173923	Basmati	*sativa*	aromatic	India, Uttar Pradesh	Basmati_4	yes	yes
PI 385403	Basmati	*sativa*	aromatic	Pakistan, Punjab	Basmati_7	yes	yes
PI 412774	Basmati 5875	*sativa*	aromatic	Pakistan, N-W Front	Basmati_6	yes	yes
PI 430924	Basmati	*sativa*	aromatic	Pakistan, Sind	Basmati_8	yes	yes
PI 584556	Basmati I	*sativa*	aromatic	Pakistan	Basmati_5	yes	yes
PI 65323		*sativa*	aromatic		Basmati_1	yes	yes
103404		*rufipogon*	wild rice	Bangladesh	W_BD_01	yes	yes
105898		*rufipogon*	wild rice	Bangladesh	W_BD_04		yes
100916		*rufipogon*	wild rice	China	W_CN_01	yes	
104624		*rufipogon*	wild rice	China	W_CN_02	yes	yes
81976		*rufipogon*	wild rice	Indonesia	W_ID_02	yes	
105567		*rufipogon*	wild rice	Indonesia	W_ID_04	yes	
105956		*rufipogon*	wild rice	Indonesia	W_ID_05	yes	
106453		*rufipogon*	wild rice	Indonesia	W_ID_06		yes
NSGC 5936		*rufipogon*	wild rice	India	W_IN_02	yes	yes
NSGC 5940		*rufipogon*	wild rice	India	W_IN_05	yes	yes
105471		*rufipogon*	wild rice	India	W_IN_27	yes	yes
105711		*rufipogon*	wild rice	India	W_IN_29	yes	
106078		*rufipogon*	wild rice	India	W_IN_31	yes	yes
106086		*rufipogon*	wild rice	India	W_IN_32	yes	yes
106103		*rufipogon*	wild rice	India	W_IN_33	yes	yes
106122		*rufipogon*	wild rice	India	W_IN_35	yes	yes
105720		*rufipogon*	wild rice	Cambodia	W_KH_01	yes	
106325		*rufipogon*	wild rice	Cambodia	W_KH_06	yes	
106321		*rufipogon*	wild rice	Cambodia	W_KH_11		yes
106163		*rufipogon*	wild rice	Laos	W_LA_06	yes	yes
104599		*rufipogon*	wild rice	Sri Lanka	W_LK_02	yes	
81990		*rufipogon*	wild rice	Myanmar	W_MM_07	yes	yes
100923		*rufipogon*	wild rice	Myanmar	W_MM_08	yes	
106346		*rufipogon*	wild rice	Myanmar	W_MM_09	yes	yes
100189		*rufipogon*	wild rice	Malaysia	W_MY_01	yes	yes
106036		*rufipogon*	wild rice	Malaysia	W_MY_03	yes	yes
81994		*rufipogon*	wild rice	Papau New Guinea	W_PG_01		yes
106262		*rufipogon*	wild rice	Papau New Guinea	W_PG_04		yes

106523		*rufipogon*	wild rice	Papau New Guinea	W_PG_08		
105568		*rufipogon*	wild rice	Philippines	W_PH_02	yes	
82040		*rufipogon*	wild rice	Thailand	W_TH_03	yes	
104714		*rufipogon*	wild rice	Thailand	W_TH_06	yes	
104815		*rufipogon*	wild rice	Thailand	W_TH_13	yes	yes
104618		*rufipogon*	wild rice	Thailand	W_TH_15	yes	yes
104833		*rufipogon*	wild rice	Thailand	W_TH_17	yes	
104857		*rufipogon*	wild rice	Thailand	W_TH_18	yes	yes
104871		*rufipogon*	wild rice	Thailand	W_TH_21	yes	yes
105855		*rufipogon*	wild rice	Thailand	W_TH_29	yes	yes
100904		*rufipogon*	wild rice	Thailand	W_TH_30	yes	yes
100588		*rufipogon*	wild rice	Taiwan	W_TW_02	yes	yes
106166		*rufipogon*	wild rice	Vietnam	W_VN_13	yes	
106168		*rufipogon*	wild rice	Vietnam	W_VN_14	yes	yes
104625		*rufipogon*	wild rice	China		yes	
104629		*rufipogon*	wild rice	China		yes	

Field collections:

Species	Race[i]	origin	DNA Label	phenotype	genotype
rufipogon	wild rice	Guangdong, China	GD_E1_05		yes
rufipogon	wild rice	Guangdong, China	GD_E1_12		yes
rufipogon	wild rice	Guangdong, China	GD_E2_16		yes
rufipogon	wild rice	Guangdong, China	GD_E2_12		yes
rufipogon	wild rice	Guangdong, China	GD_E3_04		yes
rufipogon	wild rice	Guangdong, China	GD_E3_07		yes
rufipogon	wild rice	Guangdong, China	GD_E4_09		yes
rufipogon	wild rice	Guangdong, China	GD_E4_16		yes
rufipogon	wild rice	Guangdong, China	GD_GS_01		yes
rufipogon	wild rice	Guangdong, China	GD_W1_12		yes
rufipogon	wild rice	Guangdong, China	GD_W1_17		yes
rufipogon	wild rice	Guangdong, China	GD_W2_06		yes
rufipogon	wild rice	Guangdong, China	GD_W2_17		yes
rufipogon	wild rice	Guangdong, China	GD_W3_06		yes
rufipogon	wild rice	Guangdong, China	GD_W3_13		yes
rufipogon	wild rice	Guangxi, China	GX_1_08		yes
rufipogon	wild rice	Guangxi, China	GX_GS_01		yes
rufipogon	wild rice	Hainan, China	HA_N_11		yes
rufipogon	wild rice	Hainan, China	HA_N_16		yes
rufipogon	wild rice	Hainan, China	HA_S_05		yes
rufipogon	wild rice	Hainan, China	HA_S_12		yes
rufipogon	wild rice	Hainan, China	HA_S_19		yes
rufipogon	wild rice	Hainan, China	HA_S_29		yes
rufipogon	wild rice	Hunan, China	HN_1_02		yes
rufipogon	wild rice	Jianxi, China	JX_E_13		yes
rufipogon	wild rice	Jianxi, China	JX_E_18		yes
rufipogon	wild rice	Jianxi, China	JX_E_25		yes
rufipogon	wild rice	Jianxi, China	JX_E_28		yes

