MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/j.1365-294X.2012.05625.x

Phylogeography of Asian wild rice, *Oryza rufipogon*: a genome-wide view

PU HUANG,* JEANMAIRE MOLINA,† JONATHAN M. FLOWERS,† SAMARA RUBINSTEIN,† SCOTT A. JACKSON,‡ MICHAEL D. PURUGGANAN† and BARBARA A. SCHAAL*

Department of Biology, Washington University, St. Louis, MO 63130, USA, †Department of Biology and Center for Genomics and Systems Biology, New York University, New York, NY 10003 USA, ‡Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30621, USA

Abstract

Asian wild rice (*Oryza rufipogon*) that ranges widely across the eastern and southern part of Asia is recognized as the direct ancestor of cultivated Asian rice (*O. sativa*). Studies of the geographic structure of *O. rufipogon*, based on chloroplast and low-copy nuclear markers, reveal a possible phylogeographic signal of subdivision in *O. rufipogon*. However, this signal of geographic differentiation is not consistently observed among different markers and studies, with often conflicting results. To more precisely characterize the phylogeography of *O. rufipogon* populations, a genome-wide survey of unlinked markers, intensively sampled from across the entire range of *O. rufipogon* is critical. In this study, we surveyed sequence variation at 42 genome-wide sequence tagged sites (STS) in 108 *O. rufipogon* accessions from throughout the native range of the species. Using Bayesian clustering, principal component analysis and AMOVA, we conclude that there are two genetically distinct *O. rufipogon* groups, Ruf-I and Ruf-II. The two groups exhibit a clinal variation pattern generally from north-east to south-west. Different from many earlier studies, Ruf-I, which is found mainly in China and the Indochinese Peninsula, shows genetic similarity with one major cultivated rice variety, *O. satvia indica*, whereas Ruf-II, mainly from South Asia and the Indochinese Peninsula, is not found to be closely related to cultivated rice varieties. The other major cultivated rice variety, *O. sativa japonica*, is not found to be similar to either *O. rufipogon* groups. Our results support the hypothesis of a single origin of the domesticated *O. sativa* in China. The possible role of palaeoclimate, introgression and migration–drift balance in creating this clinal variation pattern is also discussed.

Keywords: domestication, *Oryza rufipogon*, *Oryza sativa*, phylogeography

Received 8 December 2011; revision received 10 March 2012; accepted 14 March 2012

Introduction

One of the key aims of crop domestication studies is to characterize the levels, apportionment and geographic distribution of genetic variation in the wild progenitors of cultivated species (Diamond 2002; Doebley *et al.* 2006). In the past several decades, crop scientists and evolutionary biologists made key advances in understanding basic aspects of crop domestication, including fundamental features such as resolving when and where domestication initially occurred (Second 1982; Heun *et al.* 1997; Sun *et al.* 1997; Wang *et al.* 1998; Matsuoka *et al.* 2002; Londo *et al.* 2006; Zhang *et al.* 2009; also reviewed by Doebley *et al.* 2006), determining the origin of crop-related agricultural weeds (Londo & Schaal 2007; Kane & Rieseberg 2008; Reagon *et al.* 2010) and identifying potential genetic resources for crop breeding and improvement (Yuan *et al.* 1989). However, addressing these questions without a thorough understanding of the variation patterns in the wild progenitor can be misleading (Diamond 2002; Doebley *et al.* 2006; Buckley 2009). In several systems, well-defined phylogeographic patterns of progenitor species

Correspondence: Barbara Schaal, Fax: 314 935 4432;
E-mail: schaal@wustl.edu

have provided seminal insights into the process of domestication. For example, in corn (*Zea mays*), a single domestication centre in south-western Mexico was clearly identified and repeatedly verified by different genetic markers with thorough geographic sampling (Wang *et al.* 1998; Matsuoka *et al.* 2002). In einkorn wheat (*Triticum boeoticum*), archaeological excavations together with genetic data traced its origin to the Karacadağ mountains of Turkey (Heun *et al.* 1997; Dubcovsky & Dvorak 2007). However, in other study systems, there is frequent incongruence among different studies and genetic markers. Domesticated Asian rice (*Oryza sativa*) is one such species (Sweeney & McCouch 2007; Vaughan *et al.* 2008; Gross & Olsen 2010).

Asian wild rice (*Oryza rufipogon*) is widely recognized as the direct ancestor of cultivated *O. sativa*, including its two major varieties, *O. sativa japonica* and *O. sativa indica* (Chou 1948; Second 1982; Zhu & Ge 2005; also reviewed by Sweeney & McCouch 2007; Vaughan *et al.* 2008). *Oryza rufipogon* is a perennial plant commonly found in marsh or aquatic habitats of eastern and southern Asia (Fig. 1a). It is mainly wind-pollinated and has a mixed mating system compared with its derived crop that is predominantly selfing (Gao *et al.* 2002; Song *et al.* 2003; Chen *et al.* 2004). Flowering time in native populations of *O. rufipogon* usually ranges from September to November and overlaps with some cultivated varieties (Song *et al.* 2003). The annual ecotype of *O. rufipogon*, *Oryza nivara*, with an as yet undefined relationship to the perennial *O. rufipogon* (Lu *et al.* 2002; Londo *et al.* 2006), is treated as part of the same ancestral gene pool of *O. sativa* and not separated from the typical perennial *O. rufipogon* in our study.

Phylogeographic studies of *O. rufipogon* have often yielded inconsistent or conflicting results. An early study employing isozyme markers identified two genetic lineages of *O. rufipogon*, which were genetically distinct and geographically separated at the division between China and Southeast Asia; these lineages had genetic affinity to *japonica* and *indica* rice, respectively (Second & Morishima 1981; Second 1982). Another study, using restriction fragment length polymorphisms (RFLPs), found four different genetic groups in *O. rufipogon*: an *indica*-like *O. rufipogon* group in China, South and Southeast Asia, a *japonica*-like *O. rufipogon* group in China, an ancestral *O. rufipogon* group in South and Southeast Asia, and another ancestral *O. rufipogon* group in China (Sun *et al.* 1997). A more recent study examined sequence polymorphism in two nuclear and one chloroplast markers (Londo *et al.* 2006) and

