doi:10.1111/j.1420-9101.2009.01698.x

SNP deserts of Asian cultivated rice: genomic regions under domestication

L. WANG,* L. HAO,* X. LI,† S. HU,* S. GE‡ & J. YU*

*CAS Key Laboratory of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China
†State Key Laboratory of Plant Physiology and Biochemistry, Department of Plant Genetics and Breeding, China Agricultural University, Beijing, China
‡State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China

Keywords:

domestication;
genetic variation;
indica;
japonica;
Oryza sativa.

Abstract

When performing a genome-wide comparison between *indica* (93-11) and *japonica* (Nipponbare), we find 8% of the genome, which have an extremely low SNP rate (< 1 SNP/kb). Inside these 'SNP deserts', experimentally confirmed genes show increased K_a/K_s that indicate adaptive selection. To further elucidate this connection, we survey the level and pattern of genetic variation in both cultivated and wild rice groups, using 155 noncoding regions located within SNP deserts. The results suggest that cultivated rice has greatly reduced genetic variation within SNP deserts as compared to either the nondesert or corresponding genomic regions in wild rice. Consistent with this reduction in genetic variation, we find a biased distribution of derived allele frequency in the cultivated group, indicative of positive selection. Furthermore, over half of the confirmed, domestication-related genes are found within SNP deserts, also suggesting that SNP deserts are strongly related to domestication, and might be the key sites in the process of domestication.

Introduction

The domestication of many major crop species, such as rice and wheat, began approximately 10 000–12 000 years ago. These domesticated species are characterized by a reduction in genetic diversity compared to their wild ancestors (Eyre-Walker *et al.*, 1998; Tenaillon *et al.*, 2004). Basically, this reduction is the result of two processes; one is the population bottleneck at the beginning of domestication, and the other is the selection by humans for favourable agronomic traits during the process of domestication (Tenaillon *et al.*, 2004; Zeder *et al.*, 2006). It has been reported that domesticated species retain two-thirds of the diversity found in the wild counterparts (Buckler *et al.*, 2001).

Although many investigations have compared nucleotide diversity levels between crops and their wild relatives (Tenaillon *et al.*, 2001; Zhu *et al.*, 2003; Hamblin *et al.*, 2004; Liu & Burke, 2006), relatively few studies have been conducted in rice, and different studies have given rise to different results for the level of nucleotide diversity retained by cultivated rice, with estimated values ranging from 10% to 71%. In early studies, isozyme data suggested that domesticated rice retained ~71% of the diversity found in its wild relatives (Oka, 1988). Recently, Zhu *et al.* (2007) surveyed sequence variation over 10 unlinked nuclear loci in 60 cultivated and wild rice accessions and found that both *indica* (*Oryza sativa* L. ssp. *indica*) and *japonica* (*Oryza sativa* L. ssp. *japonica*) rice retained only 10–20% of the silent-site variation found in the wild. Caicedo *et al.* (2007) re-sequenced 111 randomly selected gene fragments in a diverse panel composed of 72 accessions of cultivated rice and 21 accessions of wild rice, and revealed that the cultivated held ~62% of the silent-site diversity found in the wild. In this study, we survey for genetic variation over 3 Mb noncoding regions (150 amplicons) in 43 diverse accessions of cultivated rice as well as 20 accessions of wild rice. Our estimation provides more reliable data and powerful insights than previous studies, as all sequences used to compute the genetic diversity are noncoding regions that remain relatively neutral in the evolutionary process.

In addition to comparative studies on genetic diversity of domesticates with their wild relatives, scientists have also become increasingly interested in genes involved in domestication, especially for cereal crops (Hirano *et al.*,

Correspondence: Jun Yu, CAS Key Laboratory of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China.
Tel.: +86 10 8299 5357; fax: +86 10 8299 5373; e-mail: junyu@big.ac.cn

1998; Clark *et al.*, 2004; Palaisa *et al.*, 2004; Doebley *et al.*, 2006; Konishi *et al.*, 2006; Olsen *et al.*, 2006; Sweeney *et al.*, 2007). However, despite these great efforts expended over the past several decades, relatively few domestication-related genes have been identified. For instance, only two maize genes were identified and mapped – *tb1* that controls inflorescence structure and *tga1* that regulates seed casing (Wang *et al.*, 2005). Other examples include the tomato (*Solanum lycopersicum*) *fw2.2* gene that was identified for fruit weight control, and the wheat (*Triticum aestivum* L) *Q* gene for monitoring inflorescence structure (Cong & Tanksley, 2006; Simons *et al.*, 2006). More domestication-related genes have been identified in rice than in other crops, but they are still limited in number: *q-SH1* and *Sh4* (*SHA1*) control abscission layer formation and seed shattering (Konishi *et al.*, 2006; Li *et al.*, 2006; Lin *et al.*, 2007), *Rc* gene governs the pericarp colour of seeds (Bautista *et al.*, 2006; Konishi *et al.*, 2006; Sweeney *et al.*, 2006, 2007), *Hd1*, *Hd6* and *Ehd1* influence rice flowering time (Yano *et al.*, 2000; Takahashi *et al.*, 2001; Doi *et al.*, 2004), the *Waxy* locus encodes a starch synthase (Olsen *et al.*, 2006), *GS3* regulates grain size and shape (Fan *et al.*, 2006) and *Sd-1* related to rice plant height (Foster & Rutger, 1978). In 2008, two other domestication-related genes were identified in rice: *qsW5* is involved in the determination of rice grain width and *PROG1* controls plant architecture (Jin *et al.*, 2008; Shomura *et al.*, 2008; Tan *et al.*, 2008). Although our goal here is not to identify any new domestication-related genes, we define 8% of the genome as 'SNP deserts' where domestication-related genes are believed to be enriched as these regions harbour most of the reported domestication-related genes. Our work will facilitate new discoveries of domestication-related genes by narrowing the scope of the search to focus on the candidate regions.

In this study, we perform a whole genome alignment between Beijing *indica* (93-11) and Syngenta *japonica* (Nipponbare) and define highly conservative regions as 'SNP deserts' that are composed of 12.8% of all aligned sequences and 8% of the *indica* genome. Although a prior study has performed a similar procedure with the same two subspecies, we adopt a more reliable alignment method and used a more accurate genome map (see Methods) (Feltus *et al.*, 2004). In addition, we survey the sequence variation in the three representative SNP deserts with 43 and 20 diverse accessions of cultivated rice and wild rice respectively.