[a] The identification number of the samples from IRRI; blank means no IRRI number is available.
[b] common name of the cultivated rice variety
[c] sativa is *O. sativa*. rufipogon is *O. rufipogon*.
[d] the racial designation of cultivated rice based on IRRI documentation and phenol reaction (tested by Jason Londo in Schaal lab)
[e] Country and region from which the original germplasm was donated or collected
[f] the label that I used in the study
[g] The samples that had phenotypic data is shown as yes. Otherwise, it is blank.
[h] The samples that were sequenced is shown as yes. Otherwise, it is blank.
[i] The DNA samples collected from field.

Table 2.2), Summary of number of samples

rice group	No. of accessions for phenotypic data	No. of accessions for genotypic data	No. of accessions for genotype-phenotype association
O. rufipogon	37	57	22
Indica	93	50	46
tropical japonica	114	56	54
temperate japonica	51	29	26
aromatic	14	8	8
aus	16	8	8

Table 2.3), Population statistics of *SsIIa* exon 8 in *O. rufipogon* and rice variety groups

statistics	*O. rufipogon*	*aus*	*indica*	*tropical japonica*	*temperate japonica*	*aromatic*
sample size	57	8	50	56	29	8
nonsynonymous	5	1	3	3	3	1
synonymous	5	0	3	2	2	1
silent	11	0	8	5	5	3
θ_π	0.00508	0	0.00421	0.00173	0.00239	0.00262
θ_w	0.00486	0	0.00362	0.00222	0.00261	0.00236
Tajima's D	-0.64032	-1.05482	0.61507	-0.39692	-0.20385	0.48523
Fay and Wu's H	1.40351	0.21429	-1.11373	-9.42828**	-4.41133	1.14286

**P<0.01

Table 2.4), Association between nonsysnonymous mutations and starch alkali spreading score in rice by general linear model

site	indica	*tropical japonica*	*tempereate japonica*	*O. rufipogon*
1	0.0348*	0.1291	0.0029*	0.3523
2	NA	0.2902	0.0097*	NA
3	1.10E-11**	4.96E-12**	7.12E-04**	NA

*the gene is statistically significant based on 5% significant level.

**the gene is significant even after Bonferroni correction.

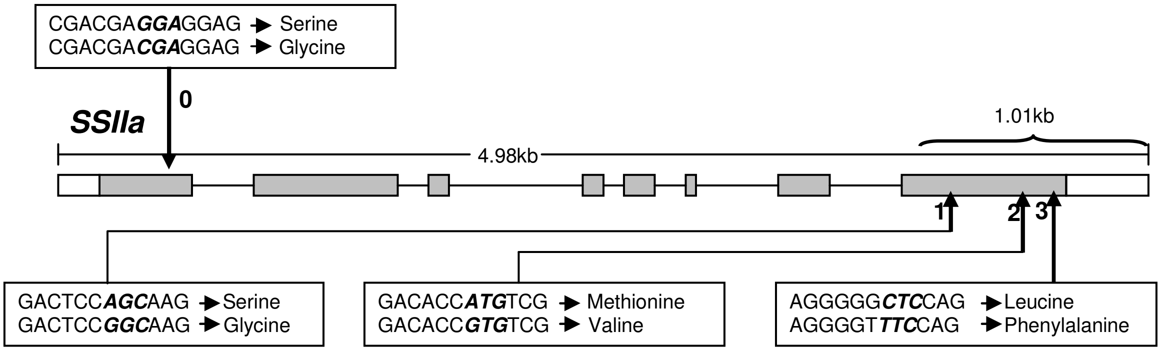
Table 2.5), Association between nonsysnonymous mutations and starch alkali spreading score in rice by nested clade analysis

comparison	*indica*	*tropical japonica*	*temperate japonica*	*O. rufipogon*
H1 vs H2	1.000000	1.000000	0.430336	0.153407
H2 vs H3	0.001027**	0.000011**	0.020414*	NA
H2 vs H4	NA	0.051782	0.102452	NA

*the gene is statistically significant based on 5% significant level.

**the gene is significant even after Bonferroni correction.

Figure 2.1), Schematic representation of gene *SsIIa* which showing the location of nonsynonymous mutations observed in cultivated rice. The box represents the exon region; The shaded ones are the translated region; The nonshased ones are the transcribed but not translated region; The line connecting the box is the intron region. The size of gene and the sequenced region is indicated above the boxes. The number is the name of nonsynonymous mutations in this study. The arrow is the location of the mutations. The bold italic sequence code is the code with nonsynonymous mutation.



The genotype and phenotype relationship in *E. coli.*

	genotype			***SsIIa* enzyme activity**
SNP0	SNP1	SNP2	SNP3	
G	G	G	C	High
G	G	A	C	Low
G	G	G	T	Medium
G	G	A	T	Low
C	G	G	C	High
C	G	A	C	Low
C	G	G	T	Medium
C	G	A	T	Low
G	A	G	C	High
G	A	A	C	Low
G	A	G	T	Low
G	A	A	T	Low
C	A	G	C	High
C	A	A	C	Low
C	A	G	T	Low
C	A	A	T	Low

Figure 2.2), Correction between starch alkali spreading score and enzyme activity of gene *SsIIa*. The solid line is the linear correlation between starch alkali spreading score and enzyme activity of gene *SsIIa*; The dashed line is the 95% confidential interval.