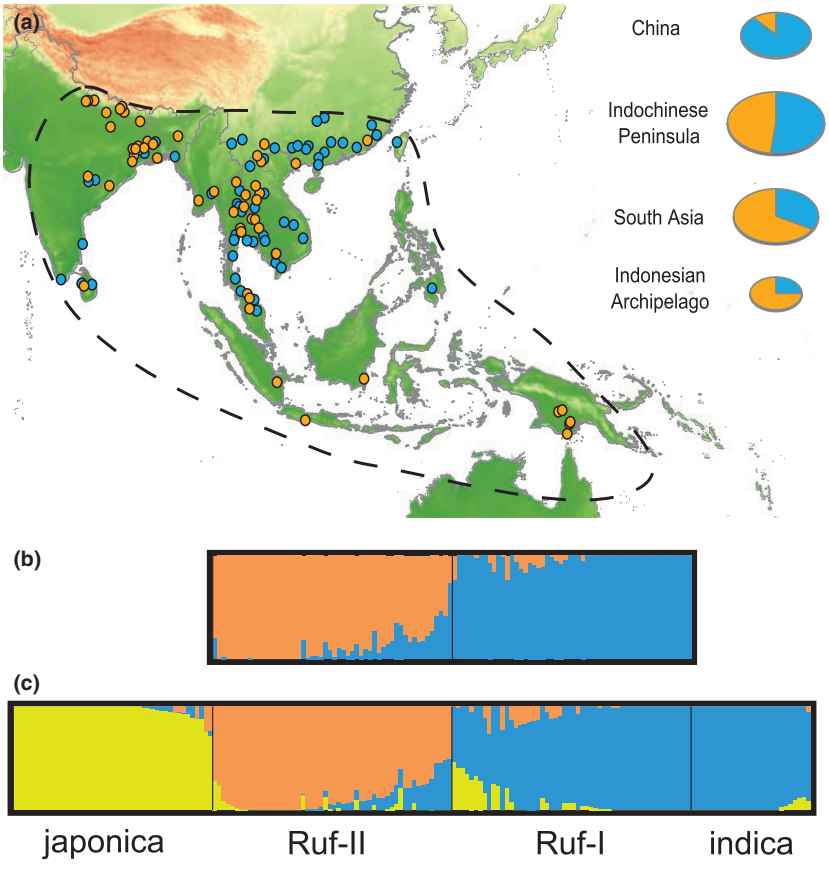

Fig. 1 (a) Geographic location of 108 *O. rufipogon* accessions used in this study. It should be noted that for a few accessions in China ($n = 15$), detailed location information was not provide by IRRI. We randomly generated 15 pseudopoints in China to represent these accessions. The area within the dashed line is the native distribution range of *O. rufipogon* in Asia (summarized from Zhou *et al.* 2003; Vaughan *et al.* 2008). Pie chart gives the number of accessions of the two genetic clusters (Ruf-I in blue and Ruf-II in orange) found in each geographic region (Table 1). Sizes of the pie charts correspond to the number of total sampled accessions in each region. (b) Bayesian clustering result using BAPS based on *O. rufipogon* samples and (c) *O. rufipogon* and *O. sativa* samples. Each vertical bar represents an *Oryza* accession. The proportion of the cluster membership is given by the length of each coloured segment in a bar.

found yet another pattern. This study identified a wide-ranging ancestral type and three geographically localized types of wild rice, including a *japonica*-like group in China, an *indica*-like group in the Indochinese Peninsula and a third group in South Asia with genetic similarity to a special drought-tolerant *indica* variety, the *aus-indica* rice (Garris *et al.* 2005).

In addition to these studies that use neutral genetic markers, the phylogeography of *O. rufipogon* has also been examined in several functional gene studies including those of seed shattering gene *Sh4* (Zhang *et al.* 2009), red pericarp gene *Rc* (Sweeney *et al.* 2007) and hybrid male sterility genes *SaF* and *SaM* (Long *et al.* 2008). While the phylogeography of *O. rufipogon* was not emphasized *per se* in these studies, a single geographic origin from China for two of these genes (*Sh4*, *Rc*) was suggested (Vaughan *et al.* 2008). Finally, many other studies address issues associated with the phylogeography of *O. rufipogon*, including the domestication history of *O. sativa* (Caicedo *et al.* 2007; Zhu *et al.* 2007; Molina *et al.* 2011), fine-scale population genetics of *O. rufipogon* (Gao *et al.* 2002; Zhou *et al.* 2003; Gao 2004; Wang *et al.* 2008) and weedy rice origin (Londo & Schaal 2007; Reagon *et al.* 2010). In general, however, the geographic coverage in these studies is limited, making comparisons of population structure difficult.

In summary, there is a rough agreement that *O. rufipogon* is geographically subdivided, but the number and the boundaries of the constituent groups and their relationship to the two cultivated varieties remain controversial. Part of these discrepancies may be due to a small number of loci employed in these studies, which reflect incomplete lineage sorting introducing misleading phylogeographic signals (Knowles & Carstens 2007; Degnan & Rosenberg 2009). Restricted geographic sampling, on the other hand, can cause biased spatial patterns of genetic variation (Buckley 2009).

A genome-wide survey of unlinked markers intensively sampled from across the entire distribution range is critical to more precisely characterize the phylogeography of *O. rufipogon* populations. In this study, we surveyed single-nucleotide polymorphisms (SNPs) at 42 genome-wide unlinked sequence tagged sites (STS) in 108 *O. rufipogon* accessions (including 12 accessions from a previous study, Caicedo *et al.* 2007) from throughout its native range. Our study addresses three questions concerning the phylogeography of *O. rufipogon*: (i) what is the population genetic structure of *O. rufipogon* and how is it related to the geographic distribution of this species, (ii) in what way the *O. rufipogon* populations are related to the two major *O. sativa* varieties and (iii) what are the possible historical causes of the contemporary population structure and the geographic pattern of genetic diversity in *O. rufipogon*?

Materials and methods

Sampling and sequencing

Germplasm of 96 *O. rufipogon* accessions (including 15 *O. nivara*) and one outgroup species, *O. meridionalis*, was obtained from the International Rice Research Institute (IRRI, Table S1). Accessions were chosen to maximize the geographic coverage of the current range of *O. rufipogon* with samples from most of the countries in southern and eastern Asia (Table 1, Fig. 1a). Initially 6–7 seeds per accession were planted in the greenhouse of Washington University to ensure germination. A final panel of 97 individuals, including a single plant from each accession, was used for analysis.