Results

Existence of SNP deserts between *indica* (93-11) and *japonica* (Nipponbare)

We provided strong evidence for the existence of SNP deserts within *indica* (93-11) and *japonica* (Nipponbare). There is a clear bimodality in the SNP rate distribution, with a major mode at about 6.5 SNP/kb and a minor mode at about zero (Fig. 1). A similar effect had been previously reported (Nasu *et al.*, 2002; Feltus *et al.*, 2004). We used a 20-kb bin size for most of our analyses, comparable to the size of an average gene island (Yu *et al.*, 2005), although the bimodality was also observed in other bin sizes (Fig. S1). The effect did become weaker as the bin size increases, mostly due to progressive inclusion of intergenic regions, which were notoriously variable in plants (Innan & Kim, 2004). One could demonstrate that it was not an artefact of the difference in SNP rates between genic and intergenic regions (Fig. S2), and that the indel rates were correlated with SNP rates (Fig. S3).

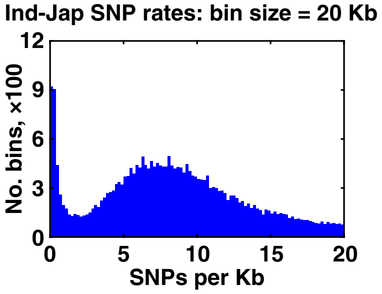

Fig. 1 SNP rate distribution on a 20 kb sliding window. The minor mode is centred at about 0 SNP/kb, and the major mode centred at about 6.5 SNP/kb. Minor mode indicates the existence of SNP deserts.

We define SNP deserts as regions with SNP rates lower than 1 SNP/kb. Many of these so-called SNP deserts are much larger than 20 kb, and some are up to a megabase in size. To compute a typical size, allowing for the highly variable intergenic regions, we defined an SNP desert as a region where 80%, 90% or 100% of the SNP rates in a 20-kb window lay below 1 SNP/kb, and the N50 sizes (above which half of the total sequence length could be found) were 201, 111 or 80 kb respectively. When SNP deserts are defined with the most stringent criteria, i.e. where 100% of the 20-kb windows have SNP rates less than 1 SNP/kb, we identified a total of 596 (33 Mb) regions within rice genome as SNP deserts (Table S1).

To see how often SNP deserts could be observed under neutral models or other complex model conditions, where recombination and/or population growth are considered, we performed a model-based simulation. The results indicated that all empirical *P*-values of observing SNP deserts under these models were below a significance level of 0.05 (Fig. S4).

Genes in SNP deserts have higher K_a/K_s ratio

If a significant fraction of SNP deserts are related to domestication, domesticates should exhibit an increased K_a/K_s ratio, where K_a and K_s represent nonsynonymous

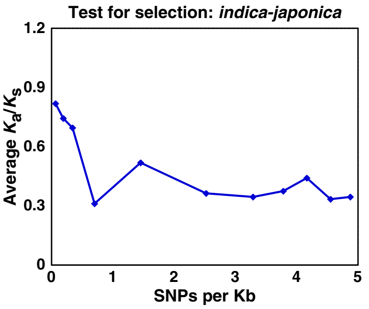

Fig. 2 Average K_a/K_s over groups of genes binned according to SNP rate in a 20-kb window. Each point represents 400 genes.

and synonymous substitutions per available site. On a per gene basis, given the paucity of coding SNP, K_a/K_s values are often undefined, zero or infinity. We addressed this problem by bringing all the genes into calculation; genes were grouped according to SNP rate and then concatenated to yield a single K_a/K_s. Results showed a substantial increase in K_a/K_s at low SNP rates, which was consistent with SNP deserts (Fig. 2).

There was a striking difference for genes in and out of SNP deserts, when classified by SNP rates computed in the 20-kb window centred about each gene (Table 1). Ninety-one per cent of genes in an SNP desert with rates of < 0.5 SNP/kb had an undefined K_a/K_s, but of those that were defined, there were 68 genes with $K_a/K_s > 1$ and 37 genes with $K_a/K_s \leq 1$, with a ratio of 1.84 (68/37, not K_a/K_s). In SNP deserts with rates above 2 SNP/kb, this ratio was 0.40. When we computed significance with the chi-squared test on the number of genes, we found $\chi^2_1 = 63.25$, $P < 0.0001$ (this is the Yates chi-squared value corrected for continuity). When bringing all the genes into calculation, for regions with rates of < 0.5 SNP/kb, we found 97 nonsynonymous and 42 synonymous SNPs, a ratio of 2.31, compared to the value of 1.15 obtained for regions with rates of > 2 SNP/kb. When we calculated significance using the chi-squared test on the number of SNPs, we obtained Yates $\chi^2_1 = 14.14$, $P < 0.001$.

Domestication-related genes in SNP deserts

We found that over half of the confirmed domesticate genes were located in SNP deserts. Loss of seed shattering was a key event in the domestication of most cereals, and thus far two genes (*Sh4*, *q-SH1*) that control seed shattering have been identified in rice. Although *q-SHI* was not observed in SNP deserts, *Sh4* was found to reside only in one of the SNP deserts on chr4 (Li *et al.*, 2006; Lin *et al.*, 2007). The other two domestication-related alleles which influence rice flowering time, *Hd1* and *Hd6*, both resided in the SNP deserts on chr6 and chr3 respectively (Yano *et al.*, 2000; Takahashi *et al.*, 2001). The *Waxy* locus, which encodes a starch synthase, lies in one SNP desert on chr6 (Hirano *et al.*, 1998; Yamanaka *et al.*, 2004; Olsen *et al.*, 2006). Gene *GS3*, which controls grain size or shape, was also identified in the one SNP desert on chr3 (Fan *et al.*, 2006). An important characteristic difference between cultivated rice and wild rice is plant architecture; wild rice usually has a prostrate growth habit whereas cultivated rice shows relatively erect growth, which greatly improves photosynthetic efficiency. Recently, scientists successfully identified the gene (*PROG1*) as controlling the plant architecture in rice, and this gene was completely within one desert on chr7.

However, there were still five genes (QTLs) that could not be found in our SNP deserts. Both *qsW5* and *q-SH1* were not observed in SNP deserts, which were reasonable as both genes were only fixed in the *japonica* cultivar group. *Rc*, required for red pericarp in wild rice, had been shown to be closely associated with seed shattering and dormancy, and the recessive allele (*rc*) of *Rc* with a 14-bp deletion had been fixed in both *indica* and *japonica* subpopulations (Sweeney *et al.*, 2006, 2007). We did not observe the *rc* allele in SNP deserts, but its status as a true 'domesticate gene' is controversial, because unlike seed shattering and plant architecture, the trait of white pericarp is not indispensable for cultivated rice. All genes related to domestication in rice are listed in Table 2.

As shown in Table 2, 11 rice genes had been reported to be associated with domestication, of which six genes were located within SNP deserts. For 45 797 predicted genes, 3789 genes were found within SNP deserts and 42 008 genes were found outside. If we computed significance with Fisher's exact test on the number of genes, we would get $P < 0.01$, which suggested it was nonrandom and that SNP deserts contain the majority of domestication-related genes. To determine any differences in gene density between SNP deserts and non-SNP desert regions, we defined gene density by *Ds* gene/kb and found *Ds* = 0.116 for SNP deserts and *Ds* = 0.117 for

Table 1 K_a/K_s computed on per-gene basis.

