Non-independent Domestication of the Two Rice Subspecies, *Oryza sativa* subsp. *indica* and subsp. *japonica*, Demonstrated by Multilocus Microsatellites

Li-zhi Gao* and Hideki Innan†, ‡

* Plant Germplasms and Genomics Center, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650204, P. R. China
† Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, 1200 Hermann Pressler, Houston, TX 77030, USA
‡ School of Advanced Sciences, Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan
Running head: Non-independent domestication of Asian cultivated rice

Key Words: Asian cultivated rice, bottleneck, domestication, microsatellite, selection

Corresponding Authors:

Li-zhi Gao
Telephone: 86 (871) 522-3277
Fax: 86 (871) 522-3277
Email: lgao@mail.kib.ac.cn

Hideki Innan
Email: Innan_hideki@soken.ac.jp
ABSTRACT

The origins of the Asian cultivated rice *Oryza sativa* from its wild ancestor *O. rufipogon* have been debated for decades. The question mainly concerns whether it originated monophyletically or polyphyletically. To shed light on the origins and demographic history of rice domestication, we genotyped a total of 92 individual plants from the two *O. sativa* subspecies and *O. rufipogon* for 60 microsatellites. An approximate Bayesian method was applied to estimate demographic parameters for *O. rufipogon* vs. *O. sativa* subsp. *indica* and *O. rufipogon* vs. *O. sativa* subsp. *japonica*. We showed that the *japonica* subspecies suffered a more severe bottleneck than the *indica* subspecies and thus a greater loss of genetic variation during its domestication. Across microsatellite loci there is a significant positive correlation in the reduction of genetic diversity between the two subspecies. The results suggest that completely independent domestication of *indica* and *japonica* subspecies may not explain our data, and that there is at least partial sharing of their ancestral populations and/or recent gene flow between them.
AMONG many domesticated crop species, Asian cultivated rice (*Oryza sativa* L.) is an important staple crop necessary to feed more than half of the world’s population. As a model species with completely sequenced genomes of two subspecies, *indica* and *japonica*, rice holds an unique position to study questions regarding domestication and the history of genetic diversity in crops. *O. sativa* was domesticated from the commonly recognized progenitor, common wild rice (CWR) *O. rufipogon* Griff. (CHANG 1976; OKA 1988; KHUSH 1997; CHENG et al. 2003; YAMANAKA et al. 2004). In contrast to the better-known domestication and evolution of modern maize (DOEBLEY 2004), the origin of *O. sativa* has been an unsolved problem for decades. The long-standing debate concerns whether the Asian cultivated rice originated monophyletically or polyphyletically. The hypothesis of monophyletic origin suggests that wild populations of *O. rufipogon* could have evolved into the *indica* subspecies in South and/or Southeast Asia, from which the *japonica* subspecies evolved through adaptation during spreading to the higher latitudes and elevations (CHANG 1976; OKA 1988; OKA and MORISHIMA 1982; LU et al. 2002). By contrast, the polyphylectic hypothesis postulates that the *indica* subspecies and the *japonica* subspecies each originated from different CWR ancestral populations (CHENG et al. 2003; YAMANAKA et al. 2004; KATO et al. 1928; SECOND 1982; WANG et al. 1992; MOCHIZUKI et al. 1993; DOI et al. 2002; VITTE et al. 2004; GARRIS et al. 2005; ZHU and GE 2005; YAMANAKA et al. 2003; LONDO et al. 2006). The different CWR lineages supposed that the *indica* - and *japonica* - diversification had already occurred before their separate domestication, and subsequently developed into these two major subspecies (SECOND 1982; WANG and SUN 1996).
It has long been believed that most domesticated plants have been reduced genetic variation in comparison with their wild progenitor species due to bottleneck events (Doebley et al. 2006). Domesticated species usually come from relatively small founder populations, i.e. from a subsample of the wild progenitor population (Eyre-Walker et al. 1998). Such bottleneck events cause a genome-wide loss of genetic variation. Another major factor decreasing genetic diversity is artificial selection. This occurs when a trait is favored under selection during the domestication process. Such selection could reduce the amount of polymorphisms in the surrounding regions of the selected target loci. The contributions of these two factors to the patterns of polymorphisms in cultivated species have been well documented in maize (Tenaillon et al. 2004; Wright et al. 2005), but less data are available for rice. Early genetic data showed an apparent loss of genetic diversity in rice during the process of domestication (Londo et al. 2006; Sun et al. 2001; Gao et al. 2006), but a number of questions regarding how selection and bottlenecks shaped the standing patterns of polymorphisms in the two rice subspecies are intriguing but largely unanswered.

In this article, we first inferred the domestication process of the two rice subspecies by using polymorphism data at multiple microsatellites. We then examined the two hypotheses on origins of Asian cultivated rice, focusing on the reduction of genetic diversity in the two rice subspecies at microsatellite loci. If the two subspecies originated from completely different ancient CWR populations, there should be no correlation between loci in the reduction of genetic variation. On the other hand, a strong correlation should be observed if the two subspecies were domesticated from the same founder populations. To test these hypotheses, we genotyped representative samples of two
subspecies of cultivated rice and CWR, and found a strong positive correlation in the
degree of reduction of genetic diversity between the two subspecies.

**MATERIALS AND METHODS**

**Plant Materials:** A total of 92 accessions of Asian cultivated rice *O. sativa* and wild
progenitor *O. rufipogon* were sampled and genotyped. They included 22 accessions of *O.
sativa* subsp. *indica*, 35 accessions of *O. sativa* subsp. *japonica*, and 35 accessions of *O.
rufipogon*. For each accession, one to five individuals were assayed to detect the
polymorphisms. The information of the studied accessions (accession name, accession
number, and geographical origins) was described elsewhere (GAO *et al.* 2005).