Young and healthy leaf material was collected from the plants about 3 months after germination. Total DNA was extracted from 1 g of fresh leaf material using a CTAB protocol (Gross *et al.* 2009) with slight modifications. Forty-two randomly chosen STS loci, approximately four loci per chromosome, were sequenced (Table S2, Fig. S1). These loci are a subset of 111 expression sequence tag (EST)-based STS loci. These 111 STS loci have been used to successfully characterize genetic variation in domesticated *O. sativa* (Caicedo *et al.* 2007).

DNA sequencing was carried out in the Beckman & Coulter Genomics facilities (Danvers, MA, USA) and in the DNA sequencing facility at Washington University. Raw base calls were manually corrected and assembled into contigs using Sequencher 4.8 (Gene Codes Corp.). Newly assembled contigs were then aligned together with those from a previous study (Caicedo *et al.* 2007), yielding 1 *O. meridionalis*, 108 *O. rufipogon* and 72 *O. sativa* accessions per STS locus. The *O. sativa* sample included 41 *O. sativa japonica* and 31 *O. sativa indica* accessions (Caicedo *et al.* 2007; Table S1). Geographically, the *O. sativa* sample covered most of the major rice production countries. Most of the accessions are landraces, but elite cultivars were also included (Caicedo *et al.* 2007; Table S1). The algorithm PHASE 2.1

Table 1 Geographic distributions of the two genetic groups of *O. rufipogon*

Genetic group*	China†	Indochinese Peninsula‡	South Asia§	Indonesian Archipelago¶
Ruf-I	20	22	10	3
Ruf-II	2	20	22	9
Total	22	42	32	12

*Groups based on BAPS analysis (Fig. 1).
†China.
‡Myanmar, Thailand, Vietnam, Laos, Cambodia.
§India, Sri Lanka, Nepal, Bangladesh.
¶Philippines, Indonesia, Malaysia, Papua New Guinea.

(Stephens *et al.* 2001; Stephens & Donelly 2003) was used to infer the two haplotypes in each individual. To eliminate possible human-induced inbreeding during *O. rufipogon* germplasm maintenance, final alignments for later analysis included only one randomly chosen haplotype per individual for all STS loci.

Population structure

The Bayesian clustering program BAPS 5.3 (Corander & Marttinen 2006; Corander *et al.* 2008) was used to identify the genetic structure of *O. rufipogon* from a model-based perspective. Compared with the commonly used program STRUCTURE that assumes free recombination among loci, BAPS provides a linked loci clustering model that provides a more precise and stable result for closely linked data sets than recombination models (Corander & Tang 2007). Accordingly, BAPS is more appropriate for STS sequences in this study. Population structure was examined in two ways in BAPS: first, using only *O. rufipogon* accessions (n = 108), and then using both *O. rufipogon* and *O. sativa* accessions (n = 180). In both cases, the cluster number, K, was tested from 2 to 10, and the optimal K was chosen by the program.

We also used a non-model-based principal component analysis (PCA) to examine population structure. Sequence alignments were combined and transformed into SNP matrices for the PCA. The analysis was then carried out by a nonlinear interactive partial least squares (Nipals) PCA function in the pcaMethods package (Stacklies *et al.* 2007) of R 2.11.1 (R Development Core Team, 2011). The Nipals PCA is used here because of its higher efficiency in handling large data sets compared with traditional PCA methods. Missing data were estimated automatically by the algorithm. Similar to the clustering analysis, PCA was performed using only the *O. rufipogon* accessions and then again with *O. rufipogon* plus *O. sativa* accessions.

Phylogeography

To detect a phylogeographic signal in *O. rufipogon*, we divided the distribution range into four regions: China, the Indochinese Peninsula, South Asia and the Indonesian Archipelago (Table 1, Fig. S2). This delimitation is based on both the topography of the region (i.e. the Tibetan Plateau and Hengduan Mountains that separate China, the Indochinese Peninsula and South Asia) and the potential boundaries of genetic groups defined in previous studies (Second & Morishima 1981; Sun *et al.* 1997; Londo *et al.* 2006). Several different methods were employed to determine phylogeographic structure. First, the accessions of each cluster generated by BAPS were partitioned geographically, and then, the resulting four regions were compared with determine differences in regional composition using Fisher's exact test (Table 1). Second, we used a locus-by-locus analysis of molecular variance (AMOVA) in the program Arlequin 3.5 (Excoffier & Lischer 2010) based on regional delimitations as well as genetic groups delimited by the Bayesian clustering algorithms to verify the geographic structure statistically. Finally, a correlation analysis was conducted using the genetic and geographic distance matrices. The genetic distance matrix was calculated based on the concatenated alignments of 42 STS loci. Because detailed sample location was sometimes not available in the IRRI database, we calculated the geographic distance matrix using randomly generated pseudopoints in the sampling areas, and this process was repeated for 1000 iterations. The average correlation coefficient between genetic and geographic distances and its standard deviation was then calculated.

Summary statistics and neutrality tests

Genetic diversity statistics were calculated for all 42 STS loci using DNASP 5.1 (Librado & Rozas 2009). Polymorphism at SNP loci was determined for each STS by the number of segregating sites (S), haplotype diversity (Hd), pairwise differences π and Watterson's θ (Watterson 1975). Each STS was compared with the annotated *O. sativa japonica* (Nipponbare) genomic sequence in GenBank for site-type determination (Rice Annotation Project Database, RAP-DB, Tanaka *et al.* 2008). The aforementioned summary statistics were calculated for each site type (nonsynonymous, synonymous and silent) and the whole sequence for each STS locus.

We applied two methods to test the neutrality of the STS loci in *O. rufipogon* samples: Tajima's D (Tajima 1989a,b) and an HKA test (Hudson *et al.* 1987). Tajima's D was calculated using DNASP 5.1 (Librado & Rozas 2009), and a multiple test Šidák correction was used to correct the P-value for all loci. The HKA test was performed with the program HKA (Wang & Hey 1996), with *O. meridionalis* as an outgroup. *Oryza barthii* sequences from a previous study (Caicedo *et al.* 2007) were used for locus sts_068 where *O. meridionalis* sequences were not available.

Results

Marker data

The final aligned sequences of the 42 STS loci have lengths ranging from 393 to 551 bp, with an average of 476 bp (Table S2). A small fraction, 1.1% (51/4536), of the sequence data was missing because of multiple indels and possible polymorphism at primer binding

sites. The total alignment length is ~20 kb. Based on the alignment after statistical phasing, the number of the haplotypes varies from 3 to 23 across 42 STS loci, with an average of 8.4 haplotypes per locus. We detected 1065 SNPs when the outgroup sequences were included. Within the *rufipogon–sativa* complex, 818 SNPs were detected, and among them, we observed 800 biallelic SNPs and 18 SNPs with more than two alleles. Finally, 763 biallelic SNPs were found in the *O. rufipogon* samples alone. Only the biallelic SNPs were used in the later PCA analysis.