		No. SNPs		Genes with K_a/K_s within the defined range r			
SNP rate	Non-KOME	NS	SY	NaN	$r = 0$	$0 < r \leq 1$	$r > 1$
< 0.5 SNP/Kb	1213	97	42	1180	30	7	68
> 2 SNP/Kb	13130	23994	20860	3093	2373	4772	2892

The number of nonsynonymous (NS) and synonymous (SY) SNPs is limited. When K_a/K_s is 0/0, we indicate that as Not-a-Number (NaN). Non-KOME refers to nonredundant KOME cDNA data set.

Table 2 List of all confirmed domestication-related genes in rice.

Gene	Location	In SNP desert	Function	Fixed in	Accession number	Reference (PMID)
Sh4(SHA1)	chr04	Yes	Shattering	*indica/japonica*	DQ383414	16527928
Hd6	chr03	Yes	Flowering time	–	DQ157463	11416158
Waxy	chr06	Yes	Starch synthase	*temperate japonica*	X62134	16547098
GS3	chr03	Yes	Grain size/shape	–	DQ355996	16453132
Hd1	chr06	Yes	Flowering time	–	AB041837	11148291
PROG1	chr07	Yes	Plant architecture	*indica/japonica*	FJ155665	18820696
Rc	chr07	No	Seed colour	*indica/japonica*	DQ204735	16399804
Ehd1	chr10	No	Flowering time	–	AB092508	15078816
Sd-1	chr01	No	Plant height	–	AF465256	11939564
q-SH1	chr01	No	Shattering	*temperate japonica*	AB071332	16614172
qSW5	chr05	No	Grain width	*japonica*	AB433345	18604208

non-SNP deserts (Fig. S5). Although SNP deserts comprised only 8% of the genome, over half (6 of 11) of the domestication-related genes were identified within them.

Level and pattern of genetic variation in cultivated and wild populations

Within 155 amplicons, 2630 SNPs were identified if all samples were taken into account. 1979 SNPs were identified if we only considered the 20 lines of wild rice. For 20 lines of *indica* and 23 lines of *japonica*, 803 and 766 SNPs were observed respectively. The number of polymorphic sites gave us an overview of the level of sequence diversity for three different populations, because the numbers of individuals in each group were nearly the same.

Within SNP deserts, nucleotide variability values (θ') for *indica*, *japonica*, and *temperate japonica* were 2.25×10^{-3}, 2.13×10^{-3} and 0.94×10^{-3} respectively. But for wild rice, including *O. nivara* and *O. rufipogon*, the average value was as high as 6.41×10^{-3} (Table 3). Our estimation for cultivated rice was slightly lower than a previous estimation (3.2×10^{-3}), but our estimation for wild populations was a bit higher than previously estimated (5.19×10^{-3}) (Caicedo *et al.*, 2007). When compared with other species, the level of polymorphism in cultivated rice was significantly lower than that in maize, a domesticated out-crossing crop (Tenaillon *et al.*, 2001), and slightly lower than that of *Arabidopsis thaliana*, a selfing, wild species (Nordborg *et al.*, 2005), but was approximately the same as that reported for sorghum (Hamblin *et al.*, 2004), and much higher than that found for soybeans (Zhu *et al.*, 2003).

Table 3 Level of genetic variation in and out of SNP deserts.

	SNP deserts (θ', $\times 10^{-3}$)	LOD in SNP deserts (%)	Control (θ', $\times 10^{-3}$)	LOD in control (%)
indica	2.25	64.90	3.89	36.54
japonica	2.13	66.77	2.91	52.53
*japonica*_TEJ	0.94	85.34	3.09	49.59
wild	6.41	0.00	6.13	0.00

Loss of genetic diversity (LOD) is calculated as $1.0 - \theta'_{\text{cultivated}}/\theta'_{\text{wild}}$. *japonica*_TEJ refers to *temperate japonica* cultivar group. Wild rice includes both *Oryza nivara* and *O. rufipogon*.

The loss of diversity (LOD) in cultivated species could be measured by $1.0 - \theta'_{\text{cultivated}}/\theta'_{\text{wild}}$. Based on previous studies, we could easily reach a conclusion that crops had lost about one-third (and retained two-thirds) of the variation found in the wild. In rice, a study using 111 randomly chose gene fragments over 72 diverse *O. sativa* accessions and 21 *O. rufipogon* accessions estimated that $\text{LOD}_{\text{rice}} = 38.34\%$ (Caicedo *et al.*, 2007). Maize lost genetic variation around 30% (Goloubinoff *et al.*, 1993; Eyre-Walker *et al.*, 1998; Hilton & Gaut, 1998), and bread wheat (durum wheat) lost 31% (16%) (Haudry *et al.*, 2007).

However, within rice SNP deserts, the LOD was much more severe than any of the above estimates, ranging from 64.90% to 85.34%. Moreover, when comparing SNP deserts with non-SNP deserts (control regions), we found that the diversity of SNP deserts was markedly lower than that of non-SNP deserts (Table 3). In combination, these findings indicated that, in addition to a population bottleneck, there must be other factors acting to lower the variation in SNP deserts. One plausible explanation would be gene introgression (which formed the SNP deserts) and the ensuing positive selection (see discussion).

Allele frequency distribution in three representative SNP deserts

The pattern of derived allele frequency spectrum (DAFS) for cultivated rice was clearly different from that of its wild relatives (Fig. 3a). Both *indica* and *japonica* presented an excessively high frequency of derived alleles (HFDs), whereas the wild rice exhibited a pattern which fits the neutral model (Fig. 3d).

