

Molecular Evolution of the Endosperm Starch Synthesis Pathway Genes in Rice (*Oryza sativa* L.) and Its Wild Ancestor, *O. rufipogon* L.

Guoqin Yu,*†¹ Kenneth M. Olsen,¹ and Barbara A. Schaal¹

¹Department of Biology, Washington University in St. Louis, Missouri

†Present address: Institute for Biotechnology and Research, University of Maryland, College Park, Maryland.

*Corresponding author: E-mail: guoqinyu@umd.edu.

Associate editor: John McDonald

Abstract

The evolution of metabolic pathways is a fundamental but poorly understood aspect of evolutionary change. One approach for understanding the complexity of pathway evolution is to examine the molecular evolution of genes that together comprise an integrated metabolic pathway. The rice endosperm starch biosynthetic pathway is one of the most thoroughly characterized metabolic pathways in plants, and starch is a trait that has evolved in response to strong selection during rice domestication. In this study, we have examined six key genes (*AGPL2*, *AGPS2b*, *SSIIa*, *SBEIIb*, *GBSSI*, *ISA1*) in the rice endosperm starch biosynthesis pathway to investigate the evolution of these genes before and after rice domestication. Genome-wide sequence tagged sites data were used as a neutral reference to overcome the problems of detecting selection in species with complex demographic histories such as rice. Five variety groups of *Oryza sativa* (*aus*, *indica*, *tropical japonica*, *temperate japonica*, *aromatic*) and its wild ancestor (*O. rufipogon*) were sampled. Our results showed evidence of purifying selection at *AGPL2* in *O. rufipogon* and strong evidence of positive selection at *GBSSI* in *temperate japonica* and *tropical japonica* varieties and at *GBSSI* and *SBEIIb* in *aromatic* varieties. All the other genes showed a pattern consistent with neutral evolution in both cultivated rice and its wild ancestor. These results indicate the important role of positive selection in the evolution of starch genes during rice domestication. We discuss the role of *SBEIIb* and *GBSSI* in the evolution of starch quality during rice domestication and the power and limitation of detecting selection using genome-wide data as a neutral reference.

Key words: positive selection, starch synthesis pathway, rice, domestication.

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 (Wright and Gaut 2005; Ramos-Onsins et al. 2008). However, most of this work focuses on individual genes or multiple neutral markers. There is relatively less understanding about the evolution of 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 determine how evolutionary forces act on multiple interacting genes that are components of molecular pathways or networks (Fraser et al. 2002). Moreover, 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 the understanding of the evolution of underlying metabolic pathways. 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 function in specific biochemical pathways and now allow for evolutionary study of functional genes in the context of an entire pathway.

Rice (*Oryza sativa*) is one of the world's most important food crops and feeds about half of the world's population (Khush 1997). 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 rice domestication, and it is likely to have occurred via selection on starch synthesis pathway genes (Whitt et al. 2002).

Asian rice is one of the oldest domesticated species, domesticated in Asia at least 11,000 years ago (Khush 1997). Previous studies have indicated that *O. sativa* is derived from its wild ancestor, *O. rufipogon* (Khush 1997). Recent studies have indicated that there were possibly at least two domestication centers: one in South China for *japonica* rice and 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).

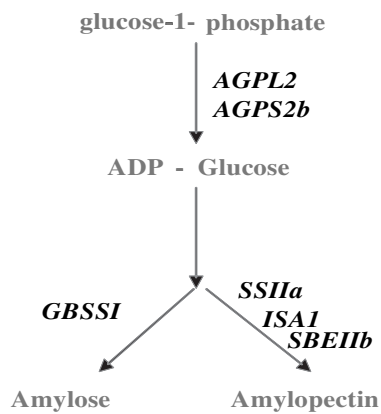


Fig. 1 Simplified starch synthesis pathway with all the studied starch genes.

Rice, *O. sativa*, is highly variable in phenotype with an estimated 120,000 varieties (Khush 1997). Most varieties of rice are placed into two subspecies or races, *O. sativa* ssp. *indica* and *O. 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 endosperm starch granules in alkali solution (Khush 1997). A recent study with 169 nuclear microsatellite (simple sequence repeat) markers and two chloroplast loci identified five rice genetically distinct variety groups: *aus*, *indica*, *tropical japonica*, *temperate japonica*, and *aromatic* (Garris et al. 2005). Among these five variety groups, close evolutionary relationships were found between *aus* and *indica* rice and between *tropical japonica*, *temperate japonica*, and *aromatic* rice (Garris et al. 2005).

Starch is the major component of yield in the world's most important cereal crop plants (James et al. 2003). It is composed of amylose and amylopectin (Buléon et al. 1998). Amylose is a linear molecule of (1→4) linked α -D-glucopyranosyl units (Buléon et al. 1998). 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). Starch quality in crops varies based on either different ratios of amylose to amylopectin content or the structure of amylopectin (Hirano et al. 1998; Umemoto et al. 2002).

The starch synthesis pathway is an ideal system for examining the evolution of biochemical pathways. It is one of the best-characterized metabolic pathways in plants (Myers et al. 2000). Over 20 genes involved in the starch synthesis pathway have been identified so far (Myers et al. 2000) with six playing a major role in rice endosperm starch synthesis: ADP-glucose pyrophosphorylase large subunit 2 (AGPL2), ADP-glucose pyrophosphorylase small subunit 2b (AGPS2b), granule-bound starch synthase I (GBSSI), starch synthase IIa (SSIIa), starch branching enzyme IIb (SBEIIb), and isoamylase1 (ISA1) (Pan and Nelson 1984; Sano 1984; Nakamura 1996; James et al. 2003; Smidansky et al. 2003; Tanaka et al. 2004; Nakamura

et al. 2005; Lee et al. 2007). The position of these starch genes in the starch synthesis pathway is shown in figure 1. All of these genes except AGPS2b, GBSSI, and ISA1 are exclusively expressed in rice endosperm (James et al. 2003).

Detection of selection at the molecular level has proved to be difficult in many species, mainly due to their complex demographic histories (Wright and Gaut 2005). Most available methods detect the signal of selection by comparing observed data with those expected by a standard neutral model (Tajima 1989; Fu and Li 1993; Fay and Wu 2000; Sabeti et al. 2002; Kim and Nielsen 2004; Wong and Nielsen 2004; Li and Stephan 2005). A standard neutral model usually assumes a simple demographic history (e.g., constant population size), whereas most species, especially domesticated species, have experienced much more complicated demographic processes (e.g., bottlenecks and/or population expansions). Therefore, instead of the action of selection, deviations from the standard neutral model at particular loci might reflect the specific demographic events of a species. An alternative method to determine selection is by employing a genome-wide scan of variation (Wright and Gaut 2005; Ramos-Onsins et al. 2008). Demographic processes would affect the entire genome, whereas selection only affects specific loci. Thus, a genome-wide scan of variation can be used as neutral reference for detecting selection. A recent genome-wide survey of variation in *O. rufipogon* and *O. sativa* indicated a complex demographic history (Caicedo et al. 2007). In this study, we have used these recently published genome-wide variation data, consisting of 111 sequence tagged sites (STSs) as a neutral reference to detect selection in cultivated rice and its wild ancestor (Caicedo et al. 2007). In order to determine selection with greater confidence, we also employed several additional approaches.