Population structure

The Bayesian clustering analysis of *O. rufipogon* (n = 108) and the analysis with *O. rufipogon* and *O. sativa* combined (n = 180) both indicated two distinct genetic groups in *O. rufipogon* (Fig. 1b,c), referred to as Ruf-I and Ruf-II. The placement of accessions into a cluster was generally consistent between the two analyses for all *O. rufipogon* individuals, except for 103308, which was assigned into the *japonica* rice cluster based on the combined BAPS analysis (Table S1). Two large genetic groups were observed in the *O. sativa* sample, corresponding to the two traditional varieties of rice, *O. sativa indica* and *O. sativa japonica* (Fig. 1c). Ruf-I is in the same genetic cluster with *O. sativa indica* when *O. sativa* samples were included in the analysis (Fig. 1c), indicating a shared genetic background of the two groups. In contrast, except for the few admixed individuals, Ruf-II showed little affinity to either *O. sativa indica* or *O. sativa japonica*. Although a small number of admixed individuals were observed, the majority of accessions showed a low level of admixture.

The result of Nipals PCA showed similar patterns to the Bayesian clustering. In the analysis using only *O. rufipogon*, the first two principal components (PCs) explained 11.7% and 7.33% of the total variance, respectively. Variances along the first PC axis correspond to the Ruf-I and Ruf-II groups, and Ruf-I showed a higher level of variation compared with Ruf-II (Fig. 2a). When *O. sativa* samples were included in the analysis, the first two PCs in this case explained 18.03% and 9.62% of the total variance. *Oryza sativa japonica* and *O. sativa indica* were each tightly clustered but distinct. A sharp distinction between Ruf-I and Ruf-II was again observed (Fig. 2b), and Ruf-I overlapped the *O. sativa indica* cluster. Finally, both *O. rufipogon* groups showed a much higher level of variation compared to the two cultivar groups.

Phylogeography

We observed a strong phylogeographic signal in *O. rufipogon*. Ruf-I has the highest frequency in China and moderate frequency in the Indochinese Peninsula, whereas in South Asia and the Indonesian Archipelago, Ruf-II is in higher frequency (Table 1, Fig. 1a). A two-sided Fisher's exact test over all four regions significantly rejects the null hypothesis of no association between a genetic cluster and geographic region (P = 0.00016). Pairwise Fisher's exact tests showed significant genetic–geographic association between China and South Asia (P = 0.000079, Table S3) and between China and the Indonesian Archipelago (P = 0.00065, Table S3), while no significant association was observed in other paired regions after Bonferroni correction. In general, a clinal pattern of variation in the two genetically distinct *O. rufipogon* groups, geographically from north-east to south-west, was observed. This geographic signal was further documented by AMOVA and correlation analysis. The AMOVA design based on the regional delimitation yielded an among-region variance component that explained approximately 8.74% of the total genetic variation (F_{ST} = 0.087, P < 0.0001, Table 2). Twenty thousand bootstraps indicated that the 95% confidence interval (CI) of F_{ST} ranged from 0.071 to 0.104 (Table 2). In comparison, the AMOVA design based on Ruf-I and Ruf-II delimitation gave an average F_{ST} = 0.138 with a 95% CI from 0.100 to 0.174.

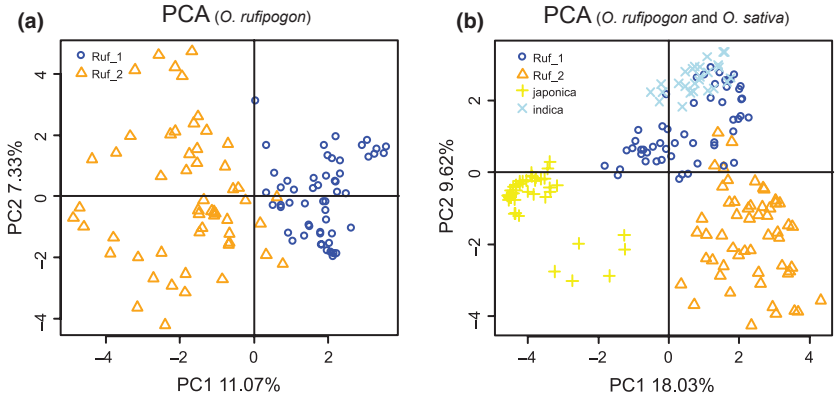

Fig. 2 Nipals principal component analysis on a) *O. rufipogon* samples and b) *O. rufipogon* and *O. sativa* samples. The first two principal components (PCs) are shown on the horizontal and vertical axes, respectively, with numbers representing the percentage variance explained by the PC. Each symbol represents an individual. The shape and colour of the symbols correspond to the group or variety to which the individual belongs, illustrated by the legend at the top left of the figure.

Table 2 Locus-by-locus AMOVA for *O. rufipogon* based on geography and BAPS groups

AMOVA design	Source of Variation	Sum of Squares	Variance Components	Percentage variation
Delimitation based on geography*	Among region	428.15	4.029	8.74[b]
	Within region	4320.39	42.054	91.26
	Total	4730.54	46.083	
Delimitation based on BAPS groups	Among group	394.45	6.671	13.76[c]
	Within group	4337.52	41.794	86.24
	Total	4731.97	48.46	

*Delimitation is the same as shown in Table 1.
[bc]$P < 0.0001$ based on significance test (10 000 permutations)

Correlation analysis between the genetic distance matrix and the geographic distance matrix revealed a positive correlation coefficient (average $r = 0.255$, SD = 0.0059).

Genetic diversity and neutrality tests

The average silent sites π and θ of *O. rufipogon* across 42 STS loci are approximately 6.41×10^{-3} and 9.62×10^{-3}, respectively (Tables 3 and S2). They are, as expected, higher than the estimates for *O. sativa* with a mean π of 2.75×10^{-3} and θ of 2.85×10^{-3} (Tables 3 and S2). Between the two groups of *O. rufipogon*, π and θ of Ruf-I are 4.71×10^{-3} and 7.01×10^{-3}, respectively, and they are approximately 30% and 20% lower compared with 7.26×10^{-3} and 8.99×10^{-3} in Ruf-II (Tables 3 and S2). Molecular diversity indices at synonymous sites are generally similar to those of silent sites in all comparisons, whereas at nonsynonymous sites, a dramatically lower diversity level (80%-89% lower compared with the corresponding indices at silent sites) was observed.