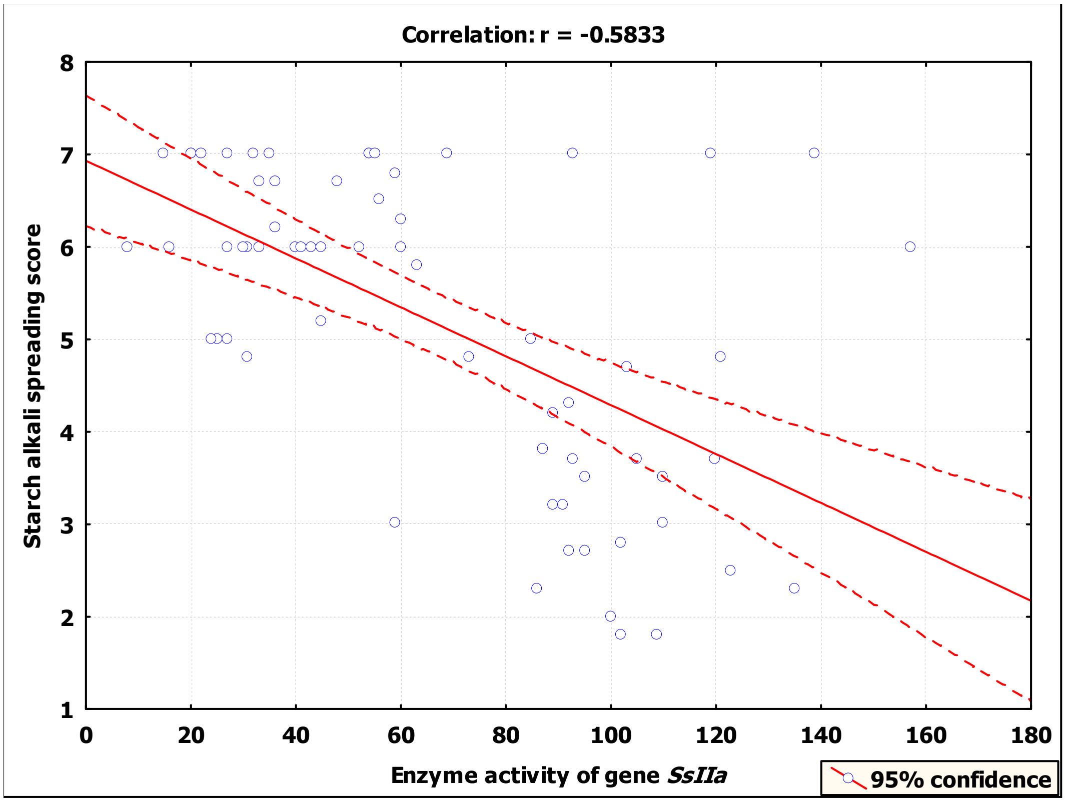
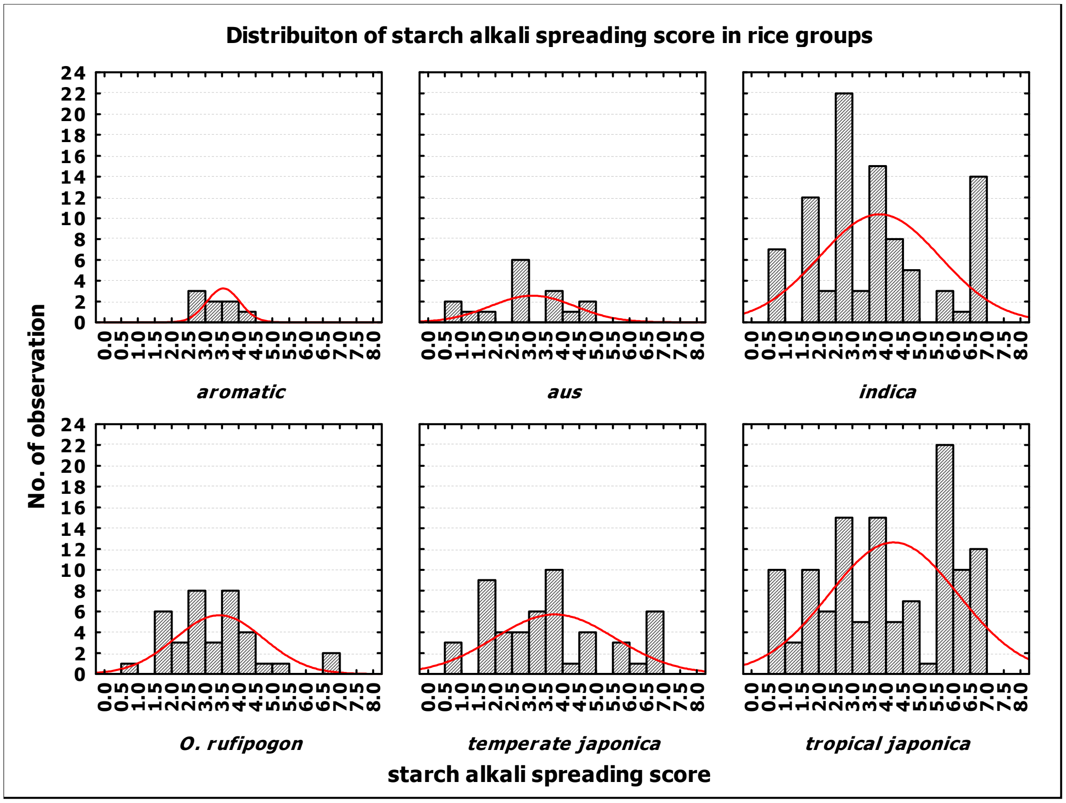