In this study, we focus on six major starch genes that play critical roles in the 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, *O. rufipogon*? 2) What is the relationship of nucleotide variation and position of genes in rice endosperm starch synthesis pathway in *O. rufipogon*? 3) What is the relative role of selection and genetic drift in shaping the variation pattern of these six starch genes before and after rice domestication?

Materials and Methods

Plant Materials

Both cultivated rice, *O. sativa*, and its wild progenitor, *O. rufipogon*, were used for this study. The accessions of *O. sativa* and *O. rufipogon* were chosen to represent the diversity found within each species. The *O. sativa* collections include representatives of five variety groups identified by a previous study (Garris et al. 2005): 34 *indica*, 26 *tropical japonica*, 21 *temperate japonica*, 10 *aus*, and 8 *aromatic* accessions. The *O. rufipogon* samples include 70 accessions representing its geographic range except for Australia. Some accessions were field collected (China

and Thailand), whereas all other accessions of *O. rufipogon* and *O. sativa* and outgroup species, *O. meridionalis*, were obtained from the International Rice Research Institute (IRRI) (Manila, Philippines). Detailed information for the accessions is listed in [supplementary tables 1 and 2](#) ([Supplementary Material](#) online).

DNA Extraction, Polymerase Chain Reaction, and Sequencing

DNA was extracted from dried leaves by a modified CTAB method (Doyle JJ and Doyle JL 1990). Except for the *O. rufipogon* leaf material collected from the field, seeds from IRRI were grown for leaf material 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 approximately 1 kb with 100 bp overlap between neighboring amplified regions for each gene. Polymerase chain reactions (PCRs) were conducted in a TX2 or PTC-100 thermal cycler. The PCR solution is 20 μ l, which includes 1 \times Tag buffer, 2 mM dNTP mix, 1 μ M primers, 1 u/20 Taq polymerase, 2.5 mM MgCl₂, 1 μ g template DNA, and sterile deionized water. PCRs were conducted under the following conditions: 95 °C for 5 min; 40 cycles of 95 °C for 50 s, 52–60 °C for 1 min, and 72 °C for 2.5 min; 10 min of extension at 72 °C. The annealing temperature for PCRs differs by primer. The primers and the annealing temperature for PCRs are listed in [supplementary table 3](#) ([Supplementary Material](#) online). 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 sequencing primers for five newly sequenced starch genes are listed in [supplementary table 4](#) ([Supplementary Material](#) online).

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 sequences of Nipponbare were from GenBank and were included in the analyses for *temperate japonica* varieties (<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 is a predominantly selfing species, and no heterozygous single nucleotide polymorphisms (SNPs) were found in our samples. However, the wild species progenitor, *O. rufipogon*, is a predominantly outcrossing species. Some samples from IRRI and almost all 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 burn-in steps (500,000) and sampling interval (500).

Nucleotide Variation

Only samples with less than 50-bp missing data were included in the analyses. The missing data are due to the heterozygosity of indels in some samples or to various other reasons during the sequencing process. Therefore, slightly different sample set were used for different starch genes. Statistics for the level of variation (number of polymorphic synonymous, nonsynonymous, and silent sites; average pairwise nucleotide diversity, θ_{π} ; average number of segregating sites, θ_w) (Watterson 1975; Tajima 1983) were performed in DnaSP v5.0 (Librado and Rozas 2009). Only silent sites that include synonymous sites within coding regions and noncoding regions were used for estimation of θ_{π} and θ_w .

Because transposable elements were observed in *SSIIa* and *GBSSI* (see Results), sliding window analyses were conducted across *SSIIa* and *GBSSI* in *O. rufipogon* to examine the contribution of transposable elements to the level of diversity. The window size was 100 bp, and the step size was 20 bp. In order to search for an association between either the level of variation (as measured by θ_{π} and θ_w) or the 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*.

The population recombination parameter ρ ($4N_e r$, where N_e is the effective population size and 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 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).

Tests of Selection

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 (NIs) were calculated as $R_p S_f / R_f S_p$ (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 evidence of positive selection, whereas 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 burn-in and sampling were extended to 5,000 steps.

The HKA test (Hudson et al. 1987) 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 known neutral loci are considered under selection. HKA tests were performed for each rice variety in addition to *O. rufipogon*. HKA tests were performed in the program HKA with 10,000 simulations (<http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/HKA>). Chi-square tests were used to determine statistical significance ($P < 0.025$).

Since the HKA test showed a statistically significant result for *tropical japonica* rice, 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 specific 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>.

In order to compare the pattern of diversity of the starch genes with that expected by a 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 v5.0 (Tajima 1989; Fay and Wu 2000; Rozas et al. 2003). The species *O. meridionalis* served as the outgroup for calculating Fay and Wu's H . Significant deviations from standard neutral model were determined by coalescent simulation based on 5% significance level.

Although the HKA test has proved to be robust, it only utilizes nucleotide polymorphism within species and divergence between species and does not consider the allele frequency spectrum. Therefore, we also compared the allele frequency spectrum estimated by Tajima's D and Fay and Wu's H of these starch genes with 111 STS loci (Caicedo et al. 2007). A gene is considered under selection when it significantly deviates from genome-wide STS loci and from a standard neutral model based on both Tajima's D and Fay and Wu's H statistics.

Among the 111 STS loci, six loci have no outgroup sequences to calculate Fay and Wu's H value. Among the remaining 105 loci, 76.2% to 94.3% of STS loci have less than three polymorphic nucleotide sites within the rice varieties, which reduces the statistical support for Tajima's D and Fay and Wu's H statistics due to the small number of polymorphic sites per STS locus (supplementary table 5, Supplementary Material online). To overcome this problem, we randomly combined five STS loci together for 1,000 replicates, which made 1,000 randomly combined STS loci (RCSTS).

In order to make the comparison meaningful, we equated our sampling for the starch genes with that for the STS loci by using accessions from a similar geographic range. Tajima's D and Fay and Wu's H values for each RCSTS locus and each resampled studied genes were calculated. Each starch gene was compared with 1,000 RCSTS loci. Genes with Tajima's D or Fay and Wu's H values lower or higher than 97.5% of the RCSTS loci are considered significantly different from RCSTS loci. A Bonferroni correction was used to correct the significance level for multiple tests. The above analyses were conducted by the scripts written in Perl.

Results

Level of Diversity in Wild and Cultivated Rice

Five genes (*AGPL2*, *AGPS2b*, *ISA1*, *SBE1IB*, and *SSIIa*) that code for components of the starch synthesis pathway were examined. Published sequence data for gene *GBSSI* was included in our analyses (Olsen et al. 2006). 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 the study. The location and size of the sequenced region for each gene are shown in figure 2.

Estimates of nucleotide polymorphism for the starch genes in *O. rufipogon* are given in table 1. Among all the starch genes, the lowest level of variation was found at *AGPL2* ($\theta_\pi = 0.00146$ and $\theta_w = 0.00334$) and the highest at *GBSSI* ($\theta_\pi = 0.0213$ and $\theta_w = 0.0173$). The level of diversity of the starch genes in *O. rufipogon* is not significantly different from that of RCSTS loci except for *GBSSI* (fig. 3), which has significantly higher variation.