**Genomic DNA Extraction and Microsatellite Genotyping:** Five loci for each
chromosome were selected totaling 60 primer pairs that are randomly distributed
throughout the rice genome (*WU* *et al.* 1993; *PANAUD* *et al.* 1996; *AKAGI* *et al.* 1996;
*CHEN* *et al.* 1997). These are also displayed on the Cornell University Rice Genes web
site (http://www.gramene.org/microsat/ssr.txt). Microsatellite polymorphisms were
assayed by specific PCR conditions (*PANAUD* *et al.* 1996). PCR products were run on 6
% polyacrylamide denaturing gels. Marker bands were detected using the silver staining
methods (*PANAUD* *et al.* 1996). The null alleles were confirmed after several repetitions
with different amplification conditions to ensure that no reaction failure existed. To
determine the allele size, the samples were directly compared with band sizes from an
allelic ladder prepared by the amplification of an artificial mixture of DNA from all the
assayed samples.
**Diversity Analysis:** In a total of 60 assayed microsatellites, the 46 di-nucleotide microsatellites were used in data analysis; the other 14 microsatellites were excluded because they are microsatellites with either more than two repeats or other motif-complicated repeats. The level of microsatellite variation was measured as the expected average heterozygosity ($H$) (Table 1). For the typical selfers, $H$ was defined as the probability that randomly chosen alleles from different individuals are not identical. Wright’s $F$-statistics were also estimated from the excess of homozygous individuals conditional on the estimates of $H$ (WRIGHT 1965). The average $H$ for *O. rufipogon*, *O. sativa* subsp. *indica* and *O. sativa* subsp. *japonica* are denoted by $H_{W,obs}$, $H_{I,obs}$, and $H_{J,obs}$, respectively.

**Genetic Relationship Tree:** We used data of all 60 microsatellite loci to construct a tree representing genetic relationships (Figure 1A). C.S. Genetic distances were estimated by using Chord distance (CAVALLI-SFORZA and EDWARDS 1967), which was shown to generate correct tree topologies regardless of the microsatellite mutation model (TAKEZAKI and NEI 1996). Phylogenetic reconstruction was based on the neighbor-joining method implemented in PowerMarker version 2.7 (LIU and MUSE 2004). Using NEIGHBOR in the PHYLIP computer package (version 3.5c) (FELSENSTEIN 1995), we constructed an unrooted Neighbor-joining (NJ) tree (Figure 1A). The robustness was assessed by creating 999 bootstrap replicates of the data set with the SEQBOOT algorithm in PHYLIP, and then generating a majority rule consensus tree in the
CONSENSE program. These distance trees were viewed by the program TREEVIEW (PAGE 1996).

**Population Structure:** Population structure among *O. sativa* and *O. rufipogon* accessions was evaluated with the model-based program STRUCTURE 2.1 (PRITCHARD *et al.* 2000). The analysis had a burn-in length of 50,000 iterations and a run length of 100,000 iterations. The correlated allele frequency model was used with asymmetric admixture allowed. The highest log likelihood scores were obtained when the number of populations was set at three. Five independent runs yielded consistent results. The result is presented in Figure 1B.

**Demographic Model:** The population model of domestication used here (Figure 2) is similar to previous studies (WRIGHT *et al.* 2005; INNAN and KIM 2004). The model approximates the demography of a domestication event of a single cultivated species from its wild progenitor. In this study, the model was used for analyzing jointly CWR and *O. sativa* subsp. *indica* and for analyzing jointly CWR and *O. sativa* subsp. *japonica*, respectively. It is assumed that the wild progenitor species has a constant-size population with size $N_0$. The domestication begins at time $T_d$ in a small founder population of the cultivated species, which is merely a subset of the members in the wild progenitor. The domestication is assumed to have occurred in this constant-size founder population with size $N_1$. Usually, $N_1$ could be much smaller than $N_0$. When the domestication is complete at time $T_e$, it is assumed that the population size changes to $N_2$, which is generally much larger than $N_1$, representing a rapid population expansion of the domesticated species. Let
Let $T_1$ be the length of time when the population size is $N_1$, therefore $T_1 = T_d - T_e$ (Figure 2). Thus, this simple domestication model includes five demographic parameters. Although *O. rufipogon* is a perennial species, we assume that the generation time is one year for this wild progenitor like the annual cultivated rice, because it is able to reproduce one generation each year.

**Simulating Microsatellite Polymorphisms:** Under the population model described above, patterns of microsatellite variation in the two species (one cultivar and its progenitor) were simulated. Since cultivated rice is highly selfing, we employed the coalescent simulation algorithm of NORDBOG and DONNELLY (1997) to construct a random genealogy with diploid sample size $n$. On a simulated genealogy from two subpopulations with the shared ancestral population (Figure 2), random mutations were placed and the changes of repeat numbers at microsatellites were simulated. To approximate the mutation process in microsatellites, we used a generalized stepwise model, under which the change of the number of microsatellite repeats follows a symmetric geometric distribution with a parameter $q$ (GOLDSTEIN et al. 1995; SLATKIN 1995; PRITCHARD et al. 1999). When $q$ is 0, the model is essentially the same as the infinite allele model (IAM) (KIMURA and CROW 1964), while when $q = 1$, the model is identical to the strict stepwise mutation model (SMM) (OHTA and KIMURA 1973). In this study, we considered five models with $q = 1$, 0.9, 0.8, 0.5, and 0, but we mainly show the results for $q = 0.8$ through this article because the results under the five models are very similar (the results for $q = 1$, 0.9, 0.5, and 0 are shown in Sup.)
Figure 1). Our choice of \( q = 0.8 \) is roughly in agreement with estimates from maize (VIGOUROUX et al. 2005) and humans (PRITCHARD et al. 1999).

**Rejection-sampling Algorithm for Estimating Demographic Parameters:** We employed a rejection-sampling algorithm to obtain estimates of the demographic parameters in the above-described model. Since the initial work by TAVARÉ et al. (1997), rejection-sampling algorithms have been extensively used in population genetics (PRITCHARD et al. 1999; WEISS and HAESELER 1998; BEAUMONT et al. 2002). Recently, Wright and his colleagues (WRIGHT et al. 2005) developed a rejection-sampling algorithm for a demographic model that is essentially the same as Figure 2, and applied it to nucleotide polymorphism data in maize and its wild progenitor. We basically followed their algorithm, but with modifications as below. Because evaluating the full likelihood of multiple microsatellites may be computationally infeasible at present to the best of our knowledge, we summarized the data by two statistics, the average heterozygosity in wild (\( H_W \)) and the average heterozygosity in cultivated rice (\( H_C \)), which could be either \( H_I \) or \( H_J \) for the analysis of the indica and japonica subspecies, respectively.

The procedure of our rejection-sampling algorithm to obtain a sample from the joint posterior distribution of the five demographic parameters is described with example parameters. We will change those to check their robustness later.