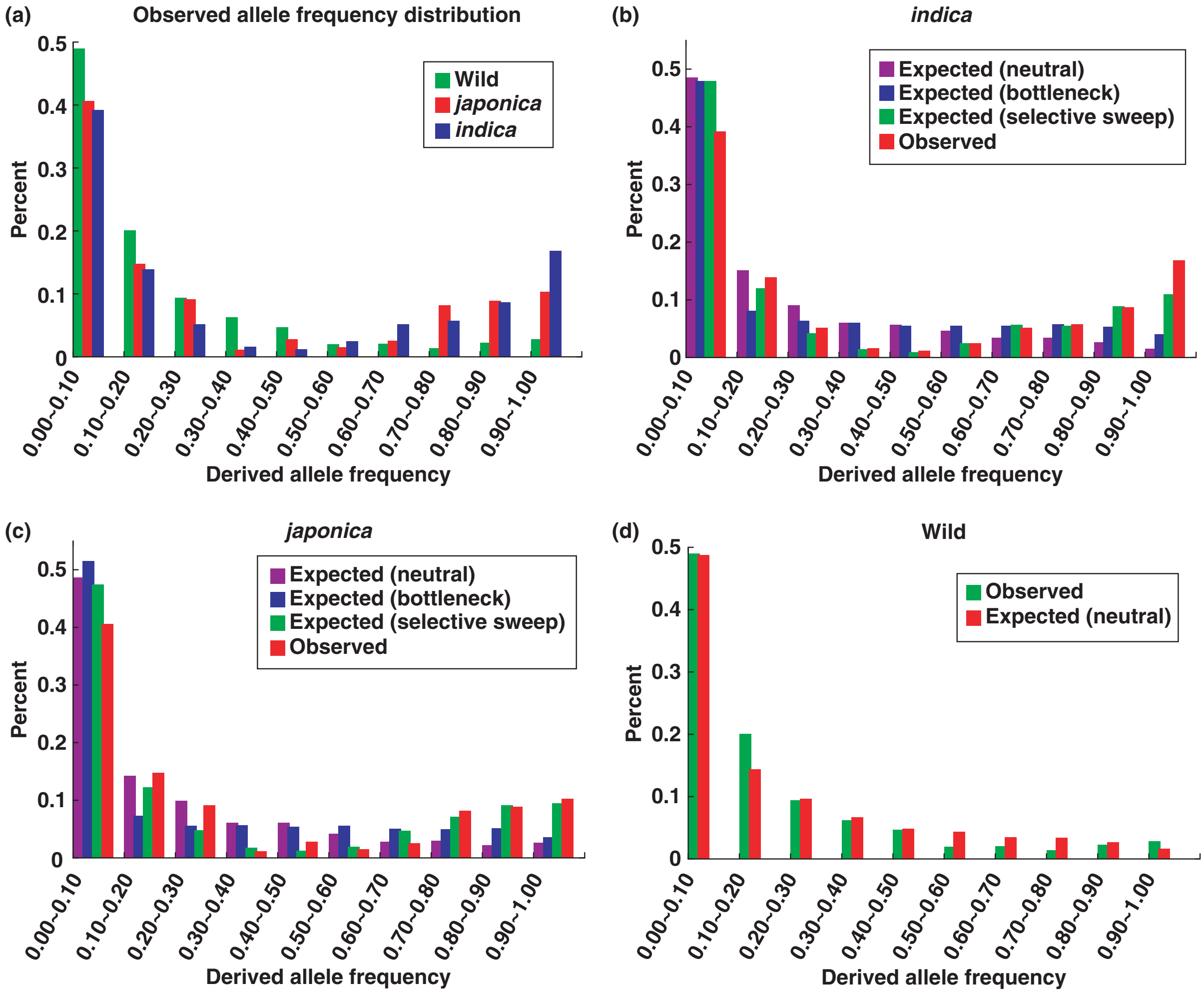

Fig. 3 (a) Observed patterns of derived allele frequency spectrum (DAFS) for *indica*, *japonica* and wild rice. (b, c) Comparisons of observed and simulated patterns of DAFS for *indica* and *japonica* respectively. Red histogram represents observed pattern. Green histogram represents pattern under selective sweep model, blue histogram represents pattern under bottleneck effect, purple histogram represents pattern under standard neutral model. (d) Comparisons of observed and simulated patterns of DAFS for wild rice. Green histogram represents observed pattern and red histogram represents pattern under standard neutral model.

A pattern of excess in HFDs is often interpreted as a unique signature of recent selective sweeps (Nielsen, 2005). To have a clearer view, we performed computer simulations based on the neutral, population bottleneck, and selective sweep models, using Sebastian E. Ramos-Onsins's program MLCOALSIM. Visual inspection of the observed and expected DAFS revealed that the selective sweep model predicts the DAFS of cultivated rice much better than the other two models (Fig. 3b,c). To test this assertion, we computed the Akaike information criterion (AIC). AIC is the standard method for comparing non-nested models; the preferred model is the one with the lowest AIC value. For the three different models, we have $\mathrm{AIC}_{\mathrm{neutral}} = -38.50$, $\mathrm{AIC}_{\mathrm{bottleneck}} = -41.00$ and $\mathrm{AIC}_{\mathrm{sweep}} = -61.84$ for *indica*, and $\mathrm{AIC}_{\mathrm{neutral}} = -44.48$, $\mathrm{AIC}_{\mathrm{bottleneck}} = -41.87$ and $\mathrm{AIC}_{\mathrm{sweep}} = -65.01$ for *japonica*. The AIC values obviously favoured the selective sweep model for both *indica* and *japonica*. In addition, we quantified the fitness among the models and the observed data using Pearson's chi-squared goodness-of-fit test. For *indica*, we derived $P = 0.98$ ($\chi_9^2 = 2.444$), $P < 0.05$ ($\chi_9^2 = 17.655$), and $P < 0.01$ ($\chi_9^2 = 25.292$) for the selective sweep, bottleneck and standard neutral models respectively. For *japonica*, we had $P = 0.93$ ($\chi_9^2 = 3.742$), $P = 0.055$ ($\chi_9^2 = 16.634$) and $P < 0.05$ ($\chi_9^2 = 17.498$) for the selective sweep, bottleneck and standard neutral models, respectively.

It should be noted that other factors such as misidentification of ancestral allelic states, trans-specific introgression and population subdivision could also lead to an excess of HFDs in the absence of selection. In each of the above cases, one would expect to see the same effect in wild rice; however, this was not observed, suggesting that the excess of HFDs was a result of positive selection. Furthermore, all HFDs in cultivated rice occurred as a low or intermediate frequency in wild rice, indicating that such alleles had only recently increased their frequency. The occurrence of multiple mutations could also increase the proportions of HFDs – to avoid this confounding factor, all polymorphic sites taken into consideration were bi-allelic sites (very few tri-allelic sites were removed).

A negative H value is also indicative of the excess of HFDs. Based on the allele frequency data, we calculated the Fay and Wu H value using the HTEST program on a local Unix computer (Fay & Wu, 2000). H was calculated for all 155 amplicons except the 23 amplicons presenting zero polymorphic sites. For *indica* and *japonica*, the average values of H across all amplicons were −2.61 and −1.41 respectively. For wild rice, the average value of H was 0.55.

Discussion

Probability of observing SNP deserts with recombination and population growth model

It is known that ascertainment bias can affect estimates of many evolutionarily important parameters, especially recombination rate (Clark *et al.*, 2005; Thornton & Jensen, 2007). This kind of ascertainment bias also exists when we search for SNP deserts and use only two rice genomes (i.e. 93-11 and Nipponbare); some rare SNPs may go undiscovered. However, we believe that the absence of rare SNPs would have little effect on defining SNP deserts, because most rare SNPs are newly occurring mutations, and scattered randomly throughout the genome. In addition, when performing our population analysis, we used 43 diverse accessions of cultivated rice and 20 accessions of wild rice, which will greatly reduce the ascertainment bias. In Fig. 3, the proportion of rare SNPs is nearly the same in the observed data as in the simulation data, indicating a very low ascertainment bias.