Estimates of nucleotide polymorphism of starch genes in cultivated rice are shown in table 2. The lowest θ_π and θ_w values were found in *tropical japonica* at *GBSSI* ($\theta_\pi = 0.00018$, $\theta_w = 0.00033$) and the highest in *indica* at *SSIIa* ($\theta_\pi = 0.00544$, $\theta_w = 0.00529$).

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

No significant association was detected in *O. rufipogon* 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 (data not shown).

Estimates of population recombination rate (ρ), minimum recombination events (R_m), and permutation tests are shown in table 3. Wild rice, *O. rufipogon*, had higher estimates of R_m (2–13) than any cultivated rice variety group at all starch genes except for *GBSSI* in *indica*. *Oryza rufipogon* also had higher ρ values (4.040–19.192 per kb) at the starch genes except for *ISA1* and *GBSSI* in *indica*.

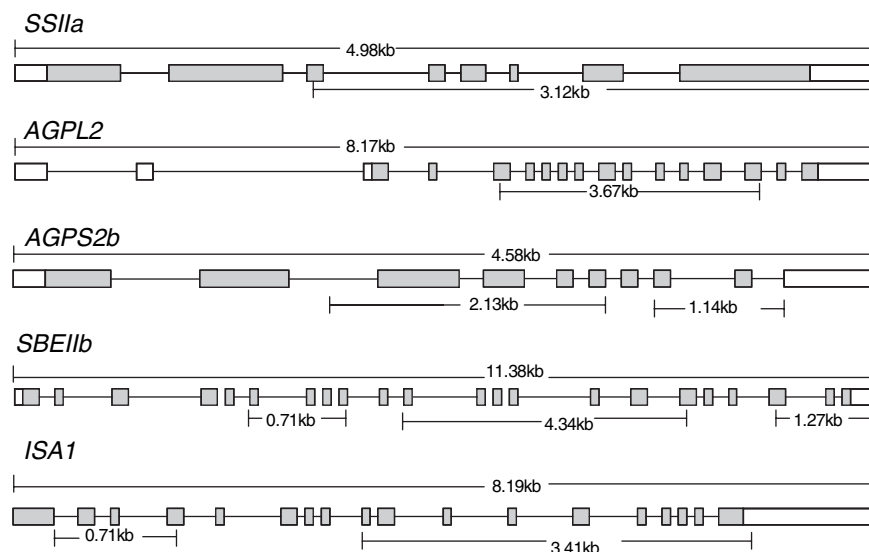


Fig. 2 Location and size of the sequenced regions; the boxes represent exons; shaded boxes correspond to translated regions; the lines connecting the boxes are introns; the gene size is indicated above the boxes. The sequenced regions and size are shown under the boxes.

Significant recombination was detected in *O. rufipogon* at *SBEIIb*, *SSIIa*, and *GBSSI* genes. Among rice varieties, *indica* and *tropical japonica* have slightly higher estimation of R_m or ρ values than do the other three rice varieties at all starch genes. No significant recombination was detected by permutation tests in *aus*, *aromatic*, and *temperate japonica* at five out of six starch genes. However, *indica* and *tropical japonica* showed significant recombination at three out of six starch genes.

By blasting against the *Oryza* repeat database (<http://rice.plantbiology.msu.edu/blast.shtml>), we detected a miniature inverted-repeat transposable element (MITE) with a size of approximately 360 bp in intron 7 of *SSIIa*. A different MITE 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 stow-away family, which is associated with the genes in Angiosperms and is AT rich (around 72%) (Bureau and Wessler 1994). Previous studies indicate a short interspersed element (about 125 bp) in intron 1 and a MITE (about 75 bp) in intron 13 of *GBSSI* gene (Umeda et al. 1991). Nucleotide variation analysis indicates that there is higher variation at these transposable element regions than in the rest of the gene regions of both *SSIIa* and *GBSSI* in *O. rufipogon* (supplementary fig. 1, Supplementary Material online).

Tests of Selection

The results of MK tests are shown in table 4. Only *AGPL2* yielded a significant result by Fisher's exact test ($P = 0.04$). The neutral indexes range from 0.0 to 2.45. The neutral indexes are not available for *AGPL2* 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 of STS loci in some rice groups are statistically significant (supplementary table 6, Supplementary Material online), suggesting that some STS loci might be under selection. This result is consistent with the previous study by coalescent simulation (Caicedo et al. 2007). Because STS loci are not a good neutral reference for HKA tests, HKA tests were performed on the starch genes alone. In order to include all the accessions of this study, HKA test were performed for five starch genes (no *GBSSI*). No significance was detected (supplementary table 7, Supplementary Material online). After equating the samples of the five starch genes with *GBSSI*, HKA tests including *GBSSI* were performed and showed significant result in *tropical japonica* (table 5), suggesting selection at some starch loci in *tropical japonica*.

MLHKA tests were then performed to detect selection in *tropical japonica* at individual loci (table 6). The results

Table 1. Population Statistics of Starch Genes in *O. rufipogon*.

| Statistics | <i>AGPL2</i> | <i>AGPS2b</i> | <i>ISA1</i> | <i>SBEIIb</i> | <i>SSIIa</i> | <i>GBSSI</i> |
|-------------------|--------------|---------------|-------------|---------------|--------------|--------------|
| 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 |
| θ_π | 0.00146 | 0.00285 | 0.00267 | 0.00296 | 0.00566 | 0.02127 |
| θ_w | 0.00334 | 0.00464 | 0.00543 | 0.00519 | 0.00738 | 0.01728 |
| Tajima's <i>D</i> | -1.80957* | -1.22540 | -1.55277 | -1.41417 | -0.82375 | 1.18644 |

* $P < 0.05$.