1. Simulate the five demographic parameters (\( T_d, T_e, N_0, N_1, N_2 \)) and the mutation rate of microsatellites from the prior distributions. Since most domestication events occurred about 5,000 – 10,000 years ago (SMITH 2001), a uniform distribution between 5,000 and
10,000 is adopted for the prior distribution of $T_d$. Once $T_d$ is determined, $T_e$ is assumed to be a random variable between 0 and $T_d$. For $N_0$, the level of nucleotide diversity in *O. rufipogon* is roughly one-third of that in teosintes (CLARK *et al.* 2004; MIYASHITA 2005), so $N_0$ of wild progenitor of rice may be about around 300,000 given that $N_2$ for wild teosinte may be about 900,000-1,000,000 (EYRE-WALKER *et al.* 1998). Therefore, we used a LogNormal distribution with mean $3 \times 10^5$ and SD $0.6 \times 10^5$ for the prior distribution of $N_0$ so that the 95% interval is approximately between $2 \times 10^5$ and $4.4 \times 10^5$. As it may be very difficult to estimate the current effective population size for cultivars, we used a rather diffuse prior distribution for $N_2$, namely, a LogNormal distribution with mean $3 \times 10^5$ and SD $1.5 \times 10^5$ (the 95% interval is approximately between $1 \times 10^5$ and $7 \times 10^5$). Since little prior information is known for $N_1$, a uniform distribution between 0 and 30,000 is assumed for the prior distribution of $N_1$. For the mutation rate ($u$), we used a uniform distribution from 0 to $2 \times 10^4$ when $q = 0.8$.

2. Simulate 46 random independent genealogies with the algorithm for partial selfers (NORDBORG and DONNELLY 1997). The diploid sample sizes of the two species at 46 loci follow those in the observed data (Table 1). The effect of inbreeding is plugged in using estimates of Wright’s $F$-statistic in our data ($F = 0.95$ and 0.65 for *O. sativa* and *O. rufipogon*, respectively).

3. Place random mutations and simulate patterns of microsatellite polymorphisms. To incorporate the variation in mutation rate across loci, the mutation rate at each locus is assumed to follow a Gamma distribution with mean $u$ (determined at Step 1) and SD $0.7 \times u$. With our prior simulations, it was found that this amount of SD roughly explains the
variation in $H$ across loci in the observed data of $O. rufipogon$ (Table 1) for a wide range of $u$.

4. Calculate the average $H_W$ and $H_C$ for the simulated polymorphisms at 46 independent microsatellite loci.

5. Keep the parameters if $|H_W - H_{W,obs}| < \varepsilon \sqrt{H_{W,obs}(1 - H_{W,obs})}$ and $|H_C - H_{C,obs}| < \varepsilon \sqrt{H_{C,obs}(1 - H_{C,obs})}$, otherwise discard the parameters and go to Step 1.

This procedure produces a sample from joint posterior distribution of the parameters conditional on $H_W$ and $H_C$. We applied this procedure to the pairs of $O. sativa$ subsp. indica vs. $O. rufipogon$ and $O. sativa$ subsp. japonica vs. $O. rufipogon$ under the five microsatellite mutation models with $q = 1, 0.9, 0.8, 0.5,$ and 0. For each case, at least 5,000 accepted sets of parameters were collected. We set $\varepsilon = 0.02$ after some preliminary runs of the simulations, which demonstrated that $\varepsilon$ values smaller than 0.02 produced essentially the same results.

**Testing for Selection:** A test of neutrality was performed by focusing on $\Delta$, which is defined by $\Delta = H_C/H_W$. $\Delta$ has been used in studies of maize and wild teosintes (VIGOUROUX et al. 2002, 2004; TENAILLON et al. 2004; WRIGHT et al. 2005). The null distributions of $\Delta$ were usually obtained under certain demographic models with fixed parameters estimated from the data (e.g., VIGOUROUX et al. 2002, 2004, TENAILLON et al. 2004). Because such an analysis might underrepresent the uncertainty in the inference of the demographic parameters (THORNTON and ANDOLFATTO 2006), we used the obtained sample from the joint posterior distribution of parameters (see the previous section) as follows. Let $M_i$ be the $i$th accepted vector of
the parameters \((i = 1, 2, 3, \ldots, B,\) where \(B\) is the number of accepted vectors obtained by the rejection-sampling algorithm). The rejection probabilities of the neutrality were obtained locus by locus, because the null distribution of \(\Delta\) is given conditional on \(H_W\). That is, under our framework, the probability that \(\Delta\) is smaller than the observed value conditional on the observed \(H_{W,i}\) is approximately given by

\[
P(\Delta < \Delta_{\text{obs}} | H_{W,i}) = \frac{1}{B} \sum_{i=1}^{B} P(\Delta < \Delta_{\text{obs}} | H_{W,i}, M_i).
\]

For each \(M_i\), \(P(\Delta < \Delta_{\text{obs}} | H_{W,i}, M_i)\) was obtained by coalescent simulations with 10 replications, therefore the unconditional probability, \(P(\Delta < \Delta_{\text{obs}} | H_{W,i})\), was evaluated from at least 50,000 replications for each microsatellite locus.

**RESULTS**

The gene diversity of the two rice subspecies and \(O. rufipogon\) is summarized in Table 1. At 44 loci of the 46 di-nucleotide microsatellite loci studies, the wild progenitor \(O. rufipogon\) possessed higher diversity \((H)\) than either of the two rice subspecies. The \(indica\) and \(japonica\) rice subspecies have about 70.8% and 62.1%, respectively, of the microsatellite diversity found in their wild progenitor, and the \(indica\) subspecies has higher diversity than the \(japonica\) subspecies at the 37 of the loci. When the \(indica\) and \(japonica\) subspecies are pooled, the cultivars have 77.0% of diversity in comparison to the wild progenitor.

A genetic relationship tree of 70 individuals was constructed using C.S. Chord distances (Figure 1A). In addition to a clear separation of the CWR, most accessions were classified into either the \(indica\) or \(japonica\) group with high bootstrap support. However,
two *japonica* varieties fell into within the *indica* cluster, while five *indica* varieties were included within the *japonica* cluster. To further determine genetic differentiation and population structure among rice groups, we applied the Bayesian clustering program STRUCTURE to analyzing microsatellite data. The highest log likelihood scores were obtained when the number of populations was set at three, and the result is shown in Figure 1B. Most CWR accessions were classified into one group, while the accessions of cultivated rice could not be clearly assigned to either *indica* or *japonica* population. The observation was consistent with the incomplete clustering pattern based on genetic distance (Figure 1A).