The so-called mismatch distribution, which depicts the pair-wise difference in a set of sequences, can also have a bimodal shape under certain conditions such as population expansion or slight recombination (Hudson, 1987; Slatkin & Hudson, 1991). However, the distribution in Fig.1 is different from the mismatch distribution; the former occurs at the genome level, and is derived by comparing two individuals at any loci throughout the genome, whereas the latter is at the population level, and is derived by comparing a locus or certain loci in any pair of individuals throughout the population.

Origin of SNP deserts

In this study, we observe the SNP deserts between *indica* (93-11) and *japonica* (Nipponbare), but how these SNP deserts originated depends on the scenario of rice domestication. If *indica* and *japonica* have a single origin, human selection for their preferred traits would leave SNP deserts in the genomes of cultivated rice. However, as mentioned earlier, increasing evidence supports the hypothesis that *indica* and *japonica* were domesticated independently. In this scenario, to explain the occurrence of SNP deserts in cultivated rice, selective sweeps (or human-favoured alleles) might have occurred in one subspecies and then introgressed into the other in one of several timeframes. Recent introgression could have created SNP deserts between the two subspecies, but this is less likely due to reproductive and geographical barriers. Introgression could also have happened a long time ago (i.e. before the divergence of the two subspecies), leaving SNP deserts that initially lay between *indica* and *japonica* but had been obscured by the following accumulation of mutations. As a third possibility, introgression occurred long ago could have been followed by positive selection within one or both species. Under this condition, the SNP deserts left by ancient introgression would remain as SNP deserts, consistent with our findings that regions approaching 0 SNP/kb present a high K_a/K_s ratio (Fig. 2).

It must be noted that not every SNP desert is necessarily a result of domestication; other factors, such as a cross between ancestors of 93-11 and Nipponbare could also leave an SNP desert in the genome. Therefore, the claim that all SNP deserts are domestication-related is not strictly valid, but we believe that most of them are, for the following reasons: (1) SNP deserts have a higher K_a/K_s ratio than other regions, (2) more than half of the domesticate alleles reside within them and (3) genetic variation of SNP deserts regions within cultivated groups is severely reduced. However, more data are needed to distinguish the parts of SNP deserts resulting from domestication from those with another origin.

Methods

Defining SNP deserts between *indica* (93-11) and *japonica* (Nipponbare)

We identified all sequence polymorphism based on a genome-wide comparison between the whole genome shotgun assemblies of Beijing *indica* (93-11) and Syngenta *japonica* (Nipponbare) (Goff *et al.*, 2002; Yu *et al.*, 2005). The genome sequence is available at http://rice.genomics.org.cn/rice/link/download.jsp. We aligned the sequence with CrossMatch, a Smith-Waterman algorithm built into Phrap (Phil Green http://www.phrap.org). To avoid paralogous sequence confusion, we used 34 190 unique sequences as anchoring points to identify all allelic counterparts. To distinguish polymorphisms from sequencing errors, we used the PHRED quality value Q ($Q = -10 \times \log(p)$, where p is the error probability), requiring $Q > 23$ at the variant sites, and

$Q > 15$ for both flanking 5 bp sequences. To avoid complexities arising from the intergenic regions where nested retrotransposons were clustered, we classified the sequence data into nonoverlapping five basic categories: unassembled, assembled-but-unmapped, mapped-but-unaligned, and aligned. We only identified sequence variations in the last category as it covered most of the rice genes. To ensure that the computer-identified SNPs were real, we manually inspected the raw sequencing traces for 97 nonsynonymous SNPs in the sequence regions under 0.5 SNP/kb; all were confirmed although five of them leveraged on single trace. A recent polymorphism study based on our *indica* draft sequence assembly reported a 98% confirmation rate for SNPs with a PCR-based re-sequencing procedure (Shen *et al.*, 2004).

We computed polymorphism rates, in different sliding windows, ranging from 5 to 200 kb and ignored gaps that are less than 2 kb in length to increase the number of usable windows up to 51.6% at the canonical size of 20 kb. We normalized the SNP rates over the aligned sequence in each bin rather than the bin size, and bins with less than 1 kb of aligned sequence were rejected.

Given the fast-evolving nature of plant intergenic sequences (Ma & Bennetzen, 2004), we must examine the extent to which the observed SNP rate bimodality is due to differences between genic and intergenic sequences. Therefore, we enriched the gene content of our sample by restricting our analysis to regions with a 20-mer copy number less than 10, where the copy number is defined as how often a 20-mer is perfectly repeated across the genome (Zhong *et al.*, 2003). This did not remove all intergenic sequences, but did capture 99.4% and 98.1% of the exon and intron sequences, respectively, based on our analysis of experimentally defined genes (Yu *et al.*, 2005). The end result was a shift in the major mode, down to about 6 SNP/kb. Conversely, we enriched the intergenic sequence content by restricting our analysis to regions identified by RepeatMasker (Smit and Green http://www.repeatmasker.org) as transposon-derived. This shifted the major mode, dramatically increasing it to about 18 SNP/kb. In both cases, the minor mode stayed fixed at zero, confirming that the existence of bimodality is not an artefact, even if the details were affected by genic and intergenic differences.

To determine the probability of observing the SNP deserts randomly, we performed a modal-based coalescent simulation using Mlcoalsim. We first performed the simulation under the recombination model, because recombination rate varied markedly within different genomic regions (Wu *et al.*, 2003). We calculated R ($R = 4Nr$, where R is the population recombination parameter, N is the population size and r the recombination rate per locus) using different values, ranging from 0 to 100. When $R = 0$, 0.01, 0.1, 1, 5, 10, 50 and 100, the empirical P-value = 0.0129, 0.016, 0.0133, 0.0120, 0.0067, 0.0057, 0.0003 and 0 respectively. All empirical P-values are below the 0.05 significance level (P-value is the probability of observing a 20-kb window with SNP/kb < 1) (Fig. S4).

Second, we considered population growth. Under this combined model, we adopt the population size parameters estimated by Caicedo *et al.* (2007). Taking *indica* as an example, during the bottleneck period the effective population size (*Nb*) is 1413, and the present population (*Np*) size is 40 229. We assume that rice domestication began 12 000 years ago, with a bottleneck lasting 3000 years. Therefore, it took 9000 years for the *indica* population to grow from 1414 to 40 229. When $R = 0$, 0.01, 0.1, 1, 5, 10, 50 and 100, the resulting empirical P-value = 0.0138, 0.0148, 0.0134, 0.0128, 0.0077, 0.006, 0.0003 and 0 respectively; again, all P-values are below 0.05 significance level (Fig. S4).