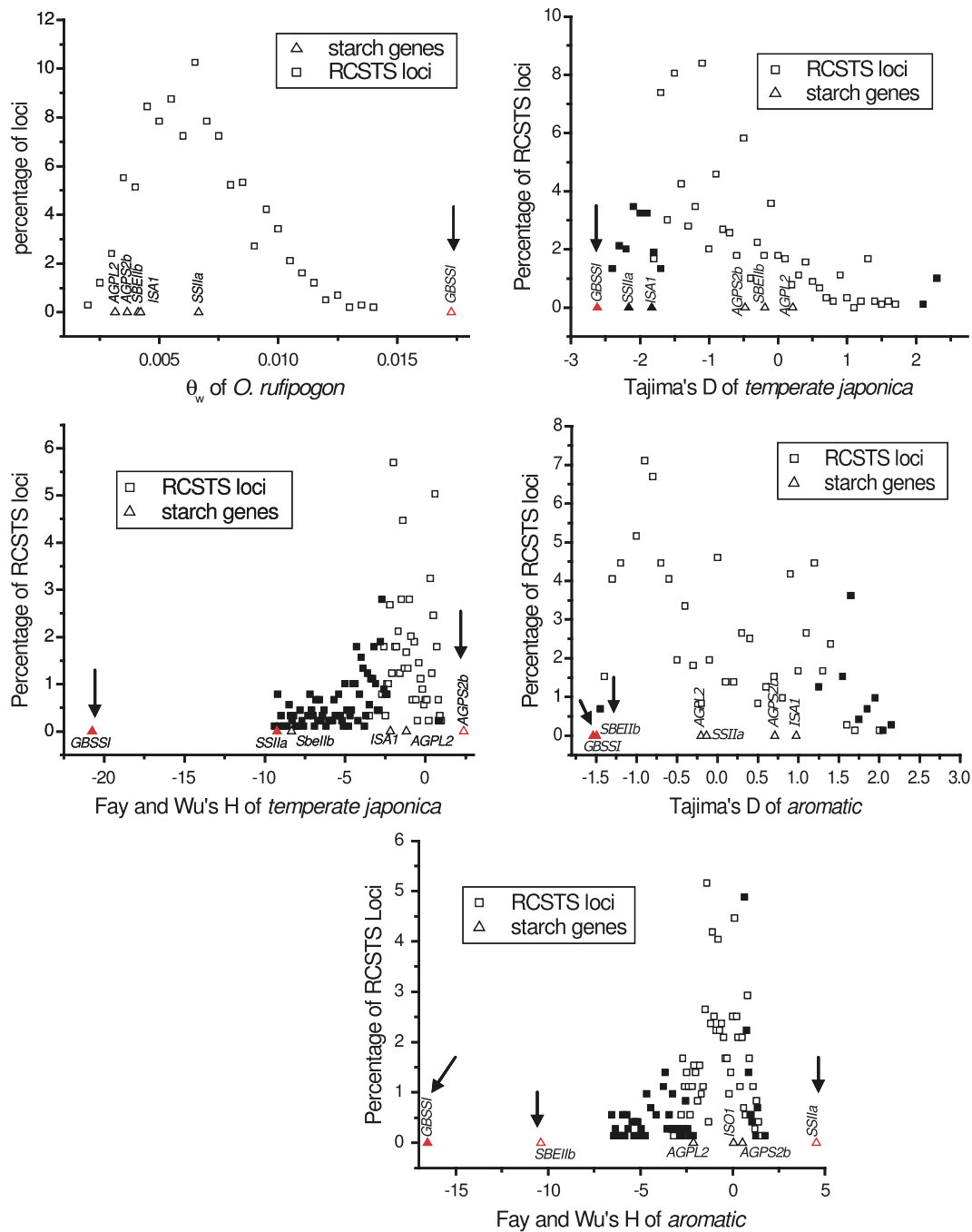


FIG. 3 The results of comparison between studied starch genes and RCSTS loci in *O. rufipogon*, *aromatic*, and *temperate japonica* rice. Triangles represent the studied starch genes. Squares represent RCSTS loci. Filled triangles or squares represent significant deviation from the standard neutral model at a 5% significant level. A red triangle indicates that the gene significantly deviates from genome-wide RCSTS loci at the 5% significant level. A gene indicated by an arrow deviates significantly from genome-wide RCSTS loci after Bonferroni correction.

showed significant differences between a neutral model and the model with selection at *GBSSI* ($P < 0.05$), which suggests selection at *GBSSI* in *tropical japonica*. The selection parameter (k value) at *GBSSI* in *tropical japonica* is 0.13, which indicates that the level of diversity at *GBSSI* 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 the other five starch genes in *tropical japonica*.

Contrasting the Observed Data Against the Standard Neutral Model

Values of Tajima's D for *O. rufipogon* are given in table 1. These statistics show a broad range of values across the starch genes. In *O. rufipogon*, *AGPL2* shows a significantly negative Tajima's D value, which reflects an excess of low frequency alleles.

Value of Tajima's D and Fay and Wu's H value of cultivated rice are given in table 7. Both Tajima's D and Fay and

Table 2. Nucleotide Polymorphism of Starch Genes in Cultivated Rice Variety Groups.

| | <i>Aus</i> | <i>Indica</i> | <i>Tropical japonica</i> | <i>Temperate japonica</i> | <i>Aromatic</i> |
|----------------|------------|---------------|--------------------------|---------------------------|-----------------|
| AGPL2 | | | | | |
| 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 |
| θ_{π} | 0.00102 | 0.00151 | 0.00160 | 0.00143 | 0.00083 |
| θ_w | 0.00123 | 0.00225 | 0.00201 | 0.00214 | 0.00089 |
| AGPS2b | | | | | |
| 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 |
| θ_{π} | 0.00122 | 0.00146 | 0.00287 | 0.00194 | 0.00117 |
| θ_w | 0.00101 | 0.00297 | 0.00308 | 0.00259 | 0.00109 |
| ISA1 | | | | | |
| 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 |
| θ_{π} | 0.00184 | 0.00159 | 0.00151 | 0.00053 | 0.00193 |
| θ_w | 0.00179 | 0.00234 | 0.00180 | 0.00141 | 0.00150 |
| SBE1b | | | | | |
| 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 |
| θ_{π} | 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 |
| θ_{π} | 0.00387 | 0.00529 | 0.00310 | 0.00276 | 0.00527 |
| θ_w | 0.00365 | 0.00544 | 0.00420 | 0.00364 | 0.00514 |
| GBSSI | | | | | |
| 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 |
| θ_{π} | 0.00279 | 0.00594 | 0.00018 | 0.00109 | 0.00372 |
| θ_w | 0.00209 | 0.00546 | 0.00033 | 0.00328 | 0.00489 |

Wu's H values deviate significantly from the standard neutral model at *GBSSI* in *temperate japonica*, at *AGPS2b* in *indica*, and at *SBE1b* and *GBSSI* in *aromatic* rice. However, no genes showed significant values of both Tajima's D and Fay and Wu's H in *aus* and *tropical japonica*.

The Tajima's D and Fay and Wu's H values for starch genes for the resampled rice groups are shown in figure 3 and supplementary figure 2 (Supplementary Material

online). Consistent patterns of deviation from the standard neutral model were found before and after resampling in *O. rufipogon*. Within each cultivated rice group, consistent patterns were found at *AGPL2* and *ISA1*, and slight, inconsistent patterns were found at *SSIIa*, *AGPS2b*, and *SBE1b* before and after resampling. These slight inconsistent patterns of Tajima's D and Fay and Wu's H value before and after resampling indicate sensitivity of those statistics to

Table 3. Population Recombination Rate and Minimum Recombination Events (Rm) Estimates in Starch Genes.

| Gene | <i>Oryza rufipogon</i> | <i>Aus</i> | <i>Indica</i> | <i>Tropical japonica</i> | <i>Temperate japonica</i> | <i>Aromatic</i> |
|---------------|------------------------|---------------|----------------|--------------------------|---------------------------|-----------------|
| <i>AGPL2</i> | 4.040 (2) | 2.020 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| <i>AGPS2b</i> | 9.091 (5) | 3.030 (0) | 0 (0) | 3.030**** (1) | 0 (0) | 0 (0) |
| <i>ISA1</i> | 11.111 (8) | 0 (1) | 27.273**** (6) | 8.081 (2) | 0 (0) | 0 (0) |
| <i>SBE1b</i> | 12.121***** (6) | 3.061**** (2) | 5.051** (2) | 3.030 (3) | 2.020 (2) | NA |
| <i>SSIIa</i> | 19.192*** (13) | 0 (0) | 6.061 (8) | 6.061***** (2) | 5.051* (2) | 16.162**** (2) |
| <i>GBSSI</i> | 4.040**** (8) | 1.010 (0) | 3.030* (10) | 5.051**** (1) | 15.152 (1) | NA |

NOTE.—Rm is shown in parentheses; test of recombination by permutation test are shown by *; $P < 0.05$; ** $P < 0.01$; *** $P < 0.005$; **** $P < 0.001$. NA, not available.