Our result of the STRUCTURE analysis seems different from that of CAICEDO *et al.*’s large-scale single nucleotide polymorphism (SNP) data (CAICEDO *et al.* 2007), where *indica* and *japonica* clearly cluster as distinct groups. One of the potential reasons would be that our sample includes some morphologically “intermediate” strains, for example, those whose characteristics are generally consistent with *indica* but some minor characteristics are similar to *japonica*. Another reason could be a lack of the power to classify the two subspecies in our data from microsatellites, at which independent mutations can occasionally create the same allele, thereby obscuring the population differentiation. More importantly, the number of loci used in our study is much smaller than that of CAICEDO *et al.* (2007). Here, it is important to notice that the clear clustering of *indica* and *japonica* in CAICEDO *et al.*’s data (2007) does not necessarily mean that the domestication processes of the two subspecies were completely independent, as CAICEDO *et al.* (2007) concluded (see also our analysis below).
To infer the bottleneck effect on cultivated rice, we used a two-subpopulation model, which approximates a domestication event of a cultivated species from its wild progenitor (Figure 2). We applied this model to *O. rufipogon* vs. *O. sativa* subsp. *indica* and *O. rufipogon* vs. *O. sativa* subsp. *japonica* for separately estimating the above-mentioned five demographic parameters (*N_2*, *N_1*, *N_0*, *T_d*, and *T_e*) plus the mutation rate (µ) using the diversity data for the di-nucleotide microsatellites. Figure 3 shows the obtained posterior distributions. Overall the distributions of *N_0*, *N_2* and µ for the two analyses are very similar (Figure 3). This is because the estimation of these parameters largely depends on the observation in *O. rufipogon*, and may not reflect that in the cultivated rice. The posterior distributions of *N_0* and *N_2* are very similar to the prior distributions, indicating that these parameters are unestimatable. Rather, we can obtain the information on the product of the population size and the mutation rate, which is the most important factor to determine the amount of polymorphism. As shown in Figures 3D and I, the joint distributions are around the line *N_2* × µ ≈ 10. The posterior distributions of *T_d* and *T_e* are also very similar to their prior distributions (not shown). This is partly because the prior distributions of *T_d* and *T_e* were set to be in very narrow ranges (less than 1/30 in units of *N_0*), so that *T_d* and *T_e* seem to have relatively minor effects in our simulations.

Among these five parameters, we are primarily interested in the effective size of the founder population during the bottleneck (*N_1*) as well as its duration (*T_1*). Figures 3D and I show the joint posterior distributions of pairs of *N_1* and *T_1*. For *O. rufipogon* vs. *O. sativa* subsp. *indica*, *N_1* and *T_1* have a distribution along the line *N_1* = 1.5 × *T_1*, while the *O. rufipogon* vs. *O. sativa* subsp. *japonica* analysis produced the posterior distribution of *N_1* and *T_1* along the line *N_1* = 0.9 × *T_1*. The results indicate that the *japonica* subspecies
experienced a more severe bottleneck than the *indica* subspecies, in agreement with the lower genetic diversity in the *japonica* subspecies. We further estimated the bottleneck severity \( k \), which is the ratio of the size of founder population during the domestication \( (N_1) \) to the duration of the bottleneck \( (T_1) \). A smaller \( k \) was observed for the *japonica* subspecies \( (k = 0.9) \) than that for the *indica* subspecies \( (k = 1.5) \), suggesting the *japonica* subspecies has suffered a severer bottleneck than the *indica* subspecies.

Artificial selection could create a local reduction in polymorphism. To evaluate this effect, we searched for outlier loci with lower genetic variation than the average in cultivated rice. We tested the neutrality for each locus focusing on \( \Delta \), the ratio of the diversity in the cultivar to that of the wild progenitor. Here, the estimated demographic models were used as null models for testing the neutral null hypothesis (see *Materials and Methods*). The \( p \)-value for each locus was obtained from coalescent simulations and the result is summarized in Table 1. There are several loci with \( p \)-values < 0.05, that is, the null model cannot explain the observed reduction of \( H \) in the cultivated rice at the 5% level. In total, three (RM47, RM167, RM247) in the *indica* subspecies and five (RM212, RM211, RM26, RM47, RM258) microsatellites in the *japonica* subspecies have \( p < 0.05 \). RM47 is found to be an outlier in both subspecies, while the other microsatellites are outlier only in the *indica* or *japonica* subspecies. The number of loci with \( p < 0.05 \) is slightly higher than the expected numbers \((0.05 \times 46 = 2.3)\), but not significant.

Although only the results for \( q = 0.8 \) are given here, almost identical results were obtained for the other four mutation models \((q = 1, 0.9, 0.5 \text{ and } 0)\). As shown in Sup. Figure 1, the posterior distributions of \( N_0, N_1, N_2 \) and \( T_1 \) for the five different values of \( q \) are vary similar, although the mutation rate depends on the mutation model. As a
consequence, it is not surprising that the results of the neutrality test ($p$-values for each locus) are in excellent agreement with one another as shown in Sup. Figures 2A and B. The $p$-values for $q = 1, 0.9, 0.5$ and 0 (plotted on the vertical axis) in almost all loci are very similar to those for $q = 0.8$ (plotted on the vertical axis).

Furthermore, such robustness is also observed when the prior distributions are changed. We repeated the same rejection-sampling procedure with wider posterior distributions of $N_0$ and $N_2$. We assumed that both distributions follow a LogNormal distribution with mean $3 \times 10^5$ and SD $2 \times 10^5$ (the 95% interval is approximately between $0.75 \times 10^5$ and $8 \times 10^5$). This analysis gave similar results for the estimates of the population mutation rates and the severity of bottleneck (data not shown) and also the results of the neutrality tests (Sup. Figures 2C and D). At almost all loci, the $p$-values for the five values of $q$ (plotted on the vertical axis) with the wider priors are very similar to those for $q = 0.8$ with the narrower priors in the Materials and Methods (i.e., the $p$-values in Table 1, plotted on the vertical axis in Sup. Figures 2C and D). It indicates that our analysis is quite robust to mutational models and the choice of prior distributions of the parameters.