Cultivated and wild rice samples used in this study

Based on breeders' nomenclature, a total of 71 accessions were used in this study, including 29 *indica*, 14 *japonica*, 12 *O. rufipogon*, 8 *O. nivara*, 5 *aus*, 2 *O. glaberrima* and 1 *O. barthii*. To represent geographical diversity, samples were collected from 22 countries or areas throughout the world. All accessions belonged to the same genome group, which includes three major varietal groups (*indica*, *temperate japonica*, and *tropical japonica*) and two additional genetically distinct variety groups (*aus* and *aromatic*) (Khush, 1997; Garris *et al.*, 2005). Wild individuals were sampled to cover the entire distribution range of two wild species, *O. rufipogon* and *O. nivara*. Two accessions grown in Africa, *O. glaberrima* and *O. barthii*, were also collected to serve as an outgroup. We removed the five *aus* accessions during the following analysis because of their incongruity in our population structure analysis. Detailed information for all samples is given in Table S2.

After re-assignment by the Bayesian clustering program structure (Pritchard *et al.*, 2000), the cultivated group included 43 accessions (20 *indica* and 23 *japonica*), whereas the wild groups included 20 accessions (15 *rufipogon* and 5 *nivara*) (Fig. S6). Although misidentification occurred within both groups, we found no misidentification between them. However, what we were concerned about was the distinction between the cultivated and the wild. So adopting the structure designation was reasonable for this purpose.

Primer design, PCR amplification and sequencing

We designed one amplicon every 20 kb to cover the three representative SNP desert regions in a collective length of 3 Mb; each amplicon was so designed that was not limited solely to exon. Moreover, biologically defined repeats, GC-rich sequences and polymorphic sites were also avoided when designing primers. Among 155 amplicons we designed, 52, 55, 48, and 9 are for the SNP desert on chromosomes 3, 6, 7 and the control regions respectively. The control regions were at least 1 Mb away from any SNP desert regions. The nine control regions are distributed as follows: four on chromosome 3, one on chromosome 6, two on chromosome 7 and two on chromosome 10. The primers sequences were designed based on the *indica* (93-11) reference sequence (Aug. 2003, BGI) by Primer3 (Rozen & Skaletsky, 2000). The average length for all amplicons was 616 bp.

PCR amplification was performed in a total volume of 25 μL with a 2720 Thermal Cycler (Applied Biosystems, Foster City, CA, USA). The reaction mix contained 5–50 ng of template DNA, 1–2 U Taq DNA polymerase, 10 mm Tris-HCl (pH 8.3), 0.25 mm dNTPs, 0.2–2 mm BSA, 1.5–2.5 mm $MgCl_2$, 0.1 μm of each primer. Because of the different T_m, GC content and priming efficiency of different primer pairs, PCR ingredients and amplification conditions were optimized separately for each primer pair (details are available upon request). Amplification products were detected by electrophoresis on 1.5% agarose gels stained with ethidium bromide using a DL2000 DNA ladder, and purified with AcroPrep^{TM} 384 Filter Plates (PALL, East Hills, NY, USA). The sequencing reaction was performed in a final volume of 7 μL containing 50–100 ng purified amplification products, 2 μL DYEnamic ET Terminator Sequencing Kit premix, 1 μL 1 pm sequencing primers. The reaction cycles was set as 95 °C for 2 min, followed by 35 cycles of 95 °C denaturation for 15 s, 50 °C annealing for 15 s and 60 °C extension for 90 s. To obtain more reliable genotype, we sequenced all amplicons from both directions with an ABI 3730 automatic DNA sequencer.

Calculating nucleotide diversity

An SNP survey can be characterized in terms of either K, the observed number of segregating sites, or θ', the observed number of segregating sites per nucleotide site. Because K increased with the number of chromosomes (n) studied and the total sequence length (L), it was preferable to use the normalized number of variant sites:

$$\theta' = K / \sum_{i=1}^{n-1} \frac{L}{i}$$

Under neutral equilibrium evolution, θ' is an unbiased estimator of the population parameter $\theta = 4\ N\mu$, where N is the population size and μ is the mutation rate per nucleotide per generation (Watterson, 1975). *Oryza nivara* and *O. rufipogon* were treated as single wild species and *aus* rice were not included in the analysis.

Allele frequency distribution

In this work, as nearly all the polymorphic sites were bi-allelic, we could easily recognize ancestral alleles from derived alleles (mutant alleles) by using *O. glaberrima* and *O. barthii* as an outgroup. Those SNP sites without outgroup data were removed, and a few tri-allelic sites were also excluded.

Population bottleneck model

Coalescent simulations were used to model the impact of the bottleneck on sequence diversity using Mlcoalsim (Mitchell-Olds, 2007) that was based on Hudson's coalescent program ms (Hudson, 2002). *Oryza nivara* and *O. rufipogon* were treated as one single population as the progenitor of *O. sativa* according to previous studies (Morishima, 2001; Londo *et al.*, 2006). We modelled the bottleneck effect for *indica* and *japonica* independently, as several studies show that *indica* and *japonica* have separate domestication origins (Second, 1982; Londo *et al.*, 2006; Caicedo *et al.*, 2007).

The impact population bottleneck on sequence diversity was determined by several parameters: d, which is the duration of the bottleneck in generations; Nb, the effective population size during the bottleneck; Np, current population size; and Na, the ancestral population size. In this analysis, we assumed $d = 3000$ as archaeological evidence suggested that the domestication of rice began ~12 000 years ago, and continued until the appearance of domesticated rice grains. The population recombination parameters (4NrL) for *indica* and *japonica* were estimated using the DnaSP package (version 4.10.9) over 155 amplicons (Rozas *et al.*, 2003). The typical sigmoidal growth curve ($ts = 0$) was adopted. Other population parameters for *indica*, *japonica* and wild rice were chosen following a recent study based on 111 randomly chosen gene fragments over 72 diverse *O. sativa* accessions (Caicedo *et al.*, 2007).

Selective sweep model

The program Mlcoalsim (v1.25) was used to simulate sequence polymorphisms with a selective sweep (Mitchell-Olds, 2007). Again, *O. nivara* and *O. rufipogon* were treated as a single gene pool of wild rice. Scaled recombination rates per nucleotide ($Rn = 4N_e r$) for both *indica* and *japonica* were estimated from the 155 resequenced amplicons with the *DnaSP* package (version 4.10.9) as $Rn_{indica} = 0.0179$ and $Rn_{japonica} = 0.0093$ (Rozas *et al.*, 2003). To model the selective sweep effect, we assumed that there was one favourable site (the target of directed selection) every 20 kb because of our choice for one amplicon for every 20 kb of the desert sequence. As a result, the average size of selective sweeps was set to 20 kb, and the mean distance (d) between the selected site and the furthest hitchhiking site would be 10 kb. Because the distance (d) between a selected site and hitchhiking site depended on the strength of selection ($\alpha = 4N_e s$, where s is selection coefficient and α is a population selection parameter), we estimated α from the relationship $d = 0.01\alpha/C$ (Kaplan *et al.*, 1989), where C ($C = Rn = 4N_e r$) is the population recombination parameter. We obtained $\alpha_{indica} = 17\,900$ and $\alpha_{japonica} = 9300$.