Table 4. MK Tests of Starch Genes in *Oryza rufipogon*.

| Gene | Polymorphism | | Fixed Differences | | P^c | NI ^d | Gamma ^e |
|--------|----------------|----------------|-------------------|----|-------|-----------------|--------------------|
| | R ^a | S ^b | R | S | | | |
| AGPL2 | 7 | 5 | 0 | 6 | 0.04* | NA ^f | 3.02 |
| AGPS2b | 0 | 8 | 1 | 0 | 0.11 | 0.00 | 2.65 |
| ISA1 | 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 |
| GBSSI | 4 | 7 | 4 | 13 | 0.67 | 1.86 | 2.29 |

^a Number of replacement sites.

^b Number 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.

* $P < 0.05$.

sampling, which also suggest the risk of detecting selection only by those two statistics in rice.

Contrasting the Observed Data Against RCSTS Data

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 (fig. 3 and supplementary fig. 2, Supplementary Material online). Based on a 5% significant level, AGPL2 and GBSSI deviate significantly from the genome-wide RCSTS data by Tajima's D value. However, none of these values is significant after Bonferroni correction (fig. 3).

Both Tajima's D and Fay and Wu's H in the resampled rice groups were compared with that of RCSTS data (fig. 3 and supplementary fig. 2, Supplementary Material online). All rice groups showed one or two genes which significantly deviate from the genome-wide RCSTS data with Fay and Wu's H and/or Tajima's D based on 5% significance level. However, after Bonferroni correction, only GBSSI in *aromatic* and *temperate japonica* continued to have significantly lower values of both Tajima's D and Fay and Wu's H than the RCSTS data. SBEIIb in *aromatic* rice had a significant lower Fay and Wu's H value than that of RCSTS data after Bonferroni correction. SBEIIb also has lower value of Tajima's D in *aromatic* varieties than the RCSTS data, although it is not a significant deviation after Bonferroni correction ($P = 0.0098$). SBEIIb in *aus* showed significantly lower value of Fay and Wu's H than the RCSTS data after Bonferroni correction. AGPS2b in *tropical japonica* and *temperate japonica*, SSIIa in *aromatic* have significantly higher values of Fay and Wu's H than the RCSTS data.

Table 5. HKA Test Results of Six Studied Starch Genes in *Oryza rufipogon* and *O. sativa*.

| Rice Group | Deviation ^a | P_{chi}^b |
|---------------------------|------------------------|--------------------|
| <i>O. rufipogon</i> | 7.3093 | 0.19863 |
| <i>Aus</i> | 4.3105 | 0.50563 |
| <i>Indica</i> | 2.1123 | 0.83340 |
| <i>Tropical japonica</i> | 17.4076 | 0.00379* |
| <i>Temperate japonica</i> | 3.5296 | 0.61891 |
| <i>Aromatic</i> | 2.8832 | 0.71799 |

^a The sum of deviations in chi-square test.

^b Chi-square distribution probability.

* $P < 0.05$.

Discussion

Pattern of Sequence Diversity at Starch Genes in Wild and Cultivated Rice

In *O. rufipogon*, the level of diversity at GBSSI is higher than that of any other starch gene and is significantly greater than that of the RCSTS data. The high diversity at GBSSI may possibly be due to transposable elements, high mutation rate, or relaxation of selection at this gene. GBSSI has two transposable elements, and the transposable element regions show high diversity in *O. rufipogon*, suggesting a possible contribution of transposable elements to the high level of diversity at GBSSI in *O. rufipogon*. The high diversity in the transposable element regions may be explained by the high mutation rate associated with transposable elements that has been documented in several species (Souames et al. 2003; Koga et al. 2006). High diversity at GBSSI might also be due to relaxation of selection at this gene. All the tests for selection including HKA, MK, Tajima's D , Fay and Wu's H , and the comparison with genome-wide RCSTS data indicated that GBSSI was under neutral evolution in *O. rufipogon*. However, a clear explanation of the high diversity at GBSSI in *O. rufipogon* requires more detailed information of genome-wide level of variation.

Our study of starch genes shows significantly higher levels of diversity in *O. rufipogon* than in *O. sativa*. This result is consistent with many previous studies (Caicedo et al. 2007; Zhu et al. 2007) and could be explained by bottleneck events during rice domestication. Genetic bottlenecks associated with domestication have been documented in many crop species, such as rice, corn and wheat, as well as in domesticated animal species (Eyre-Walker et al. 1998; Briggs and Goldman 2006; Hyten et al. 2006; Miller and Schaal 2006; Haudry et al. 2007; Zhu et al. 2007; Kavar and Dovic 2008).

Association between Nucleotide Variation and Position in the Metabolic Pathway

Recent studies suggest that genes that code for upstream enzymes are more selectively constrained than downstream genes in a metabolic pathway (Campbell et al.

Table 6. The Results of MLHKA Tests in *Tropical japonica*.

| Model | Description | lnL ^a | Comparison | Likelihood-Ratio Statistics (df) | p ^b | K ^c |
|-------|----------------------------|------------------|------------|----------------------------------|----------------|----------------|
| A | Neutral (all K = 1) | -41.5428 | | | | |
| B | Selection at <i>AGPL2</i> | -41.3914 | A vs. B | 0.1514 (1) | >0.05 | 1.37 |
| C | Selection at <i>AGPS2b</i> | -40.7597 | A vs. C | 0.7831 (1) | >0.05 | 1.94 |
| D | Selection at <i>SSIIa</i> | -41.6383 | A vs. D | -0.0955 (1) | >0.05 | 1.28 |
| E | Selection at <i>SBEIIb</i> | -41.256 | A vs. E | 0.2868 (1) | >0.05 | 1.46 |
| F | Selection at <i>ISA1</i> | -41.4092 | A vs. F | 0.1336 (1) | >0.05 | 1.66 |
| G | Selection at <i>GBSSI</i> | -35.5304 | A vs. G | 6.0124 (1) | <0.05* | 0.13 |

NOTE.—df, degrees of freedom.

^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.

1998; Rausher et al. 1999; Fraser et al. 2002; Hahn and Kern 2005). This pattern could be predicted from classical biochemical theories of metabolic regulation: first, biochemical theories assert that the enzymes at upstream points of a metabolic pathway control larger flux and end product concentration (and hence the phenotype) than do downstream enzymes (Atkinson 1977; Crabtree and Newsholme 1987). Second, the enzymes on upstream points of a pathway potentially influence a greater number of end products than do the downstream enzymes (Rausher et al. 1999). However, our results showed no significant association between the levels or patterns of diversity and position within the starch metabolic pathway of *O. rufipogon*. The results should be viewed with caution since only a subset of the starch synthesis pathway genes (six genes) were included in the analysis and the analysis only considers gene position in the pathway and does not consider the other parts of the pathway reticulation.