One drawback in our analysis incorporating the uncertainty in the inference of demography is that it might unnecessarily lose the power to detect selection because the real population experienced only one demographic history. To check this effect, we repeated the selection test in which demography was fixed to a single parameter set. The parameter set were chosen based on the posterior distributions (Figure 3) such that it explains the genome-wide reduction in genetic variation in the two subspecies. Namely, the ancestral population size is $N_0 = 300,000$. At time $T_d = 7,500$, two independent founder populations were formed, and their sizes increased to 300,000 at time $T_e = 5,000$. 
We assumed that the sizes of the two founder populations are 3,750 and 2,250 for the *indica* and *japonica* subspecies, respectively. We found that this setting produced almost identical *P*-values for all loci (not shown). Similar results were also obtained with other possible sets of parameters.

A more interesting observation in our results is that there is a strong positive correlation between the *p*-values for the two subspecies with a correlation coefficient of 0.59 (Figure 4A). To test whether this observed correlation coefficient can be explained under a neutral model of two independent domestication events, additional coalescent simulations were performed. We used the fixed setting of the demographic parameters as described above. In each replication, microsatellite polymorphisms at 46 independent loci were simulated (see *Materials and Methods*) and the correlation coefficient for the simulated data was computed. The distribution of the correlation coefficient for the simulated data suggests that the observed correlation coefficient cannot be explained under this simple independent domestication model (*p* < 0.001, Figure 4B). Simulations with other parameter sets gave very similar results (not shown).

To demonstrate that a positive correlation of the *p*-values is indeed created when the bottleneck population is shared, sets of similar simulations were further performed using the model illustrated in Figure 5A. The model considers a single bottleneck population with size \( N_{DMC} \) between time \( T_d \) and \( T_{DMC} \). At time \( T_{DMC} \), the *indica* and *japonica* subspecies were split into the *indica* and *japonica* populations (subspecies) with sizes \( N_{1i} \) and \( N_{1j} \), and then the sizes of the two populations change to \( N_{2i} \) and \( N_{2j} \) at time \( T_{ei} \) and \( T_{ej} \), respectively. As well as the simulations for Figure 4, we assumed \( N_0 = N_{2i} = N_{2j} = 300,000 \) and \( T_d = 7,500 \). Here, we assume \( N_{DMC} = N_{1i} = N_{1j} = 2,250 \) such that \( T_{ei} \) and \( T_{ej} \)
simply reflect the severities of bottleneck in the two populations. Following our estimates from Figure 3, we set $T_{ei} = 6,500$ and $T_{e1} = 5,000$. Then, we used various values of $T_{DMC}$ between 6,500 and 7,500, which represents the proportion of the bottleneck effect shared by the two species. Figure 5B shows the distributions of the correlation coefficient for $T_{DMC} = 6,500, 7,000, 7,500$. When $T_{DMC} = T_d = 7,500$ (i.e., there is no bottleneck sharing), the correlation coefficient, $r$, distributes symmetrically around 0, and the distribution is almost identical to that in Figure 4B. As the duration of bottleneck sharing increases (i.e., $T_{DMC}$ decreases), the distribution moves towards positive values as expected. When the entire bottleneck process that the indica subspecies experienced was shared with the japonica subspecies ($T_{DMC} = T_{ei} = 6,500$), the observed high correlation coefficient ($r = 0.59$) is within the 95% interval, indicating a substantial amount of sharing of genetic information between the two subspecies. The trend of the negative correlation between $r$ and $T_{DMC}$ was confirmed by simulations with a wide range of parameters (not shown).

**DISCUSSION**

Maize has long been an excellent model for studying the domestication of cultivated plants (DOEBLEY 2004; TENAILLON et al. 2004; WRIGHT et al. 2005). It represents a simple domestication model of “one wild progenitor – one crop group”, but the situations for the majority of cultivated plants may be more complicated, including the rice domestication that resulted in two subspecies, indica and japonica, from the wild taxon, O. rufipogon. The finding that the japonica subspecies has experienced a more severe loss of genetic variation during domestication than the indica subspecies is in
agreement with published empirical data, which demonstrated that the *indica* subspecies is more genetically diverse than the *japonica* subspecies (LU et al. 2002; SECOND 1982; GARRIS et al. 2005). Our results suggest a smaller founder wild population of the *japonica* subspecies than that of the *indica* subspecies and/or a longer bottleneck (i.e, $T_1$ in Figure 3), assuming the effect of selection reducing polymorphisms is much smaller than the founder effect (WRIGHT et al. 2005). In comparison, a larger estimated $k$ in maize ($k = 2.45$) (WRIGHT et al. 2005) than that in rice ($k = 0.9$ for the *japonica* subspecies and $k = 1.5$ for the *indica* subspecies) indicates that different crops might have experienced a variable severity of domestication bottleneck during their different population history. The finding seems consistent with a less serious relative loss of genetic diversity in maize ($\Delta = 0.88$) (VIGOUROUX et al. 2005) than that in rice ($\Delta = 0.77$ when the *indica* and *japonica* subspecies are pooled) based on the estimate of dinucleotide microsatellite diversity.

The two mechanisms, demography and selection, that remove genetic variation during domestication differ in that the effects of selection should be limited to the surrounding region of the target genes of selection while demographic history affects genetic variation in a similar way throughout the whole genome. Therefore, it is predicted that microsatellites closely linked to loci that were subject to artificial selection will have low diversity in cultivated species (WANG et al. 1999; INNAN and KIM 2004). The results of our neutrality test applied to each locus found three loci with $p < 0.05$ in the *indica* and five in *japonica* subspecies. These numbers are not significantly higher than those expected ones under neutrality, suggesting probably only a weak effect of selection on the genome as a whole. This result seems in agreement with the findings in maize, in
which only 2 to 4% of 774 genes suggest the effects of artificial selection (WRIGHT et al. 2005). However, the power to detect signatures of artificial selection may be limited if selective sweeps are from standing variation (INNAN and KIM 2004).