Acknowledgments

We thank IRRI and Professor Lihuang Zhu for providing rice seeds and DNA samples. We are grateful to Dr Gane Ka-Shu Wong (University of Alberta, Canada) and Rusmus Nielsen (University of Copenhagen, Denmark) for their valuable advice. This work received financial support from the Chinese Academy of Science (KSCX1-SW-03), the Ministry of Science and Technology (2005AA235110 and 2006CB910404) and the National Natural Science Foundation of China (30221004 and 30600111).

References

Bautista, N., Vaughan, D., Jayasuriya, A., Liyanage, A., Kaga, A. & Tomooka, N. 2006. Genetic diversity in AA and CC genome *Oryza* species in southern south Asia. *Genet. Resour. Crop Evol.* **53**: 631–640.

Buckler, E.S.t., Thornsberry, J.M. & Kresovich, S. 2001. Molecular diversity, structure and domestication of grasses. *Genet. Res.* **77**: 213–218.

Caicedo, A.L., Williamson, S.H., Hernandez, R.D., Boyko, A., Fledel-Alon, A., York, T.L., Polato, N.R., Olsen, K.M., Nielsen, R., McCouch, S.R., Bustamante, C.D. & Purugganan, M.D. 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**: 1745–1756.

Clark, R.M., Linton, E., Messing, J. & Doebley, J.F. 2004. Pattern of diversity in the genomic region near the maize domestication gene tb1. *Proc. Natl Acad. Sci. USA* **101**: 700–707.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**: 1496–1502.

Cong, B. & Tanksley, S.D. 2006. FW2.2 and cell cycle control in developing tomato fruit: a possible example of gene co-option in the evolution of a novel organ. *Plant Mol. Biol.* **62**: 867–880.

Doebley, J.F., Gaut, B.S. & Smith, B.D. 2006. The molecular genetics of crop domestication. *Cell* **127**: 1309–1321.

Doi, K., Izawa, T., Fuse, T., Yamanouchi, U., Kubo, T., Shimatani, Z., Yano, M. & Yoshimura, A. 2004. Ehd1, a B-type response regulator in rice, confers short-day promotion of flowering and controls FT-like gene expression independently of Hd1. *Genes Dev.* **18**: 926–936.

Eyre-Walker, A., Gaut, R.L., Hilton, H., Feldman, D.L. & Gaut, B.S. 1998. Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl Acad. Sci. USA* **95**: 4441–4446.

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., Li, X. & Zhang, Q. 2006. GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theor. Appl. Genet.* **112**: 1164–1171.

Fay, J.C. & Wu, C.I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.

Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N. & Paterson, A.H. 2004. An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Res.* **14**: 1812–1819.

Foster, K.W. & Rutger, J.N. 1978. Inheritance of semidwarfism in rice, ORYZA SATIVA L. *Genetics* **88**: 559–574.

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

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. & Briggs, S. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92–100.

Goloubinoff, P., Paabo, S. & Wilson, A.C. 1993. Evolution of maize inferred from sequence diversity of an Adh2 gene segment from archaeological specimens. *Proc. Natl Acad. Sci. USA* **90**: 1997–2001.

Hamblin, M.T., Mitchell, S.E., White, G.M., Gallego, J., Kukatla, R., Wing, R.A., Paterson, A.H. & Kresovich, S. 2004. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics* **167**: 471–483.

Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., Hochu, I., Poirier, S., Santoni, S., Glemin, S. & David, J. 2007. Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* **24**: 1506–1517.

Hilton, H. & Gaut, B.S. 1998. Speciation and domestication in maize and its wild relatives: evidence from the globulin-1 gene. *Genetics* **150**: 863–872.

Hirano, H.Y., 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, R.R. 1987. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **50**: 245–250.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.

Innan, H. & Kim, Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl Acad. Sci. USA* **101**: 10667–10672.

Jin, J., Huang, W., Gao, J.P., Yang, J., Shi, M., Zhu, M.Z., Luo, D. & Lin, H.X. 2008. Genetic control of rice plant architecture under domestication. *Nat. Genet.* **40**: 1365–1369.

Kaplan, N.L., Hudson, R.R. & Langley, C.H. 1989. The ''hitchhiking effect'' revisited. *Genetics* **123**: 887–899.

Khush, G.S. 1997. Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.* **35**: 25–34.

Konishi, S., Izawa, T., Lin, S.Y., Ebana, K., Fukuta, Y., Sasaki, T. & Yano, M. 2006. An SNP caused loss of seed shattering during rice domestication. *Science* **312**: 1392–1396.

Li, C., Zhou, A. & Sang, T. 2006. Rice domestication by reducing shattering. *Science* **311**: 1936–1939.

Lin, Z., Griffith, M.E., Li, X., Zhu, Z., Tan, L., Fu, Y., Zhang, W., Wang, X., Xie, D. & Sun, C. 2007. Origin of seed shattering in rice (*Oryza sativa* L.). *Planta* **226**: 11–20.

Liu, A. & Burke, J.M. 2006. Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics* **173**: 321–330.

Londo, J.P., Chiang, Y.C., Hung, K.H., Chiang, T.Y. & Schaal, B.A. 2006. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc. Natl Acad. Sci. USA* **103**: 9578–9583.

Ma, J. & Bennetzen, J.L. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA* **101**: 12404–12410.

Mitchell-Olds, S.E. R.-O. a. T. 2007. Mlcoalsim: multilocus coalescent simulations. *Evol Bioinform* **3**: 41–44.

Morishima, H. 2001. *Evolution and Domestication of Rice*. International Rice Research Institute, Manila.

Nasu, S., Suzuki, J., Ohta, R., Hasegawa, K., Yui, R., Kitazawa, N., Monna, L. & Minobe, Y. 2002. Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa, Oryza rufipogon*) and establishment of SNP markers. *DNA Res.* **9**: 163–171.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**: 197–218.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N.A., Shah, C., Wall, J.D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M. & Bergelson, J. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**: e196.

Oka, H.I. 1988. *Origin of Cultivated Rice*. Japan Scientific Societies Press, Tokyo.

Olsen, K.M., Caicedo, A.L., Polato, N., McClung, A., McCouch, S. & Purugganan, M.D. 2006. Selection under domestication: evidence for a sweep in the rice waxy genomic region. *Genetics* **173**: 975–983.

Palaisa, K., Morgante, M., Tingey, S. & Rafalski, A. 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc. Natl Acad. Sci. USA* **101**: 9885–9890.