Evolution of Starch Genes in *O. rufipogon* Before Domestication

Significant excess of nonsynonymous mutations in *O. rufipogon* was detected at *AGPL2*, suggesting either positive or balancing selection that favors amino-acid replacements in the protein product or the relaxation of purifying selection. 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 (number of fixed replacement mutations: 0–4, number of fixed synonymous mutations: 0–13) (Andolfatto 2008). Furthermore, if *AGPL2* is under positive or balancing selection, a high frequency of derived alleles is expected among the seven nonsynonymous mutations within *O. rufipogon*. However, we observed a very low frequency of derived alleles in all seven nonsynonymous mutations (1–3 plants of 63 *O. rufipogon* individuals). These data suggest that the alleles at the seven nonsynonymous sites are either under strong purifying selection or were recently derived. However, it seems unlikely that all seven of those nonsynonymous alleles would be recently derived.

Based on all the tests used for detecting selection, no strong evidence of selection was detected in five *O. rufipogon* starch genes (except *AGPL2*) before rice domestication. In contrast, *AGPL2* might have experienced stronger selective constraints than the other genes. First, among all the genes, *AGPL2* has the lowest diversity, and its level of diversity based on θ_w value is lower than 95.17% of RCSTS loci. Second, significant negative Tajima's *D* value was observed at *AGPL2*, and this value deviates significantly from genome-wide RCSTS data at the 5% significance level. A strong selective constraint on *AGPL2* was also documented in *Zea mays* (Whitt et al. 2002; Manicacci et al. 2007). A strong selective constraint at *AGPL2* in *O. rufipogon* might be due to its large effect on starch phenotype and its position in the starch synthesis pathway. Mutations in *AGPL2* can result in shrunken seeds in both corn

Table 7. Tajima's *D* and Fay and Wu's *H* Statistics in Cultivated Rice Groups.

| Gene | Statistics | <i>Aus</i> | <i>Indica</i> | <i>Tropical japonica</i> | <i>Temperate japonica</i> | <i>Aromatic</i> |
|---------------|---------------------------|------------|---------------|--------------------------|---------------------------|-----------------|
| <i>AGPL2</i> | <i>D</i> ^a | -0.739 | -1.14134 | -0.73957 | -1.24016 | -0.34485 |
| | FW- <i>H</i> ^b | -2.93333 | -2.73084 | -2.56 | -4.75238 | -2.57143 |
| <i>AGPS2b</i> | <i>D</i> ^a | 0.89971 | -1.83339* | -0.33271 | -0.94746 | 0.34043 |
| | FW- <i>H</i> ^b | 0.08889 | -6.16756** | -3.85231 | -4.38571 | 0.64286 |
| <i>ISA1</i> | <i>D</i> ^a | -1.55277 | -1.19512 | -0.52283 | -2.27545** | 1.50937 |
| | FW- <i>H</i> ^b | -4.53333 | -4.6631 | 2.51077 | -1.46667 | -0.85714 |
| <i>SBEIIb</i> | <i>D</i> ^a | 1.41918 | 0.81449 | 1.47224 | 2.37741* | -1.84038** |
| | FW- <i>H</i> ^b | -1.77778 | -2.61174 | -2.35077 | -0.99524 | -14.14286** |
| <i>SSIIa</i> | <i>D</i> ^a | 0.13836 | -0.11018 | -0.86069 | -0.95777 | 0.14506 |
| | FW- <i>H</i> ^b | 1.07143 | 0.25747 | -6.0751* | -4.70526 | 4.21429* |
| <i>GBSSI</i> | <i>D</i> ^a | 2.08735* | 0.37264 | -1.58998 | -2.61492*** | -1.53736* |
| | FW- <i>H</i> ^b | 0.53333 | -1.19048 | 0.22059 | -20.74459*** | -16.53333** |

^a Tajima's *D*.^b Fay and Wu's *H*.**P* < 0.05, ***P* < 0.01, ****P* < 0.001.

and rice (Bhave et al. 1990; Lee et al. 2007). Both *AGPL2* and *AGPS2b* 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 *GBSSI* only controls amylose production, and enzymes coded by *ISA1*, *SBEIIb*, and *SSIIa* only control amylopectin production (fig. 1).

AGPase is a heterotetramer with two small and two large subunits. Previous studies (Giroux et al. 1995; Georgelis et al. 2008) suggest that the small subunit is more selectively constrained than is the large subunit among angiosperm species because the overall rate of synonymous evolution is ~ 2.7 -fold greater for the large subunit than for the small subunit. Based on this hypothesis, *AGPL2* that codes for one AGPase large subunit should have higher diversity than that of *AGPS2b* in *O. rufipogon*. Our results, however, showed the opposite; *AGPL2* has lower diversity and is more selectively constrained than does *AGPS2b* in *O. rufipogon*, which is also consistent with the results from *Z. mays* (Whitt et al. 2002). Previous studies (Georgelis et al. 2008) indicated that both small and large subunits show branch-specific rates of variation among sites, which suggests that a few angiosperm species such as rice and *Z. mays* may show different patterns of evolution compared with the overall pattern in angiosperms.

The pattern of neutrality at the remaining five *O. rufipogon* starch genes suggests that genetic drift may play a more important role than does selection in shaping the pattern of diversity for these genes before domestication. In contrast, the diversity pattern of *AGPL2* in *O. rufipogon* suggests a relatively more important role for purifying selection. The lack of evidence for positive selection at six starch genes in *O. rufipogon* suggests that these starch genes did not recently contribute to any adaptive shift of starch traits in *O. rufipogon*. Given how important starch genes are to a seed and hence plant fitness, it seems likely that selection on starch genes has occurred at some time in the past. If selection occurred long ago the signal of positive selection may be lost. Finally, it is also possible that other starch genes are responsible for the adaptive shift in *O. rufipogon*. These hypotheses require further studies with more starch genes or characterization of starch quality in *O. rufipogon* and its close relatives.

Evolution of Starch Genes during Domestication

Strong evidence of positive selection was found at *GBSSI* in *tropical japonica*, *temperate japonica*, and *aromatic* rice. However, the nature of evidence for selection was different for these rice varieties. Evidence of positive selection at *GBSSI* in *tropical japonica* is based on levels of diversity: the MLHKA test in *tropical japonica* suggests significant deviation at *GBSSI* compared with the other five starch genes, the level of diversity is the lowest at *GBSSI* in *tropical japonica* among all the rice groups at any gene, and the reduction of diversity at *GBSSI* in *tropical japonica* relative to *O. rufipogon* is the most extreme. In contrast, evidence of positive selection at *GBSSI* in *aromatic* and *temperate japonica* is based on the allele frequency spectrum. *GBSSI*

shows significant deviation in *aromatic* and *temperate japonica* both by comparison with the standard neutral model and with genome-wide RCSTS data for both Tajima's *D* and Fay and Wu's *H* values. Strong selection may result in evidence of positive selection at *GBSSI* in *tropical japonica* detected in the level of diversity but may not be detected in the allele frequency spectrum. Strong selection at *GBSSI* in *tropical japonica* results in a large reduction of diversity and results in the loss of selection signal from the pattern of allele frequency spectrum. In contrast, positive selection at *GBSSI* in *aromatic* and *temperate japonica* might not be strong enough to show significant reduction in the level of diversity in the face of bottleneck events during rice domestication, but selection may be strong enough to show a significant pattern of selection in the allele frequency spectrum. Furthermore, *aromatic* and *temperate japonica* rice have most likely experienced a more severe bottleneck compared with *tropical japonica* because they appear to be derived from *tropical japonica*.