The variation in $\Delta$ across microsatellite loci also relates to the unsettled controversy about origins of Asian cultivated rice. The hypothesis of independent origins of Asian cultivated rice CWR that the indica and japonica subspecies originated from different wild populations of their progenitor is supported by insertion patterns of SINEs (YAMANAKA et al. 2003; YAMANAKA et al. 2004), LTR-retrotransposons (VITTE et al. 2004) and MITEs (ZHU and GE 2005), as well as polymorphisms of intron sequences (ZHU and GE 2005). These data indicated that the two subspecies shared more polymorphisms with different accessions of CWR than with each other, making this hypothesis become dominant. Recently, phylogeographic analysis of three distinct gene regions further supports that O. sativa was domesticated from CWR at least twice, in at least two different geographic regions in eastern Asia (LONDO et al. 2006). This argument is not conclusive because these differences between the indica and japonica subspecies may be explained by ancestral polymorphism in the CWR so that we cannot exclude a monophyletic origin and subsequent development of the two subspecies occurred after domestication (CHANG 1976; OKA 1988; OKA and MORISHIMA 1982; LU et al. 2002). As shown in Figure 4, the significant correlation between the two subspecies disagrees with the independent hypothesis ($p < 0.001$), suggesting that the model of completely independent domestication of the indica and japonica subspecies may be rejected, and that they at least partially shared their ancestral populations. Our simulation results shown in Figure 5 demonstrated that there should be quite an amount
of sharing of genetic information, which could be explained by the bottleneck sharing and/or gene flow between the *indica* and *japonica* subspecies.

Strictly speaking, however, the argument holds under neutrality. Artificial selection could potentially create a positive correlation. It is worthy of noting that some traits are frequently observed in cultivated crops, such as grain shattering, seed dormancy, synchronization of seed maturation, decrease in tiller number and branches, and increase in inflorescence and seed sizes (HARLAN 1975; HANCOCK 2004; LI *et al.* 2006). These loci responsible for such phenotypes may be likely targets of artificial selection (PATERSON *et al.* 1995). SWEENEY *et al.* (2007) recently showed strong evidence that a selected allele has originated in the *japonica* subspecies and then spread to the *japonica* varieties. It is suggested that the domestication process is still underway and the two subspecies could share genetic variation and selection. Thus, if the two subspecies share selective pressure in the same direction, even under the independent domestication scenario, we could observe a positive correlation. To exclude this effect, we reanalyzed the correlation in the *P*-values by removing loci with *P*<0.05 for both of the two subspecies. The correlation coefficient was 0.57, still highly significant (*p* < 0.001).

Furthermore, to be very conservative, we removed loci with *P*<0.2 for both and found the significance still remained even though the correlation coefficient was reduced (*r*=0.43, *p*=0.004). These results suggest that the effect of artificial selection on the observed positive correlation should be small especially when our data are from non-genic regions (*i.e.*, microsatellites).

In addition, selection could work in only one of the two subspecies, which should have contributed to the differences between the *indica* and *japonica* subspecies. Asian
cultivated rice is widely planted throughout the world under extremely diversified socioeconomic, cultural and geographic conditions. The deep-seated human food preference under divergently cultural societies put cultivars of the two subspecies under extensively strong artificial selection (GAO 2003). This scenario might well explain one of the investigated loci. At RM258, for example, \( \Lambda \) is quite low in the *japonica* subspecies \( (p = 0.038) \) but high for the *indica* subspecies \( (p = 0.7131) \). This microsatellite is closely linked with \( Rf-1 \) (only about 29 bp downstream) on Chromosome 10, a gene controlling cytoplasmic male sterility in rice (SHINJYO 1975; YU et al. 1995; AKAGI et al. 1995; WANG et al. 2006). Most *indica* varieties have dominant allele \( Rf-1 \), but the typical *japonica* subspecies carries the nonrestoring allele \( rf-1 \) (SHINJYO 1975; ZHU 2000). Therefore, our observation at RM258 may be well explained by the *japonica*-specific selection at \( Rf-1 \).

Thus, this article presents an analysis supporting the non-independent domestication more than the popularly recognized independent hypothesis of Asian cultivated rice. Our analysis rejected a strict independent domestication model. These results can be understood with at least two factors. (i) The founder populations of *indica* and *japonica* were at least partly shared. (ii) There was substantial gene flow between *indica* and *japonica* after domestication. Our results are consistent with the recent analysis of SNP data. CAICEDO et al. (2007) demonstrated that demographic models with recent gene flow explain their large-scale SNP data better than those without gene flow, and SWEENEY et al. (2007) provided a direct evidence for recent gene flow. At this moment, it may be difficult to distinguish these two factors or to elucidate their relative
contributions. More SNP data from the rice genomes would provide further opportunities to better understand origins and domestication of Asian cultivated rice.

We thank D. Charlesworth and two anonymous reviewers for valuable comments and suggestions, which improved this work substantially. We wish to thank Dao-yuan Li of the Institute of Crop Germplasm Resources, Guangxi Academy of Agricultural Sciences; Da-jian Pan of Guangdong Rice Research Institute, Guangdong Academy of Agricultural Sciences; Zong-en Qiu and Han-hua Pang of the Institute of Crop Germplasm Resources, Chinese Academy of Agricultural Sciences, and Yi-ping Wang of the China Rice Research Institute for providing wild and cultivated rice materials. This research was funded by the International Foundation for Sciences (IFS) (C/2738-1, 2), the 1999 Vavilov-Frankel Fellowship for 1999 from International Plant Genetic Resources Institute (IPGRI), the China National Science Foundation (39800013), Hundreds Talents Program of Chinese Academy of Sciences, and a grant from Chinese Department of Science and Technology (973 Program 2007CB815701) to L.-Z. GAO. H. Innan was supported by grants from the University of Texas at Houston, the Graduate University for Advanced Studies, JSPS (Japan Society for the Promotion of Science) and NSF.

LITERATURE CITED


INNAN, H. and Y. KIM, 2004 Pattern of polymorphism after strong artificial selection

KATO, S., H. KOSAKA and S. HARA, 1928 On the affinity of rice varieties as shown by the fertility of rice plants. Central Agricultural Institute of Kyushu Imperial University 2: 241–276.


MIYASHITA, N. T., 2005 DNA variation in the metallothionein genes in wild rice


TAKEZAKI, N. and M. NEI, 1996 Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144: 389–399.