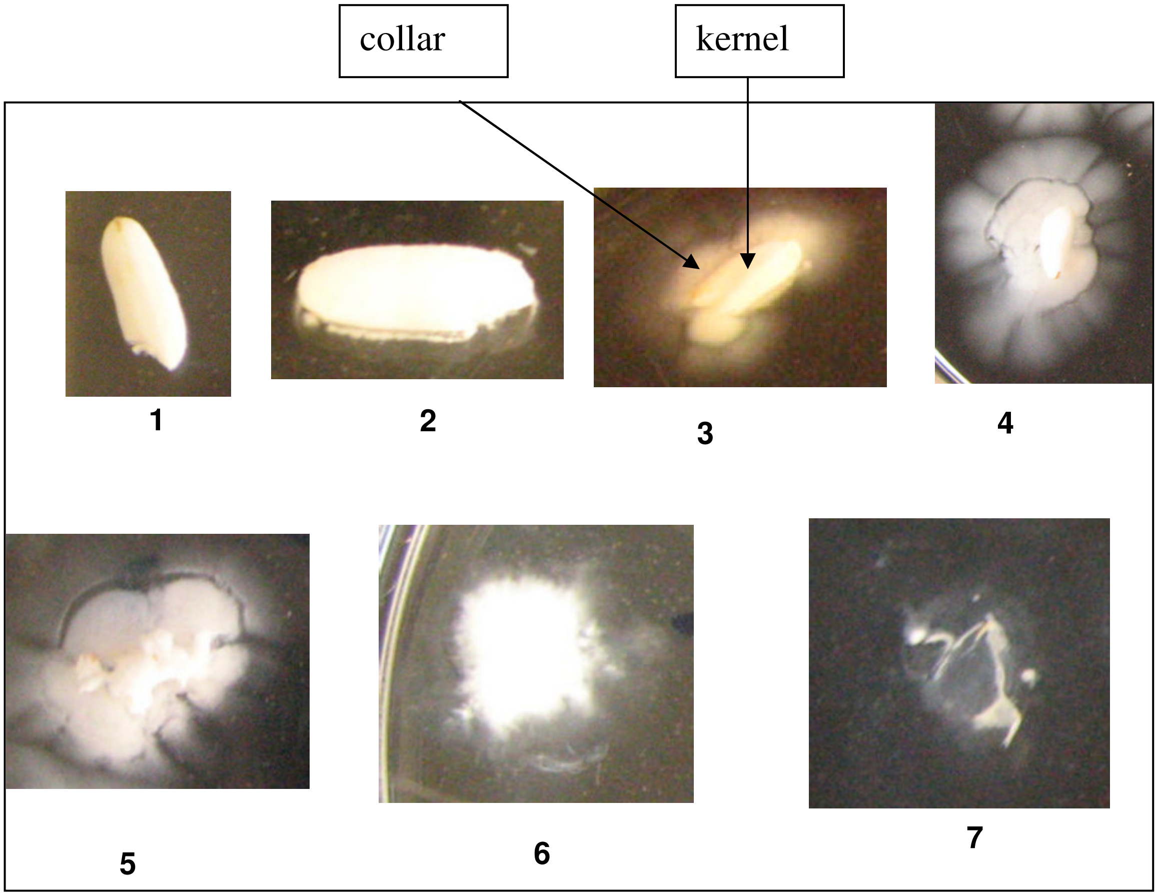