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

Evidence of positive selection was found at some starch genes in *tropical japonica*, *temperate japonica*, and *aromatic* varieties but not in *aus* or *indica* rice. Previous studies suggest independent domestication events for *aus*, *indica*, and *japonicas* (Londo et al. 2006). Therefore, it is possible that starch quality is a trait that is under independent selection during the domestication of *aromatic* and *japonica* rice. However, *GBSSI* is under selection in *aromatic* varieties as it is in *japonica* rice. These results could be explained as a single selection event at *GBSSI* during domestication of *japonicas* and *aromatic* rice, assuming a single origin for *japonicas* and *aromatic*. It is also possible that the *GBSSI* gene is independently selected in *aromatic* and *japonicas* after their divergence. Furthermore, *SBEIIb* is under positive selection in *aromatic* rice not in any other rice varieties, suggesting that it was under positive selection for unique starch quality in *aromatic* rice.

Although the same gene (*GBSSI*) was selected in both *tropical* and *temperate japonica*, and domestication for *japonicas* is a single event, it is unlikely that selection at *GBSSI* in *japonicas* is a single event. Different alleles or sites might be selected for both *japonica* groups. 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 *GBSSI* intron 1, which leads to incomplete posttranscriptional processing of the pre-mRNA and causes undetectable levels of spliced mRNA in glutinous *temperate japonica* individuals (Wang et al.

1995; Frances et al. 1998). The size of the selective sweep caused by selection for this allele in *temperate japonica* is about 250 kb, which suggests 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 allele has increased to a high frequency in *temperate japonica* (17/20) but has a very low frequency in *tropical japonica* (1/18), which suggests that either a different allele or site is under selection at *GBSSI* in *tropical japonica*. Therefore, further investigation is required to understand the exact target of selection at *GBSSI* in *tropical japonica*. It could be concluded here that selection at *GBSSI* 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 *SBE11b*, *GBSSI* in *aromatic* rice, and at *GBSSI* in *tropical japonica* and *temperate japonica* rice suggests that selection on these genes plays a major role during rice domestication. This also suggests that *SBE11b* and *GBSSI* contribute to the adaptive shift of starch traits in *aromatic* rice and that *GBSSI* contributes to the adaptive shift of starch traits in *tropical* and *temperate japonica*. All the other starch candidate genes within cultivated rice show a pattern consistent with neutrality, which suggests that genetic drift may be the determinant of evolution in these genes. No evidence of selection was discovered in *aus* and *indica* rice for any of the six candidate starch genes, which suggests either that there is no adaptive shift of starch traits during domestication or that other starch genes affect starch quality in the endosperm, and these genes were involved in the adaptive shift of starch traits in *aus* and *indica*. These two hypotheses could be tested with further studies of additional candidate genes.

Genes that were under positive selection during domestication were all located downstream in the starch synthesis pathway (see fig. 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. These results are consistent with strong selective constraints on these upstream genes before domestication. A similar case for *AGPL2* has been documented in *Z. mays* (Manicacci et al. 2007). However, selection at downstream genes of the starch synthesis pathway might be simply the efficient response to selection for starch quality during domestication. The starch quality difference could be 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 suggests (fig. 1). Therefore, it might be more efficient for selection to act on downstream genes than on upstream genes. However, more information about the target nucleotides for selection and their effects on starch quality in rice is required to test the above hypothesis.

The Power and Limitation of Detecting Selection by Comparisons With the Genome-Wide Variation Data

Instead of comparing with a “true” demographic model of a species that reflects the demographic events a species experienced, our study indicates the power of a method that uses genome-wide variation data as neutral reference. This method could be applied to more species given the increasing availability of genome-wide data sets which detail population-level variation. These data sets will allow the development of the better demographic models for species with complex demographic histories and can foster the use of new strategies for detecting selection at candidate loci. These strategies will provide not only a better understanding of more species’ demographic histories but also provide more information on the genetic basis of their adaptive traits.

However, there are obvious limitations. First, this approach requires that the comparisons be done for the same set of samples. Such comparison could be straightforward for model species, such as rice or maize given their public germplasm banks. Second, this approach assumes that the pattern of genome-wide variation should mainly reflect the species neutral demographic history. However, the pattern of genome-wide variation might also be affected by selective forces, particularly in selfing species. High selfing rates will decrease recombination rates, thus amplifying the size of selective sweeps and affecting the pattern of genome-wide variation. Domesticated rice has such a pattern. Variation of genome-wide STS loci in *tropical japonica* and *indica* is better explained by demography plus selective sweeps instead of a neutral demographic model (Caicedo et al. 2007). Therefore, it might be too conservative to detect positive selection in rice by using STS data as neutral reference. Furthermore, a neutral gene might show significantly higher positive value of Tajima’s *D* or Fay and Wu’s *H* values than RCSTS data due to the effect of selective sweeps. Therefore, it is potentially risky to determine balancing selection by this genome-wide approach. In this study, we also compared the observed data with the standard neutral model to overcome this problem.

Supplementary Material

Supplementary figures S1–S2 and tables S1–S7 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We thank Dr Justin Fay for valuable suggestions on data analysis and Dr Ana Caicedo for use of the genome-wide STS data alignments. This work was supported by the National Science Foundation (DBI 0701382 to B.A.S.).

References

- Andolfatto P. 2008. Controlling type-I error of the McDonald-Kreitman test in genome wide scans for selection on noncoding DNA. *Genetics* 180:1767–1771.