TAVARE, S., D. J. BALDING, R. C. GRIFFITHS and P. DONNELLY, 1997 Inferring
coalescence times from DNA sequence data. Genetics 145: 505–518.


WANG, R. L., A. STEC, J. HEY, L. LUKENS and J. F. DOEBLEY, 1999 The limits of

Cytoplasmic male sterility of rice with Boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. Plant Cell **18**: 676–687.


YAMANAKA, S., I. NAKAMURA, K. N. WATANABE and Y.I. SATO, 2004
Identification of SNPs in the waxy gene among glutinous rice cultivars and their evolutionary significance during the domestication process of rice. Theor Appl Genet **108**: 1200–1204.


Association of morphological and RFLP markers in rice (*Oryza sativa* L.).

*Genome* **38**: 566–574.


### TABLE 1

Genetic diversity estimates and sample size of *O. sativa* subsp. *indica*, *O. sativa* subsp. *japonica* and *O. rufipogon* for each of the forty-six di-nucleotide microsatellite loci studied

<table>
<thead>
<tr>
<th>Locus</th>
<th><em>O. sativa</em> subsp. <em>indica</em></th>
<th><em>O. sativa</em> subsp. <em>japonica</em></th>
<th><em>O. rufipogon</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>n</strong></td>
<td><strong>A</strong></td>
<td><strong>H</strong></td>
</tr>
<tr>
<td>RM212</td>
<td>16</td>
<td>3</td>
<td>0.575</td>
</tr>
<tr>
<td>OSR23</td>
<td>16</td>
<td>2</td>
<td>0.458333</td>
</tr>
<tr>
<td>RM200</td>
<td>16</td>
<td>11</td>
<td>0.916667</td>
</tr>
<tr>
<td>RM211</td>
<td>16</td>
<td>3</td>
<td>0.575</td>
</tr>
<tr>
<td>OSR11</td>
<td>16</td>
<td>4</td>
<td>0.658333</td>
</tr>
<tr>
<td>RM263</td>
<td>15</td>
<td>9</td>
<td>0.933333</td>
</tr>
<tr>
<td>OSR8</td>
<td>16</td>
<td>5</td>
<td>0.75</td>
</tr>
<tr>
<td>RM233A</td>
<td>15</td>
<td>5</td>
<td>0.809524</td>
</tr>
<tr>
<td>RM55</td>
<td>16</td>
<td>4</td>
<td>0.6</td>
</tr>
<tr>
<td>RM232</td>
<td>14</td>
<td>5</td>
<td>0.769231</td>
</tr>
<tr>
<td>RM231</td>
<td>14</td>
<td>7</td>
<td>0.879121</td>
</tr>
<tr>
<td>RM255</td>
<td>14</td>
<td>4</td>
<td>0.659341</td>
</tr>
<tr>
<td>RM252</td>
<td>16</td>
<td>6</td>
<td>0.825</td>
</tr>
<tr>
<td>RM261</td>
<td>16</td>
<td>7</td>
<td>0.75</td>
</tr>
<tr>
<td>RM241</td>
<td>16</td>
<td>7</td>
<td>0.841667</td>
</tr>
<tr>
<td>OSR15</td>
<td>16</td>
<td>2</td>
<td>0.458333</td>
</tr>
<tr>
<td>RM233B</td>
<td>16</td>
<td>4</td>
<td>0.775</td>
</tr>
<tr>
<td>RM26</td>
<td>15</td>
<td>3</td>
<td>0.647619</td>
</tr>
<tr>
<td>RM249</td>
<td>14</td>
<td>12</td>
<td>0.950549</td>
</tr>
<tr>
<td>RM164</td>
<td>14</td>
<td>7</td>
<td>0.89011</td>
</tr>
<tr>
<td>RM225</td>
<td>16</td>
<td>8</td>
<td>0.891667</td>
</tr>
<tr>
<td>OSR21</td>
<td>15</td>
<td>2</td>
<td>0.47619</td>
</tr>
<tr>
<td>OSR19</td>
<td>16</td>
<td>4</td>
<td>0.691667</td>
</tr>
<tr>
<td>OSR25</td>
<td>16</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RM253</td>
<td>16</td>
<td>7</td>
<td>0.816667</td>
</tr>
<tr>
<td>RM234</td>
<td>15</td>
<td>4</td>
<td>0.619048</td>
</tr>
<tr>
<td>RM47</td>
<td>16</td>
<td>2</td>
<td>0.125</td>
</tr>
<tr>
<td>RM248</td>
<td>14</td>
<td>8</td>
<td>0.895604</td>
</tr>
<tr>
<td>OSR22</td>
<td>16</td>
<td>3</td>
<td>0.675</td>
</tr>
<tr>
<td>OSR07</td>
<td>16</td>
<td>4</td>
<td>0.691667</td>
</tr>
<tr>
<td>RM25</td>
<td>16</td>
<td>7</td>
<td>0.7875</td>
</tr>
</tbody>
</table>