The majority of STS loci showed a negative Tajima's *D* in *O. rufipogon* (39 of 42, Table S2). Although some STS loci (10 of 42) showed a negative *D* with a *P*-value <0.05, significant deviation from neutral expectation was not observed after the Šidák correction (adjusted *P*-value for a single test is 0.0012). A neutral hypothesis was accepted in the HKA test, using both the standard chi-square statistic ($P = 0.090$, Hudson *et al.* 1987) and maximum cell value ($P = 0.061$, Wang & Hey 1996).

Discussion

Genetic diversity in O. rufipogon and *O. sativa*

Molecular diversity indices such as π and θ calculated from silent sites are indicators of the level of neutral polymorphisms within a species. In this study, the average silent site diversity ($\pi = 6.41 \times 10^{-3}$ and $\theta = 9.62 \times 10^{-3}$) is generally higher than that from several other studies of *O. rufipogon* using the same STS loci but with a more restricted geographic sample ($\pi = 5.19 \times 10^{-3}$ and $\theta = 5.42 \times 10^{-3}$ in Caicedo *et al.* 2007; $\pi = 6.35 \times 10^{-3}$ and $\theta = 7.787 \times 10^{-3}$ in Reagon *et al.* 2010). This difference in levels of diversity suggests that previously unsampled genetic variation was detected in our study, probably due to the wider geographic coverage of our samples. We also expected that the diversity

Table 3 Summary of the average molecular diversity indices across 42 sequence tagged sites (STS) loci

Statistic	Category	*O. rufipogon*			*O. sativa**
		Combined	Ruf-I	Ruf-II	
π per kb	All sites	4.41	3.38	4.80	2.21
	Silent sites[†]	6.41	4.71	7.26	2.78
	Synonymous sites	6.60	4.49	7.34	3.50
	Nonsynonymous sites	0.77	0.57	0.84	0.50
θ per kb	All sites	7.48	5.25	6.53	2.22
	Silent sites[†]	9.62	7.01	8.67	2.85
	Synonymous sites	9.78	7.37	8.99	2.84
	Nonsynonymous sites	1.98	1.22	1.61	0.87
Tajima's D[‡]	All sites	−1.19	−1.08	−0.80	0.11
	Silent sites[†]	−1.00	−0.90	−0.63	0.20
	Synonymous sites	−0.70	−0.83	−0.47	0.24
	Nonsynonymous sites	−1.05	−1.06	−0.82	−0.38

*Indices of *O. sativa* were calculated from corresponding STS loci in Caicedo *et al.* (2007).
[†]Silent site estimates include both synonymous and noncoding sites.
[‡]Average Tajima's *D* is based on the loci for which a Tajima's *D* can be calculated.

level in *O. rufipogon* would be higher than in *O. sativa*. The fact that the majority of the biallelic SNPs (763 of 800) in the total *rufipogon/sativa* complex are polymorphic in *O. rufipogon* reconfirmed the results of many previous studies that *O. rufipogon* is the ancestral gene pool for the cultivated rice at a genome-wide scale (Sun *et al.* 1997; Cheng *et al.* 2003; Londo *et al.* 2006; Sweeney *et al.* 2007; Tan *et al.* 2008; Zhang *et al.* 2009). Within *O. rufipogon*, Ruf-II had a higher genetic diversity (measured by π and θ) compared with Ruf-I, which is consistent with the results of the PCA analysis (Fig. 2a,b).

Another difference between our study and previous works (Garris *et al.* 2005; Caicedo *et al.* 2007; Reagon *et al.* 2010) is that the BAPS analysis did not separate out the five main variety groups of *O. sativa*, most likely due to the high genetic diversity in *O. rufipogon* in the combined analysis. A STRUCTURE analysis (Pritchard *et al.* 2000) of our data yielded a similar outcome, even with very high cluster numbers (K = 10 and 20, Fig. S3). In a hierarchical island model of genetic structure (Slatkin & Voelm 1991), high-level structure might cause Bayesian clustering programs to be insensitive to fine-scale structures (e.g. Evanno *et al.* 2005). Here, the larger genetic differences among *japonica* rice, *indica* rice and the two *O. rufipogon* groups probably obscured the minor differences among the smaller groups in the two cultivars.

Results of both Tajima's *D* and HKA tests provided no evidence of strong selection across the 42 STS loci as a whole, confirming that these STS loci follow neutral expectations and are appropriate for a phylogeographic study. On the other hand, the fact that Tajima's *D* values in most STS loci are negative (Table S2) also indicates a general trend towards excess high-/low-frequency SNPs (Tajima 1989a,b). Many factors may cause a negative *D* value, including positive/purifying selection, population growth, selective sweeps and gene flow from another population (Tajima 1989a; Fay & Wu 2000; Zeng *et al.* 2006). The most probable reasons for the widespread negative *D* values in *O. rufipogon* are population growth and/or gene flow, because the 42 STS loci are random markers (Caicedo *et al.* 2007), and they are both physically unlinked and functionally unrelated to each other (Table S2 and Fig. S1). Thus, it is implausible that a uniform selective regime caused the negative *D* values across the majority of the STS loci.

Phylogeography of O. rufipogon and its implications for rice domestication

The results of this study confirm the existence of two genetically distinct *O. rufipogon* groups, Ruf-I and Ruf-II. The north-eastern group, Ruf-I, has affinity to cultivated *O. sativa indica* based on BAPS clustering (but see below). In contrast, Ruf-II, mainly from South Asia and Indochinese peninsula, is not similar to cultivated rice varieties (Fig. 1). Neither of the two *O. rufipogon* groups shows close genetic similarity to *O. sativa japonica* as a whole, but a *japonica*-like component is found in a few Ruf-I individuals, mainly from China (Fig. 1 and Table S1). The fact that *O. nivara* accessions were not distinguished as a separate genetic group from *O. rufipogon* accessions verifies our assumption that *O. nivara* is in the same ancestral gene pool for domesticated rice.