- Atkinson DE. 1977. Cellular energy metabolism and its regulation. New York: Academic Press.
- Bhave MR, Lawrence S, Barton C, Hannah LC. 1990. Identification and molecular characterization of shrunken-2 cDNA clones of maize. *Plant Cell*. 2:581–588.
- Briggs WH, Goldman IL. 2006. Genetic variation and selection response in model breeding populations of *Brassica rapa* following a diversity bottleneck. *Genetics* 172:457–465.
- Buléon A, Colonna P, Planchot V, Ball S. 1998. Starch granules: structure and biosynthesis. *Int J Biol Macromol*. 23:85–112.
- Bureau TE, Wessler SR. 1994. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*. 6:907–916.
- Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in *Arabidopsis*. *Nature* 416:531–534.
- Caicedo A, Williamson S, Hernandez RD, et al. (12 co-authors). 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet*. 3:e163.
- Campbell SL, Khosravi-Far R, Rossman KL, Clark GJ, Der CJ. 1998. Increasing complexity of Ras signaling. *Oncogene* 17:1395–1413.
- Clegg MT. 1997. Plant genetic diversity and the struggle to measure selection. *J Hered*. 88:1–7.
- Collado-Vides J, Hofstad R. 2002. Gene regulation and metabolism: post-genomic computational approaches. Cambridge: MIT Press.
- Crabtree B, Newsholme EA. 1987. The deviation and interpretation of control coefficient. *J Biochem*. 247:113–120.
- Doyle JJ, Doyle JL. 1990. Isolation of plant DNA from fresh tissue. *Focus*. 12:13–15.
- Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online*. 1:47–50.
- Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS. 1998. Investigation of the bottleneck leading to the domestication of maize. *Proc Natl Acad Sci U S A*. 95:4441–4446.
- Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
- Frances H, Bligh H, Larkin PD, Roach PS, Jones CA, Fu HY, Park WD. 1998. Use of alternate splice sites in granule-bound starch synthase mRNA from low-amylose rice varieties. *Plant Mol Biol*. 38:407–415.
- Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296:750.
- Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.
- Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S. 2005. Genetic structure and diversity in *Oryza sativa* L. *Genetics* 169:1631–1638.
- Georgelis N, Braun EL, Hannah LC. 2008. Duplications and functional divergence of ADP-glucose pyrophosphorylase genes in plants. *BMC Evol Biol*. 8:232.
- Giroux M, Smithwhite B, Gilmore V, Hannah LC, Preiss J. 1995. The large subunit of the embryo isoform of Adp glucose pyrophosphorylase from maize. *Plant Physiol*. 108:1333–1334.
- Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol*. 22:803–806.
- Haudry A, Cenci A, Ravel C, et al. (11 co-authors). 2007. Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol Biol Evol*. 24:1506–1517.
- Hirano HY, Eiguchi M, Sano Y. 1998. A single base change altered the regulation of the waxy gene at the posttranscriptional level during the domestication of rice. *Mol Biol Evol*. 15:978–987.
- Hudson RR. 2001. Two-locus sampling distributions and their application. *Genetics* 159:1805–1817.
- Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA-sequences. *Genetics* 111:147–164.
- Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
- Hyten DL, Song QJ, Zhu YL, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci U S A*. 103:16666–16671.
- James MG, Denyer K, Myers AM. 2003. Starch synthesis in the cereal endosperm. *Curr Opin Plant Biol*. 6:215–222.
- Kavar T, Dovc P. 2008. Domestication of the horse: genetic relationships between domestic and wild horses. *Livest Sci*. 116:1–14.
- Khush GS. 1997. Origin, dispersal, cultivation and variation of rice. *Plant Mol Biol*. 35:25–34.
- Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167:1513–1524.
- Koga A, Iida A, Hori H, Shimada A, Shima A. 2006. Vertebrate DNA transposon as a natural mutator: the medaka fish Tol2 element contributes to genetic variation without recognizable traces. *Mol Biol Evol*. 23:1414–1419.
- Lee SK, Hwang SK, Han M, et al. (13 co-authors). 2007. Identification of the ADP-glucose pyrophosphorylase isoforms essential for starch synthesis in the leaf and seed endosperm of rice (*Oryza sativa* L.). *Plant Mol Biol*. 65:531–546.
- Li HP, Stephan W. 2005. Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* 171:377–384.
- Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451–1452.
- Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA. 2006. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc Natl Acad Sci U S A*. 103:9578–9583.
- Manicacci D, Falque M, Le Guillou S, Piegu B, Henry AM, Le Guilloux M, Damerval C, De Vienne D. 2007. Maize Sh2 gene is constrained by natural selection but escaped domestication. *J Evol Biol*. 20:503–516.
- McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.
- McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231–1241.
- Miller AJ, Schaal BA. 2006. Domestication and the distribution of genetic variation in wild and cultivated populations of the Mesoamerican fruit tree *Spondias purpurea* L. (Anacardiaceae). *Mol Ecol*. 15:1467–1480.
- Myers AM, Morell MK, James MG, Ball SG. 2000. Recent progress toward understanding biosynthesis of the amylopectin crystal. *J Plant Physiol*. 122:989–997.
- Nakamura Y. 1996. Some properties of starch debranching enzymes and their possible role in amylopectin biosynthesis. *Plant Sci*. 121:1–18.
- Nakamura Y, Francisco PB, Hosaka Y, Sato A, Sawada T, Kubo A, Fujita N. 2005. Essential amino acids of starch synthase IIa differentiate amylopectin structure and starch quality between *japonica* and *indica* rice varieties. *Plant Mol Biol*. 58:213–227.
- Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD. 2006. Selection under domestication: evidence for a sweep in the rice waxy genomic region. *Genetics* 173: 975–983.
- Olsen KM, Purugganan MD. 2002. Molecular evidence on the origin and evolution of glutinous rice. *Genetics* 162:941–950.

- Pan D, Nelson OE. 1984. A debranching enzyme efficiency in endosperms of the sugray-1 mutants of maize. *Plant Physiol.* 74:324–328.
- Ramos-Onsins SE, Puerma E, Balana-Alcaide D, Salguero D, Aguade M. 2008. Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in *Arabidopsis thaliana*. *Mol Ecol.* 17:1211–1223.
- Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol Biol Evol.* 13:735–748.
- Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol.* 16:266–274.
- Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.
- Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.
- Sano Y. 1984. Differential regulation of waxy gene expression in rice endosperm. *Theor Appl Genet.* 68:467–473.
- Smidansky ED, Martin JM, Hannah LC, Fischer AM, Giroux MJ. 2003. Seed yield and plant biomass increases in rice are conferred by deregulation of endosperm ADP-glucose pyrophosphorylase. *Planta* 216:656–664.
- Sokal RR, Rohlf FJ. 1995. *Biometry: the principles and practice of statistics in biological research*, 3rd ed. New York: W.H. Freeman and Co.
- Souames SM, Bonnivard E, Bazin C, Higuete D. 2003. High mutation rate of TPE repeats: a microsatellite in the putative transposase of the hobo element in *Drosophila melanogaster*. *Mol Biol Evol.* 20:1826–1832.
- 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.
- Tanaka N, Fujita N, Nishi A, Satoh H, Hosaka Y, Ugaki M, Kawasaki S, Nakamura Y. 2004. The structure of starch can be manipulated by changing the expression levels of starch branching enzyme Iib in rice endosperm. *Plant Biotech J.* 2:507–516.
- Umeda M, Ohtsubo H, Ohtsubo E. 1991. Diversification of the rice waxy gene by insertion of mobile DNA elements into introns. *Jpn J Genet.* 66:569–586.
- Umamoto T, Yano M, Satoh H, Shomura A, Nakamura Y. 2002. Mapping of a gene responsible for the difference in amylopectin structure between japonica-type and indica-type rice varieties. *Theor Appl Genet.* 104:1–8.
- Wang ZY, Zheng FQ, Shen GZ, Gao JP, Snustad DP, Li MG, Zhang JL, Hong MM. 1995. The amylose content in rice endosperm is related to the posttranscriptional regulation of the waxy gene. *Plant J.* 7:613–622.
- Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7:256–276.
- Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES. 2002. Genetic diversity and selection in the maize starch pathway. *Proc Natl Acad Sci U S A.* 99:12959–12962.
- Wong WSW, Nielsen R. 2004. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* 167:949–958.
- Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* 168:1071–1076.
- Wright SI, Gaut BS. 2005. Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol.* 22:506–519.
- Zhang ZC, Zhang SF, Yang JC, Zhang JH. 2008. Yield, grain quality and water use efficiency of rice under non-flooded mulching cultivation. *Field Crops Res.* 108:71–81.
- Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. 2007. Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol Biol Evol.* 24:875–888.