33
<table>
<thead>
<tr>
<th>Sample</th>
<th>n</th>
<th>A</th>
<th>H</th>
<th>P</th>
<th>Expected homogeneity</th>
<th>Expected heterozygosity</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM223</td>
<td>16</td>
<td>8</td>
<td>0.891667</td>
<td>0.9464</td>
<td>27 8</td>
<td>0.786325 0.8571</td>
</tr>
<tr>
<td>RM215</td>
<td>16</td>
<td>6</td>
<td>0.675</td>
<td>0.2811</td>
<td>28 7</td>
<td>0.809524 0.9535</td>
</tr>
<tr>
<td>RM245</td>
<td>14</td>
<td>5</td>
<td>0.752747</td>
<td>0.6522</td>
<td>27 3</td>
<td>0.145299 0.0585</td>
</tr>
<tr>
<td>RM242</td>
<td>15</td>
<td>4</td>
<td>0.761905</td>
<td>0.9561</td>
<td>28 5</td>
<td>0.637566 0.8937</td>
</tr>
<tr>
<td>RM244</td>
<td>16</td>
<td>4</td>
<td>0.591667</td>
<td>0.4223</td>
<td>27 4</td>
<td>0.430199 0.3579</td>
</tr>
<tr>
<td>RM222</td>
<td>15</td>
<td>5</td>
<td>0.790476</td>
<td>0.6811</td>
<td>26 13</td>
<td>0.910769 0.9998</td>
</tr>
<tr>
<td>RM258</td>
<td>16</td>
<td>6</td>
<td>0.833333</td>
<td>0.7131</td>
<td>28 4</td>
<td>0.267196 0.038</td>
</tr>
<tr>
<td>RM228</td>
<td>14</td>
<td>5</td>
<td>0.824176</td>
<td>0.6481</td>
<td>27 8</td>
<td>0.680912 0.4942</td>
</tr>
<tr>
<td>RM202</td>
<td>16</td>
<td>7</td>
<td>0.879167</td>
<td>0.8201</td>
<td>28 7</td>
<td>0.780423 0.7226</td>
</tr>
<tr>
<td>RM206</td>
<td>13</td>
<td>9</td>
<td>0.935897</td>
<td>0.9894</td>
<td>23 10</td>
<td>0.873518 0.9838</td>
</tr>
<tr>
<td>RM167</td>
<td>16</td>
<td>4</td>
<td>0.516667</td>
<td>0.0433</td>
<td>27 8</td>
<td>0.783476 0.8162</td>
</tr>
<tr>
<td>RM224</td>
<td>16</td>
<td>6</td>
<td>0.85</td>
<td>0.9904</td>
<td>23 4</td>
<td>0.644269 0.8075</td>
</tr>
<tr>
<td>RM235</td>
<td>16</td>
<td>10</td>
<td>0.9</td>
<td>0.9978</td>
<td>25 7</td>
<td>0.806667 0.9869</td>
</tr>
<tr>
<td>OSR20</td>
<td>16</td>
<td>6</td>
<td>0.816667</td>
<td>0.5943</td>
<td>26 8</td>
<td>0.874615 0.9829</td>
</tr>
<tr>
<td>RM247</td>
<td>16</td>
<td>4</td>
<td>0.575</td>
<td>0.0489</td>
<td>25 9</td>
<td>0.82 0.8591</td>
</tr>
</tbody>
</table>

\(^a\) Sample size;  
\(^b\) Observed numbers of alleles;  
\(^c\) Heterozygosity;  
\(^d\) Probability values for the test of neutrality.
Figure Legends:

**FIGURE 1.** — (A) Unrooted neighbor-joining tree based on C.S. Chord distance using 46 nuclear microsatellites. The admixed varieties are indicated with an asterisk. Bootstrap support values are shown; (B) Estimated population structure for 70 accessions of *O. sativa* and *O. rufipogon* from 46 microsatellite loci. Horizontal bars along vertical axis represent each accession. For all the analyzed accessions, the proportion of ancestry under $K = 3$ clusters that be attributed each cluster is given by the length of each colored segment in a bar.

**FIGURE 2.** — The demographic model and parameters used in this study. A typical genealogy of samples from the progenitor and cultivar populations (species) is also illustrated. The underlying model was used twice: CWR vs *O. sativa* subsp. *indica* and CWR vs *O. sativa* subsp. *japonica*. See text for details.

**FIGURE 3.** — The posterior distributions of the demographic and mutation parameters when $q = 0.8$. The results for the CWR vs *indica* analysis are shown in panels A-E and those for the CWR vs *japonica* analysis are in F-J. In A-C and F-H, the posterior distributions are shown by bars, while the priors are by lines. The prior distributions are as described in Materials and Methods. Panels D, E, I and J show two dimensional posterior distributions, in which the gray scale represents the density as indicated in the right side of each panel. The results for the other four $q$ values are shown in Sup. Figure 1.

**FIGURE 4.** — (A) Joint distribution of the $p$-values for the CWR vs *indica* analysis and CWR vs *japonica* analysis. The $p$-values are from Table 1 (i.e., $q = 0.8$ with the priors in the Materials and Methods). The $p$-values from the two analysis are highly correlated ($r = 0.59$); (B) The null distribution of the correlation coefficient ($r$) under a neutral model of independent domestication. See text for details.
**FIGURE 5.** — Effect of bottleneck sharing (non-independent domestication) on the correlation coefficient ($r$) between the p-values for the CWR vs *indica* analysis and CWR vs *japonica* analysis. (A) Illustration of the model of bottleneck sharing used in our simulation. (B) Distribution of the correlation coefficient ($r$) when $T_{DMC} = 7,500$, $7,000$ and $6,500$. See text for the other parameters.

**Supplementary Figure 1.** — See the legend of FIGURE 3.

**Supplementary Figure 2.** — (A-B) Effect of the choice of the microsatellite mutation models on the p-values for the neutrality test in the CWR vs *indica* analysis (A) and CWR vs *japonica* analysis (B). The horizontal axis represents the p-values for $q = 0.8$, which are shown in Table 1. The vertical axis represents the p-values for the other four mutation models ($q = 1$, $0.9$, $0.5$ and $0$) with the same priors described in *Materials and Methods*. (C-D) The p-values for the five mutation models ($q = 1$, $0.9$, $0.8$, $0.5$ and $0$) with wider priors, which are compared with those in Table 1 (*i.e.*, $q = 0.8$ with the priors in the *Materials and Methods*). C and D show the results for the CWR vs *indica* analysis and CWR vs *japonica* analysis, respectively.
$N_0$

Progenitor

$N_1$

$N_2$

Cultivar

$T_e$

$T_d$

$T_1$
Indica Japonica

SMM ($q = 0.8$)

A

Indica

Current population size ($N_0$)

Density

0 0.1 0.2

0 50,000 100,000

B

Japonica

Ancestral population size ($N_2$)

Density

0 0.1 0.2

0 50,000 100,000

C

Mutation rate ($\mu \times 10,000$)

Density

0 0.1 0.2

0 1

D

Ancestral population size ($N_2$)

Mutation rate ($\mu \times 10,000$)

Density

0 0.25 0.5 0.75 1

0 5000 10000 15000 20000 25000

E

Bottleneck population size ($N_1$)

Duration of bottleneck ($T_1$)

Density

0 0 5000 10000 15000 20000

0 0 5000 10000 15000 20000
A

P-values for Japonica

P-values for Indica

B

Probability Density

Observed
A

Progenitor

japonica

indica

B

Correlation coefficient

Probability Density

$T_{DMC} = 7,500$

$T_{DMC} = 7,000$

$T_{DMC} = 6,500$