It should be pointed out that the genetic composition of Ruf-I is probably to be more complicated than a simple resemblance to *indica* rice. Notably, Ruf-I is genetically much more variable than *indica* rice based on all measures of genetic diversity (Table 3). This trend is also obvious from the PCA plot, in which Ruf-I has a much wider spread than *O. sativa indica* (Fig. 2b). There are six *japonica* accessions showing different genetic background compared with the other *japonica* accessions based on the PCA plot (Fig. 2b). These individuals belong to the *aromatic* group (or Group V, Garris *et al.* 2005). Aromatic rice is placed within the broad sense *japonica*-type cultivars, but is known to have mixed genetic origins (Garris *et al.* 2005; Caicedo *et al.* 2007). Another general tendency observed from the PCA plot is that Ruf-I is generally more *sativa*-like compared with Ruf-II. This implies both cultivars are descendents of Ruf-I. The strong differentiation of *japonica* from the other groups (Figs 1c and 2b) may derive from the loss of ancestral alleles in *japonica* because of stronger bottlenecks or possibly more intense episodes of selection (Garris *et al.* 2005; Caicedo *et al.* 2007). Conversely, our BAPS clustering showed Ruf-II is not closely related to either of the cultivars, and in the PCA plot, Ruf-II is further away from *O. sativa* compared with Ruf-I. Accordingly, Ruf-II is probably to represent populations of extant *O. rufipogon* that were not intensely involved in the domestication of rice.

The notion that *O. rufipogon* is geographically structured is in accordance with most previous studies (Second 1982; Sun *et al.* 1997; Londo *et al.* 2006). However, the geographic pattern revealed in our study is very different from these previous results (Second 1982; Sun *et al.* 1997; Londo *et al.* 2006). First of all, the pattern of geographic variation is clinal from north-east to south-west, with the humid tropical plain areas in the Indochinese Peninsula being a transitional region where the two groups coexist. A clear geographic boundary between the two genetic groups was not observed. Second, deep divergence of *indica*-like and *japonica*-like *O. rufipogon* lineages was not observed in the *sativa*-like Ruf-I. This result shows that the *indica*-like and *japonica*-like divergence in *O. rufipogon*, which has been detected

in other studies using a small number of genes (Zhu & Ge 2005; Londo *et al.* 2006), does not emerge when large-scale sampling and markers are employed. Incomplete lineage sorting can be an important source of bias that causes phylogenetic incongruence (Knowles & Carstens 2007; Degnan & Rosenberg 2009), and this effect is stronger when fewer loci are sampled. As incomplete lineage sorting has already been reported in the genus *Oryza* (Zou *et al.* 2008), it is most likely to be the reason for the incongruent geographic pattern between our study and the others. Third, South Asia, previously thought to be one of the distribution centres of the *indica*-like *O. rufipogon*, shows a higher frequency of the more 'ancestral' Ruf-II genotypes rather than the *sativa*-like Ruf-I. This trend is especially prominent in the southern Himalayan regions.

The similarity between Ruf-I and *O. sativa* suggests that at least part of Ruf-I was probably the ancestral gene pool for rice domestication, whereas the involvement of Ruf-II in rice domestication is less likely. As Ruf-I is still found in the Indochinese Peninsula and South Asia at low frequency, these results do not rule out the possibility that *O. rufipogon* from this geographic region contributed to rice domestication. However, the area around China, which has the highest frequency of the *sativa*-like Ruf-I, is a more likely centre of rice domestication. A domestication scenario consistent with our results is that a primitive cultivar was originally domesticated from the Ruf-I group in China, and both *indica* and *japonica* rice are its descendents. Ancestral polymorphism in *japonica* rice was largely swept out by both intense selection and strong bottlenecks, leaving only slight traces of its similarity to Ruf-I, whereas in *indica* rice, more ancestral polymorphism is maintained. Supporting evidence for this hypothesis comes from the previous study that provided the STS data for *O. sativa* used here (Caicedo *et al.* 2007). The demographic modelling result from Caicedo *et al.* showed that the effects of selective sweeps were necessary, in addition neutral bottlenecks, to sufficiently explain the observed SNP frequency spectrum of both *indica* and *japonica*. Also, *japonica* showed much more derived high-frequency SNPs compared with *indica* (Caicedo *et al.* 2007), again indicating a stronger effect of selection in *japonica*. *Indica* rice, under this overall scenario, could have been brought westward along the Silk Road as a primitive cultivar and may have experienced weaker selection compared with that of *japonica, thus* retaining more Ruf-I alleles. Such a scenario is consistent with archaeological studies (Fuller *et al.* 2010). This result is in accordance with the single-origin hypothesis of cultivated rice, which is supported by some recent genetic (Molina *et al.* 2011) and archaeological (Fuller & Sato 2008; Fuller *et al.* 2009) work.

Historical causes of the clinal variation pattern in O. rufipogon

The historical causes of the north-east–south-west clinal variation in the two groups of *O. rufipogon* can be considered in the light of three possible mechanisms. The most intuitive hypothesis is that the clinal variation is caused by the drift–migration equilibrium ('equilibrium' hypothesis) between the Ruf-I group in China and the Ruf-II group in South Asia. Fine-scale population genetic studies of *O. rufipogon* generally show a relatively high F_{ST} (0.15 in Xu *et al.* 2006; 0.39–0.47 in Gao *et al.* 2002), indicating structured populations at local scale and restricted gene flow among populations. Additionally, as has been pointed out in several studies, *O. rufipogon* has a mixed mating system and in some cases can have high levels of self-pollination (Gao *et al.* 2002; Xu *et al.* 2006). Vegetative propagation is also very common (Grillo *et al.* 2009), which can be spatially restricted within a local body of water. All of these factors could limit the effective distance of gene flow in *O. rufipogon* populations. Thus, genetic drift or local adaptation may contribute to the pattern of geographic subdivision in *O. rufipogon* populations.

The second possible explanation for a clinal pattern of variation is the 'secondary contact' hypothesis. During the last glacial maximum (LGM, 15 000–26 500 years ago, Ray & Adams 2001; Clark *et al.* 2009), lowered sea levels resulted in an increase in the land area of East Asia and also markedly different climatic conditions compared with today's climate in the region (Ray & Adams 2001). The monsoon forests of the Indochinese Peninsula area were divided into eastern and western regions by intervening tropical savanna vegetation, which corresponds to the Cambodia, Thailand, area today (Ray & Adams 2001). Aquatic-living *O. rufipogon* is not probably to survive in the dry savanna vegetation regions because of its high water dependency, but it can survive in some monsoon forest regions (e.g. present day southern China and northern India). Thus, it is reasonable to assume that during the LGM, *O. rufipogon* was divided into two parts, southern China and south of the Himalayas. This geographic vicariance would lead to the establishment of Ruf-I and Ruf-II as two genetically differentiated groups. With the retreat of the glaciers beginning about 17 000 years ago (Clark *et al.* 2009), sea level rose and the climate of the Indochinese Peninsula changed to suitable conditions for *O. rufipogon*, similar to today's climate. This newly opened area would have suitable habitats for migrants from both southern China and South Asian populations, creating a mixed Ruf-I and Ruf-II pattern. Also, a growing population size could account for the generally negative Tajima's *D* observed in our sample.