Correlation: r = -0.5833
Starch alkali spreading score
8
7
6
5
4
3
2
1
0
20
40
60
80
100
120
140
160
180
Enzyme activity of gene *SsIIa*
95% confidence

Figure 2.3), Distribution of SASS and frequency distribution of SASS groups in rice groups. The red line is the fitting of normal distribution.



SASS	*aus*	*indica*	*tropical japonica*	*temperate japonica*	*aromatic*	*O. rufipogon*
Low	25.00%	20.43%	19.30%	23.53%	7.14%	18.92%
Medium	75.00%	60.22%	42.11%	56.86%	92.86%	75.68%
High	0.00%	19.35%	38.60%	19.61%	0.00%	5.41%

Figure 2.4, The phenotypes for starch quality in alkali. The number below the picture is the starch alkali spreading score for the phenotype shown in the picture.



Alkali spreading scores (numerical scales 1–7)

Scores	Spreading behavior
1	Kernel not affected
2	Kernel swollen
3	Kernel swollen; collar complete or narrow
4	Kernel swollen; collar complete and wide
5	Kernel split or segregated; collar complete and wide
7	Kernel dispersed; merging with collar
8	Kernel completely dispersed and intermingled

Figure 2.5), distribution of variance of starch alkali spreading score

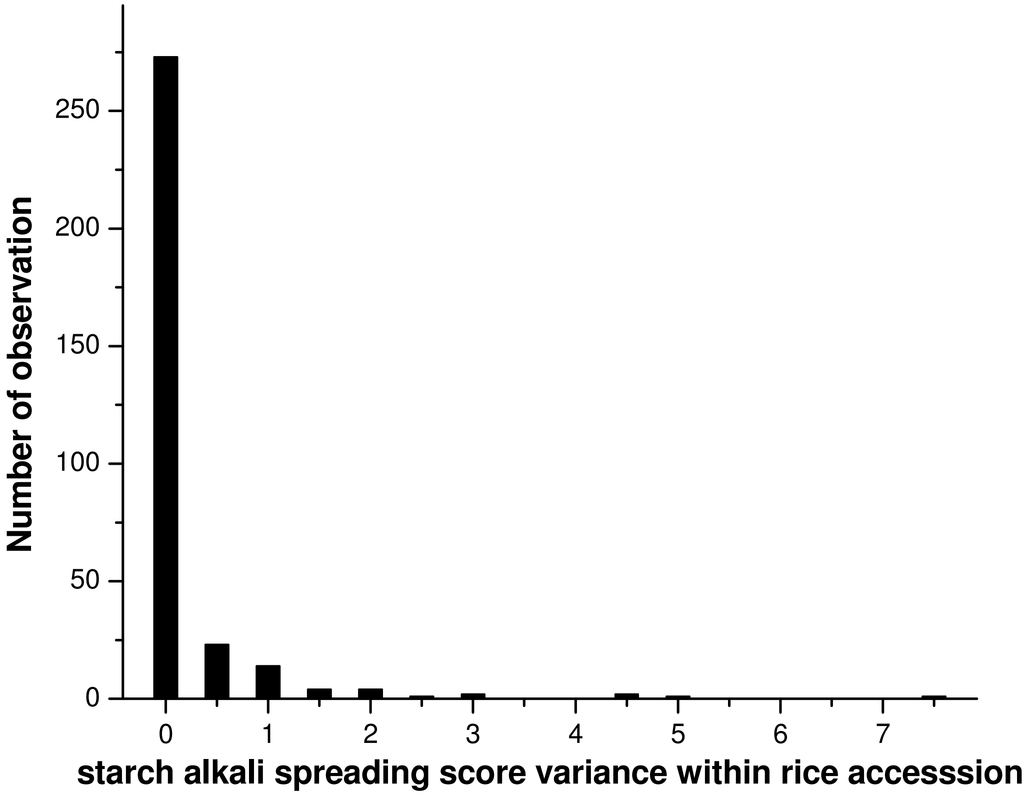

Number of observation
0
50
100
150
200
250
0
1
2
3
4
5
6
7
starch alkali spreading score variance within rice accesssion

Figure 2.6), The haplotype network of *SsIIa* exon8 region. *aus* (□), *indica* (🟦), *tropical japonica* (🟩), *temperate japonica* (🟪), *aromatic* (▨), *O. barthii* (⬛), *O. rufipogon* (🟨). The circles represent the nodes. The size the circle is proportional to the number of individuals. The line connecting the nodes is roughly proportional to the number of mutations. The short line on the line which connects node represents the extinct node. The number on the line is the name of nonsynonymous mutations in the study. The name of the haplotypes was shown as A to P.

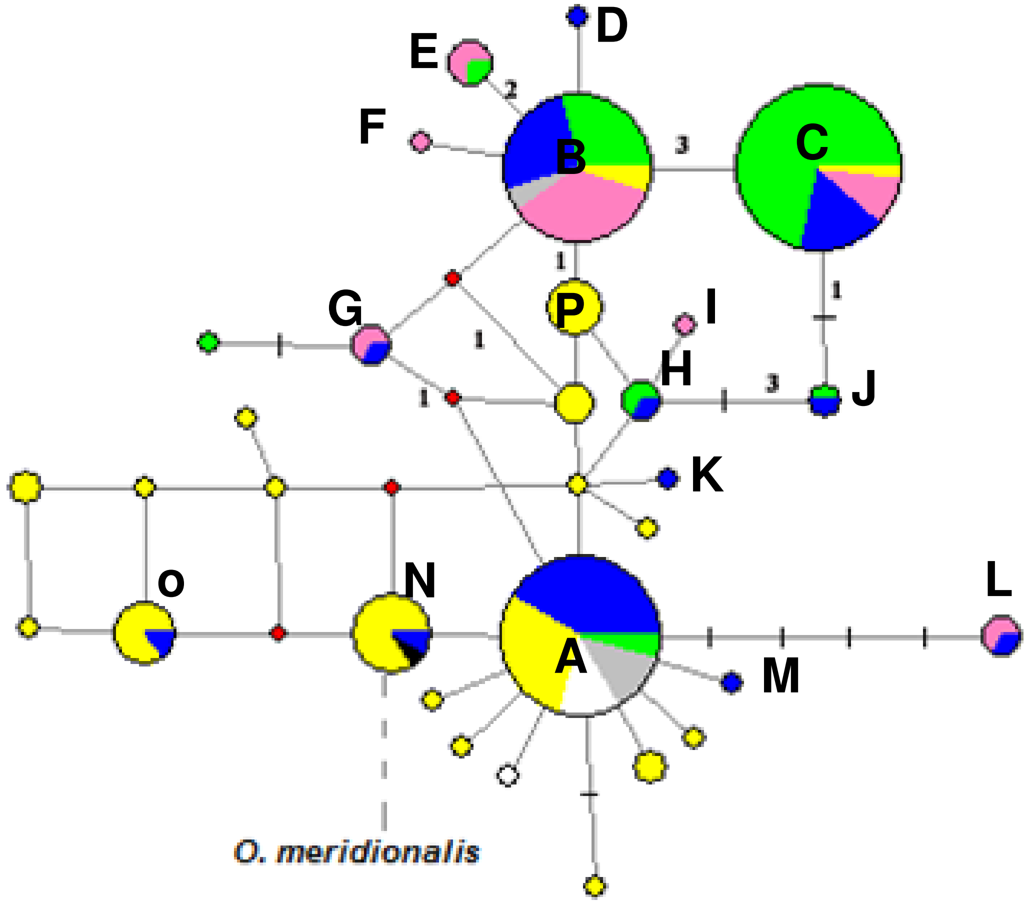
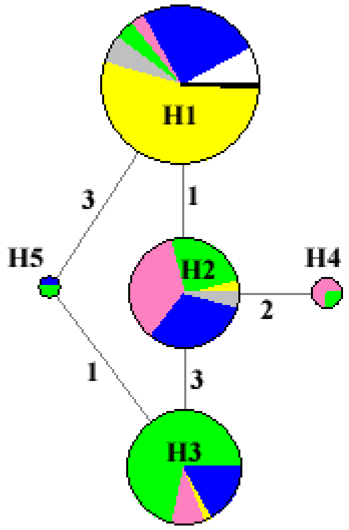