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. & Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.

Rozen, S. & Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**: 365–386.

Second, G. 1982. Origin of the genic diversity of cultivated rice (*Oryza* spp.): study of the polymorphism scored at 40 isozyme loci. *Jpn. J. Genet.* **57**: 25–57.

Shen, Y.J., Jiang, H., Jin, J.P., Zhang, Z.B., Xi, B., He, Y.Y., Wang, G., Wang, C., Qian, L., Li, X., Yu, Q.B., Liu, H.J., Chen, D.H., Gao, J.H., Huang, H., Shi, T.L. & Yang, Z.N. 2004. Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol.* **135**: 1198–1205.

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S. & Yano, M. 2008. Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* **40**: 1023–1028.

Simons, K.J., Fellers, J.P., Trick, H.N., Zhang, Z., Tai, Y.S., Gill, B.S. & Faris, J.D. 2006. Molecular characterization of the major wheat domestication gene *Q*. *Genetics* **172**: 547–555.

Slatkin, M. & Hudson, R.R. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–562.

Sweeney, M.T., Thomson, M.J., Pfeil, B.E. & McCouch, S. 2006. Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell* **18**: 283–294.

Sweeney, M.T., Thomson, M.J., Cho, Y.G., Park, Y.J., Williamson, S.H., Bustamante, C.D. & McCouch, S.R. 2007. Global dissemination of a single mutation conferring white pericarp in rice. *PLoS Genet.* **3**: e133.

Takahashi, Y., Shomura, A., Sasaki, T. & Yano, M. 2001. Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase CK2. *Proc. Natl Acad. Sci. USA* **98**: 7922–7927.

Tan, L., Li, X., Liu, F., Sun, X., Li, C., Zhu, Z., Fu, Y., Cai, H., Wang, X., Xie, D. & Sun, C. 2008. Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* **40**: 1360–1364.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. & Gaut, B.S. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl Acad. Sci. USA* **98**: 9161–9166.

Tenaillon, M.I., U'Ren, J., Tenaillon, O. & Gaut, B.S. 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* **21**: 1214–1225.

Thornton, K.R. & Jensen, J.D. 2007. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* **175**: 737–750.

Wang, H., Nussbaum-Wagler, T., Li, B., Zhao, Q., Vigouroux, Y., Faller, M., Bomblies, K., Lukens, L. & Doebley, J.F. 2005. The origin of the naked grains of maize. *Nature* **436**: 714–719.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**: 256–276.

Wu, J., Mizuno, H., Hayashi-Tsugane, M., Ito, Y., Chiden, Y., Fujisawa, M., Katagiri, S., Saji, S., Yoshiki, S., Karasawa, W., Yoshihara, R., Hayashi, A., Kobayashi, H., Ito, K., Hamada, M., Okamoto, M., Ikeno, M., Ichikawa, Y., Katayose, Y., Yano, M., Matsumoto, T. & Sasaki, T. 2003. Physical maps and recombination frequency of six rice chromosomes. *Plant J.* **36**: 720–730.

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

Yano, M., Katayose, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., Baba, T., Yamamoto, K., Umehara, Y., Nagamura, Y. & Sasaki, T. 2000. Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. *Plant Cell* **12**: 2473–2484.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, S., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Wang, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Wang, J., Wang, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Liu, J., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Zhang, J., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Zhang, Z., Zhang, Y., Huang, X., Su, Z., Tong, W., Li, J., Tong, Z., Li, S., Ye, J., Wang, L., Fang, L., Lei, T., Chen, C., Chen, H., Xu, Z., Li, H., Huang, H., Zhang, F., Xu, H., Li, N., Zhao, C., Li, S., Dong, L., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Zhang, B., Hu, W., Zhang, Y., Tian, X., Jiao, Y., Liang, X., Jin, J., Gao, L., Zheng, W., Hao, B., Liu, S., Wang, W., Yuan, L., Cao, M., McDermott, J., Samudrala, R., Wang, J., Wong, G.K. & Yang, H. 2005. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**: e38.

Zeder, M.A., Emshwiller, E., Smith, B.D. & Bradley, D.G. 2006. Documenting domestication: the intersection of genetics and archaeology. *Trends Genet.* **22**: 139–155.

Zhong, L., Zhang, K., Huang, X., Ni, P., Han, Y., Wang, K., Wang, J. & Li, S. 2003. A statistical approach designed for finding mathematically defined repeats in shotgun data and determining the length distribution of clone-inserts. *Genomics Proteomics Bioinformatics* **1**: 43–51.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D. & Cregan, P.B. 2003. Single-nucleotide polymorphisms in soybean. *Genetics* **163**: 1123–1134.

Zhu, Q., Zheng, X., Luo, J., Gaut, B.S. & 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.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1 SNP rate bimodality vs. window size ranges from 5 to 200 kb. The comments given at the top-left corners are the peak magnitudes at 0 SNP/kb.

Figure S2 SNP rate bimodality after restricting to genic and intergenic proxies. (a) Gene enriched regions with 20-mer copy numbers less than 10. (b) Intergenic enriched regions identified by RepeatMasker as being of transposon origin.

Figure S3 Indel rate bimodality and correlation with SNP rates. For the latter, each point represents the rate in a single 20-kb window. For clarity, we show only 10 000 random points. The best fit through the origin has a slope of 1–6.19.

Figure S4 Model-based coalescent simulation. Distributions were obtained from 10 000 coalescent simulations in a sample of $n = 4$ chromosome (corresponding to two diploid individuals). The mutation parameter was set to $\theta = 3.2 \times 10^{-3}$. R (recombination rate per locus) was set to 0, 0.01, 0.1, 1, 5, 10, 50 and 100. (a) Simulation based on recombination model. (b) Simulation based on recombination plus population growth model.

Figure S5 A graphical depiction of all predicted genes and SNP deserts along the 12 chromosomes. Green bars represent all predicted genes, red bars represent highly conserved genes (with 0 SNPs) between *indica* (93-11) and *japonica* (Nipponbare), and blue bars represent SNP deserts.

Figure S6 Estimated population structure for 71 accessions of *O. sativa*, *O. rufipogon*, *O. nivara*, *O. glaberrima* and *O. barthii* from 155 noncoding fragments. Population parameter $K = 5$. Colours indicate the five different clusters given by software structure, with horizontal orange lines are used to separate six traditional groups designated by breeders. Each horizontal box represents an accession.

Table S1 Chromosome coordinates for all SNP deserts (SNP/kb < 1) identified. Positions are with regard to *indica* (93-11) reference genome (Aug. 2003, BGI).

Table S2 All rice samples used in this study. The samples were collected from 22 different countries or areas throughout the world to capture genetic diversity.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Received 10 August 2008; revised 9 December 2008; accepted 2 January 2009