Third, introgressive hybridization between *O. rufipogon* and *O. sativa* may also contribute to the clinal variation pattern (the 'introgression' hypothesis). The rice domestication system is complicated because the distribution range of the rice progenitor is embedded within the traditional range of cultivation. Furthermore, the traditional range of cultivation became subsequently embedded within a much larger, modern range of cultivation (Diamond 2002; Vaughan *et al.* 2008). This temporal and spatial overlap creates ample opportunity for gene flow between the crop and progenitor species. Both historical (Molina *et al.* 2011) and on-site (Song *et al.* 2003; Chen *et al.* 2004) gene flow between *O. sativa* and *O. rufipogon* have been documented, and hybrids between wild and cultivated rice are fertile (Niruntrayakul *et al.* 2009). At a larger scale, crossability has been observed among all AA genome species (including *O. rufipogon*, *O. sativa* and a few other species) in the genus *Oryza* (Khush 1997). Demographic reconstruction studies of our STS data also show a relatively high level of gene flow between *O. sativa* and *O. rufipogon* (data not shown). In fact, one possible explanation for the lowered overall genetic diversity and more negative Tajima's *D* in Ruf-I compared with Ruf-II (Tables 3 and S2) is the influx of derived high-frequency SNPs from *O. sativa*. If Ruf-I is a product of introgressions from *O. sativa*, with a longer domestication history, there would have been greater opportunities for the local *O. rufipogon* populations to gain alleles from *O. sativa* through introgression, leading to a higher proportion of Ruf-I in the local *O. rufipogon* populations. Again this process would help establish a cline if the original rice domestication was in China.

Finally, these three hypotheses are not mutually exclusive. The time period of 'secondary contact' after LGM (more recent than 17 000 years ago, Ray & Adams 2001; Clark *et al.* 2009) overlaps with the time of rice domestication estimated from genetic data (8200–13 500 years ago, Molina *et al.* 2011), allowing introgression to reinforce the formation of the geographic cline. And equilibrium between migration and genetic drift can dynamically maintain this clinal variation pattern.

Gene flow between O. rufipogon and *O. sativa*

As pointed out by several researchers (Song *et al.* 2003; Vaughan *et al.* 2008; Molina *et al.* 2011), gene flow between *O. rufipogon* and *O. sativa* might be a key confounding factor that obscures the true phylogenetic relationships in the rice domestication system. It should be noticed that there are two types of gene flow: the historical gene flow, which is the continuous gene exchange between *O. rufipogon* and *O. sativa* populations throughout the history of domestication, vs. occasional recent hybridization events that could happen during the germplasm maintenance process (Zhu & Ge 2005; Vaughan *et al.* 2008). One important point we would like to emphasize here is that the long-term historical gene flow, which potentially greatly influenced the population structure and phylogeography of *O. rufipogon*, should be regarded as part of the history of *O. rufipogon* instead of being excluded arbitrarily. The two major analysis methods we used in this study, BAPS and PCA, both require no prior knowledge of the species identity of individuals (Corander & Tang 2007; Stacklies *et al.* 2007). Accordingly, the result should be robust to different species delimitations criteria and gene flow, although how these results are interpreted could vary (see above discussions). On the other hand, recent accidental hybridization that could happen at a germplasm centre is a confounding factor and needs to be accounted for. We examined our *O. rufipogon* samples morphologically to avoid potential hybrids with many *O. sativa* traits. Also, a detectable phylogeographic signal itself is proof that such confounding effects are not strong; otherwise, a geographically randomly distributed *sativa*-like *O. rufipogon* should be observed because of recent hybridizations.

Conclusions

In summary, our genome-wide survey of SNP polymorphism reveals a new phylogeographic pattern within *O. rufipogon*. The two genetic groups of *O. rufipogon*, Ruf-I and Ruf-II, show a generally north-east–south-west clinal variation pattern across its native range in Asia. The *sativa*-like Ruf-I is centred in southern China; Ruf-II, with little similarity to cultivated rice, is mainly distributed in South Asia; while in the Indochinese Peninsula, the two groups coexist. Our results are consistent with the hypothesis of a single origin of domesticated *O. sativa* in China. Three potential factors, including the palaeoclimatic conditions, introgression from *O. sativa* to *O. rufipogon* and migration–drift balance over the native range of *O. rufipogon*, may contribute to the clinal variation pattern we observe.

Acknowledgements

We thank S.C. Hsu and L. Small for laboratory advice, K.M. Olsen, B.R. Lu (Fudan University), I. Jiménez (Missouri Botanical Garden), Y.C. Chiang (National Sun Yat-sen University), S.C. Hsu and P. Tonapha for inspiring discussions and advice, Washington University greenhouse staff for plant care, I. Jiménez for R instructions, and the B.A. Schaal and K.M. Olsen laboratory members and three anonymous reviewers for valuable comments and editing advice. Funding for this project was provided by the National Science Foundation, Plant Genome Research Project 0701382.

References

Buckley D (2009) Toward an organismal, integrative, and iterative phylogeography. *BioEssays*, **31**, 784–793.

Caicedo AL, Williamson SH, Hernandez RD *et al.* (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genetics*, **3**, 1745–1756.

Chen LJ, Lee DS, Song ZP, Suh HS, Lu BR (2004) Gene Flow from Cultivated Rice (*Oryza sativa*) to its Weedy and Wild Relatives. *Annals of Botany*, **93**, 67–73.

Cheng C, Motohashi R, Tsuchimoto S, Fukuta Y, Ohtsubo H, Ohtsubo E (2003) Polyphyletic origin of cultivated rice: based on the interspersion pattern of SINEs. *Molecular Biology and Evolution*, **20**, 67–75.

Chou CL (1948) China is the place of origin of rice. *Zhong Guo Dao Zuo*, **7**, 53–54. (in Chinese).

Clark PU, Dyke AS, Shakun JD *et al.* (2009) The Last Glacial Maximum. *Science*, **325**, 710–714.