D
E
2
F
B
3
C
1
G
1
P
I
1
H
3
J
1
K
o
N
L
A
M
O. meridionalis

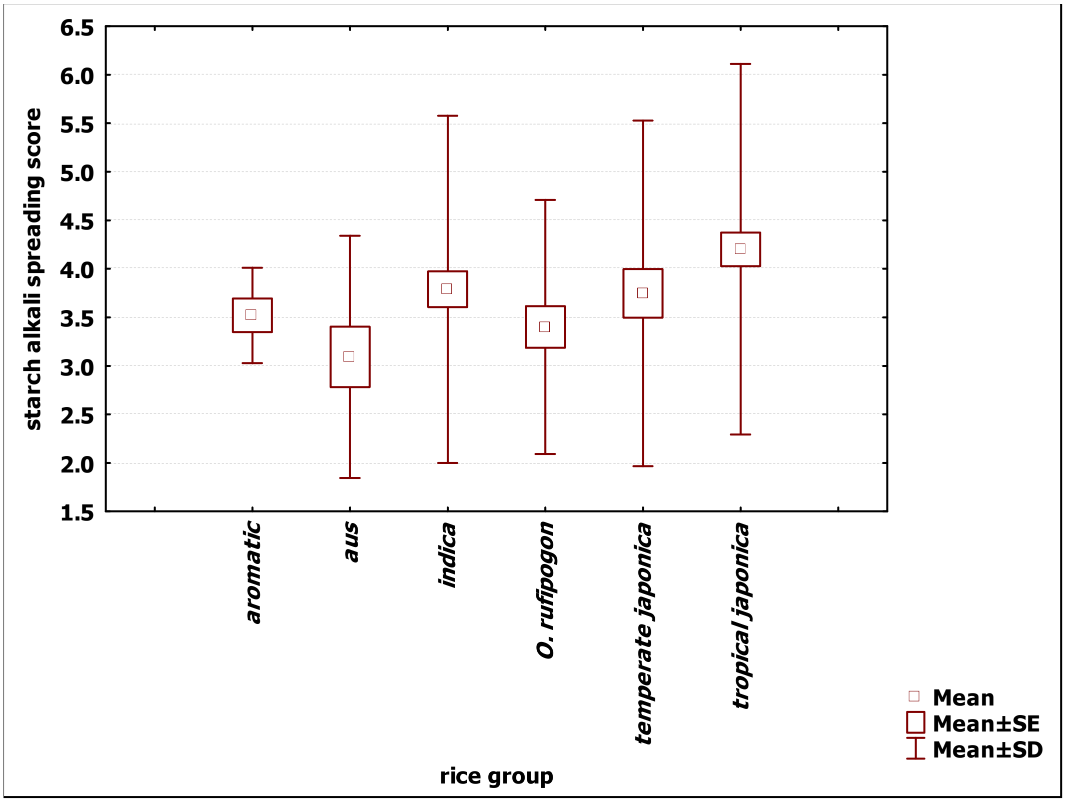
Figure 2.7), The haplotype network of the nonsynonymous muations in *SsIIa* exon8 region. *aus* (□), *indica* (🟦), *tropical japonica* (🟩), *temperate japonica* (🟪), *aromatic* (▨), *O. barthii* (⬛), *O. rufipogon* (🟨). The circles represent the nodes. The size the circle is proportional to the number of individuals. The line connecting the nodes is roughly proportional to the number of mutations. The short line on the line which connects node represents the extinct node. The number on the line is the name of the mutation in the study. H1 to H5 is the name of the haplotypes.



Haplotype frequency in *O. rufipogon* and cultivated rice variety groups

haplotype	*aus*	*indica*	*tropical japonica*	*temperate japonica*	*aromatic*	*O. rufipogon*
H1	100.0%	50.0%	7.1%	10.7%	75.0%	94.7%
H2	0.0%	30.0%	21.4%	60.7%	25.0%	3.5%
H3	0.0%	18.0%	67.9%	17.9%	0.0%	1.8%
H4	0.0%	0.0%	1.8%	10.7%	0.0%	0.0%
H5	0.0%	2.0%	1.8%	0.0%	0.0%	0.0%

Figure 2.8), The mean value of starch alkali spreading score in rice groups and P value of student t test between rice groups. SE is standard error; SD is standard deviation. The red P value is lower than 0.05.



P value of student t test between rice groups

	aus	*indica*	*tropical japonica*	*temperate japonica*	*aromatic*
indica	0.138529				
tropical japonica	0.021431	0.077491			
temperate japonica	0.177222	0.891926	0.113499		
aromatic	0.550662	0.351509	0.081608	0.403848	
O. rufipogon	0.429825	0.232443	0.013466	0.318538	0.858841

Figure 2.9), Distribution of SASS in haplotype groups

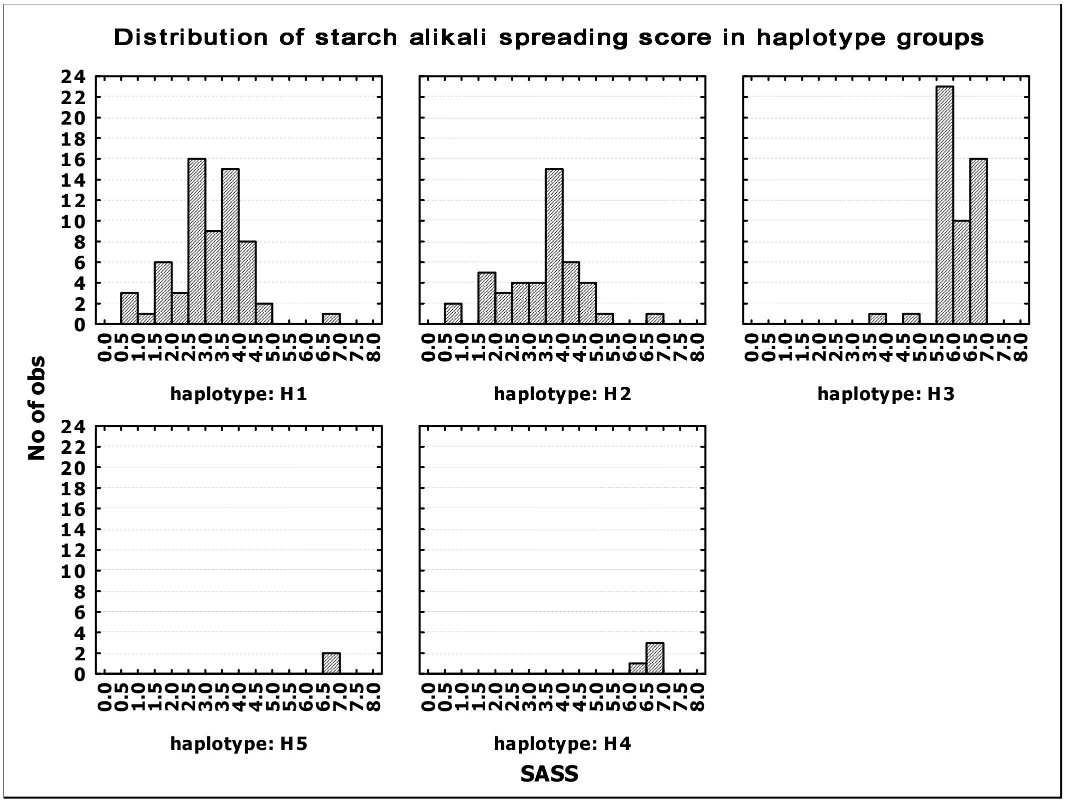
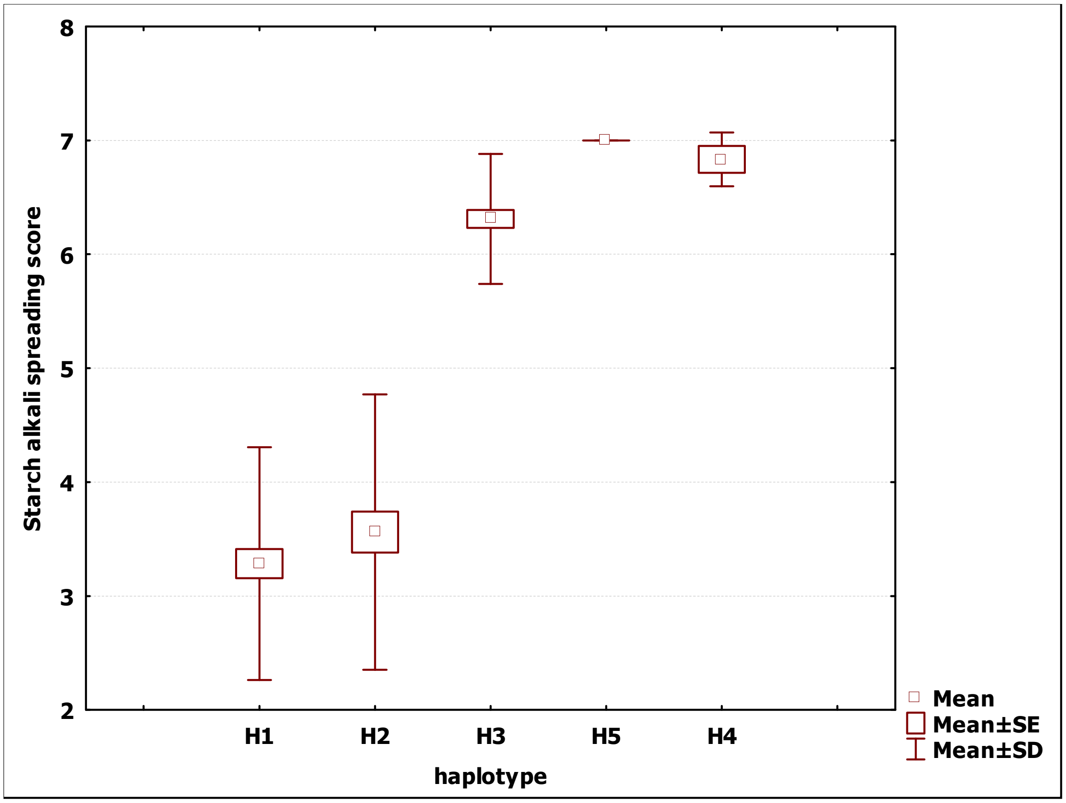

Distribution of starch alikali spreading score in haplotype groups
No of obs
0
2
4
6
8
10
12
14
16
18
20
22
24
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
haplotype: H1
haplotype: H2
haplotype: H3
haplotype: H5
haplotype: H4
SASS

Figure 2.10), The mean value of starch alkali spreading score of each genotype and P value of comparisons between haplotype groups. SE is standard error; SD is standard deviation. The red P value is lower than 0.05.



P value of pairwise comparisons among haplotype groups

	H1	H2	H3	H4
H2	0.1982			
H3	0.0001	0.0001		
H4	0.0001	0.0001	0.0768	
H5	0.0001	0.0002	0.0965	0.397

Conclusion of the Dissertation

Great morphological and functional divergence exists between species, much of which is thought to be adaptive. The study of population genetics of functional genes is essential to understand the origin and the evolution of adaptive traits. Domesticated species have been considered as excellent model systems for understanding the evolution of adaptive traits. Moreover, there is great interest in identifying the genes involved in the evolution of agricultural traits because of the potential agricultural benefits their manipulation could bring.

One approach of studying the evolution of agricultural traits during crop domestication is through their underlying functional genes. Rice (*Oryza sativa*) endosperm starch quality is one of the important agricultural traits, and the key genes (*shrunken2, Sh2; brittle2, Bt2; waxy, Wx; starch synthase IIa, SsIIa; Starch branching enzyme IIb, SbeIIb;* and *Isoamylase1, Iso1*) involved in rice endosperm starch synthesis pathway have been identified. However, the evolution of these key genes and their contributions to starch quality evolution before and after rice domestication are still unknown. Chapter One of this dissertation attempts to fill in this gap by investigating the evolutionary forces that have influenced the evolution of endosperm key starch genes in cultivated rice, *Oryza sativa* and its wild ancestor, *Oryza rufipogon*.

In Chapter One, the level and pattern of diversity for six starch genes are examined in the five major rice variety groups (*aus*, *indica*, *tropical japonica*, *temperate japonica* and *aromatic*) and their wild ancestor, *O. rufipogon*. Results indicate significantly higher diversity of the starch genes in *O. rufipogon* than those in any other rice variety groups, which might be the indication of bottleneck events during rice domestication. The level of diversity of starch genes is slightly higher in modern cultivars than traditional landraces in *indica, tropical* and *temperate japonica*, which might be the result of the genetic introgression during modern improvement.

This dissertation is also one of the few studies which study the evolution of functional genes in the context of metabolic pathways. No general association between nucleotide variation at a gene and position of that gene in the rice endosperm starch synthesis pathway was found in *O. rufipogon*. However, upstream genes *Sh2* do show low diversity and significant deviation from a standard neutral model by Tajima's D value, which might suggest strong selective constraints at this gene before domestication. Among these six starch genes, *Wx* has significantly higher level of diversity than those of the available genome-wide sequence data in *O. rufipogon.* This elevated diversity might be a result of high diversity of the transposable elements in *Wx*.

Detecting selection is a challenge in species with complex demographic histories, such as domesticated species. This study is one of the few studies which detect selection by comparing the studied genes with both genome-wide sequence data and a standard neutral model. No evidence of selection is found at any of the six starch genes in *O. rufipogon* and at four starch genes (*Sh2, Bt2, SsIIa, Iso1*) in *O. sativa*. Evidence of directional selection is detected at *Wx* in *tropical japonica*, *temperate japonica,* and at *Wx* and *SbeIIb* in *aromatic,* which suggests the contribution of these starch genes to starch quality evolution during the domestication for *aromatic* and *japonicas* rice. Although the same gene (*Wx*) was selected in *aromatic*, *tropical* and *temperate japonica* rice, it is unlikely to be a single selection event (See page 52 for reasons). This study not only demonstrates the role of selection in the evolution of starch genes during rice domestication, but also suggests the complex history of rice domestication.

This study, for the first time, uses the genome-wide sequence data as control of neutral reference for detecting selection in rice. The genome-wide sequences used here are 111 sequenced tagged sites, distributed across the whole rice genome. It was believed that genome-wide sequence data would mainly reflect the demographic history of a species. However, in order to fully understand the diversity pattern of the starch genes in rice, a genome-wide sequence data with increasing coverage across the rice genome is required. A larger sequence data set will allow the establishment of the most likely demographic scenarios for rice, which could help the interpretation of nucleotide variation of starch genes and other functional genes in rice to the full extent.

This dissertation has indicated the contribution of *Wx* and *SbeIIb* to the evolution of starch quality evolution during rice domestication. However, in order to fully understand starch quality evolution during rice domestication, further research is required. For example, over 20 genes involved in the starch synthesis pathway have been identified. Their effects on the starch quality evolution during rice domestication need to be determined. In addition, the specific targets of selection at *SbeIIb* in *aromatic* and at *Wx* in *tropical japonica* and *aromatic rice* remain unknown, which also requires further investigation.

Chapter One exhibits the pattern of variation of six key starch genes. Starch quality shows not only difference between *O. sativa* and its wild ancestor, *O. rufipogon,* but also variation within *O. sativa* and *O. rufipogon*. However, the relationship between genetic diversity of starch genes and starch quality diversity is unknown. Chapter Two is a genotype-phenotype association study, which seeks to determine the genetic basis of starch quality variation in rice. Candidate region identified by previous studies, *SsIIa exon 8* was sequenced in *O. rufipogon* and five variety groups of *O. sativa*. Starch alkali spreading score (SASS) was used to quantify the rice endosperm starch quality. Genotype-phenotype association analyses were performed in each of the five rice variety groups and *O. rufipogon* to reduce the effect of population structure on the analyses. Both general linear model and nested clade analyses of genotype-phenotype association show consistent results. Mutation 3 (see Fig 2.1) contributes to SASS diversity in rice and mutation 1 does not, which is statistically supported. Mutation 2 alters SASS from low to high in our samples. However, this result is tentative since statistical support is lacking due to limited amount of accessions which contain mutation 2. While there is a significant association between genotype and phenotype, the relationship is not absolute. More samples and more candidate genes are required in order to understand the genetic basis of SASS diversity and hence starch quality diversity in rice.

In addition, Chapter Two also demonstrates the evolutionary history of those candidate mutations in rice, and surveys starch quality difference among rice groups. The haplotype network of *SsIIa* exon 8 suggests that mutation 1 is evolutionary older than mutations 2 and 3. SASS comparison among rice groups showed that *tropical japonica* is different from all other rice groups with more high SASS category individuals. However, no statistical difference of SASS among rice groups is found. Starch quality can be measured by more advanced and accurate quantitative methods. This study suggests that an advanced starch quality phenotyping method and more samples are required to understand the starch quality difference among *O. rufipogon* and the other five rice variety groups.

The work presented in this dissertation uses population genetics techniques to understand the genetic basis of phenotypic divergence between species, and phenotypic variation within species in an important crop plant model system. Very few population studies of functional genes in the context of a biosynthesis pathway have been previously undertaken. This study examines the relationship between the evolution and the position of a gene in a biosynthesis pathway. The challenges of detecting selection in species with complex demographic histories are addressed here. This work demonstrates the importance of knowing population structure of a species for the sampling and analyses of genotype-phenotype association within a species.

Different populations of a species might have different frequencies of alleles. Lacking the samples from one population would results in the lack of statistical support of association analysis for some alleles, which only exist in high frequency in that population. This work emphasizes the importance of knowing the demographic history and population structure of a species for the studying of functional gene evolution, and for the study of the relationship between genetic and phenotypic variation within a species. Many challenges are foreseeable for the study of evolution of adaptive traits. But with the development of high-throughput sequencing, the population genetics techniques promise to bring much more understanding of adaptation and the great life variation on Earth.