Corander J, Marttinen P (2006) Bayesian identification of admixture events using multi-locus molecular markers. *Molecular Ecology*, **15**, 2833–2843.

Corander J, Tang J (2007) Bayesian analysis of population structure based on linked molecular information. *Mathematical Biosciences*, **205**, 19–31.

Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modeling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, **9**, 539.

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, **24**, 332–340.

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature*, **418**, 700–707.

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell*, **127**, 1309–1321.

Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, **316**, 1862–1866.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

Fuller DQ, Sato Y (2008) *Japonica* rice carried to, not from, Southeast Asia. *Nature Genetics*, **40**, 1264–1265.

Fuller DQ, Qin L, Zheng Y *et al.* (2009) The domestication process and domestication rate in rice: spikelet bases from the Lower Yangtze. *Science*, **323**, 1607–1610.

Fuller DQ, Sato Y, Castillo C *et al.* (2010) Consilience of genetics and archaeobotany in the entangled history of rice. *Archaeological and Anthropological Sciences*, **2**, 115–131.

Gao LZ (2004) Population structure and conservation genetics of wild rice *Oryza rufipogon* (Poaceae): a region-wide perspective from microsatellite variation. *Molecular Ecology*, **13**, 1009–1024.

Gao LZ, Schaal BA, Zhang CH, Jia JZ, Dong YS (2002) Assessment of population genetic structure in common wild rice *Oryza rufipogon* Griff. using microsatellite and allozyme markers. *Theoretical and Applied Genetics*, **106**, 173–180.

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. *Genetics*, **169**, 1631–1638.

Grillo MA, Li C, Fowlkes AM *et al.* (2009) Genetic architecture for the adaptive origin of annual wild rice, *Oryza nivara*. *Evolution*, **63**, 870–883.

Gross BL, Olsen KM (2010) Genetic perspectives on crop domestication. *Trends in Plant Science*, **15**, 529–537.

Gross BL, Skare KJ, Olsen KM (2009) Novel *Phr1* mutations and the evolution of phenol reaction variation in US weedy rice (*Oryza sativa* L.). *New Phytologist*, **184**, 842–850.

Heun M, Schäfer-Pregl R, Klawan D *et al.* (1997) Site of einkorn wheat domestication identified by DNA fingerprinting. *Science*, **278**, 1312–1314.

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Kane NC, Rieseberg LH (2008) Genetics and evolution of weedy *Helianthus annuus* populations: adaptation of an agricultural weed. *Molecular Ecology*, **17**, 384–394.

Khush GS (1997) Origin, dispersal, cultivation and variation of rice. *Plant Molecular Biology*, **35**, 25–34.

Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. *Systematic Biology*, **56**, 887–895.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Londo JP, Schaal BA (2007) Origins and population genetics of weedy red rice in the USA. *Molecular Ecology*, **16**, 4523–4535.

Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA (2006) Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proceedings of the National Academy of Sciences, USA*, **103**, 9578–9583.

Long Y, Zhao L, Niu B *et al.* (2008) Hybrid male sterility in rice controlled by interaction between divergent alleles of two adjacent genes. *Proceedings of the National Academy of Sciences, USA*, **105**, 18871–18876.

Lu BR, Zheng KL, Qian HR, Zhuang JY (2002) Genetic differentiation of wild relatives of rice as assessed by RFLP analysis. *Theoretical and Applied Genetics*, **106**, 101–106.

Matsuoka Y, Mitchell SE, Kresovich S, Goodman M, Doebley J (2002) Microsatellites in Zea—variability, patterns of mutations, and use for evolutionary studies. *Theoretical and Applied Genetics*, **104**, 436–450.

Molina J, Sikora M, Garud N *et al.* (2011) Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences, USA*, **108**, 8351–8356.

Niruntrayakul S, Rerkasem B, Jamjod S (2009) Crossability between cultivated rice (*Oryza sativa*) and common wild rice (*O. rufipogon*) and characterization of F1 and F2 populations. *ScienceAsia*, **35**, 161–169.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www,R-project,org.

Ray N, Adams JM (2001) A GIS-based vegetation map of the world at the last glacial maximum (25,000-15,000 BP).

Internet Archaeology, **11**. Availablae at http://intarch.ac.uk/journal/issue11/rayadams_toc.html.

Reagon M, Thurber CS, Gross BL, Olsen KM, Jia Y, Caicedo AL (2010) Genomic patterns of nucleotide diversity in divergent populations of U.S. weedy rice. *BMC Evolutionary Biology*, **10**, 180.

Second G (1982) Origin of the genetic diversity of cultivated rice (*Oryza* spp.): study of the polymorphism scored at 40 isozyme loci. *Japanese Journal of Genetics*, **57**, 25–57.

Second G, Morishima H (1981) Geographical and ecological pattern of variation of *Oryza perennis* at 7 isozyme loci. *National Institute of Genetics, Mishima, Japan*, 1980 Annual Report, 95–96.

Slatkin M, Voelm L (1991) F_{st} in hierarchical island model. *Genetics*, **127**, 627–629.

Song ZP, Lu BR, Zhu YG, Chen JK (2003) Gene flow from cultivated rice to the wild species *Oryza rufipogon* under experimental field conditions. *New Phytologist*, **157**, 657–665.

Stacklies W, Redestig H, Scholz M, Walther D, Selbig J (2007) pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164–1167.

Stephens M, Donelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, **73**, 1162–1169.

Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Sun CQ, Wang XK, Yoshimura A, Iwata N (1997) RFLP analysis of nuclear DNA in common wild rice (*Oryza rufipogon* Griff.) and cultivated rice (*Oryza sativa* L.). *Scientia Agricultura Sinica*, **30**, 37–44 (in Chinese).

Sweeney MT, McCouch SR (2007) The complex history of the domestication of rice. *Annals of Botany*, **100**, 951–957.

Sweeney MT, Thomson MJ, Cho YG *et al.* (2007) Global dissemination of a single mutation conferring white pericarp in rice. *PLoS Genetics*, **3**, 1418–1424.

Tajima F (1989a) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tajima F (1989b) The effect of change in population size on DNA polymorphism. *Genetics*, **123**, 597–601.

Tanaka T, Antonio BA, Kikuchi S *et al.* (2008) The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Research*, **36**, D1028–D1033.

Tan L, Li X, Liu F *et al.* (2008) Control of a key transition from prostrate to erect growth in rice domestication. *Nature Genetics*, **40**, 1360–1364.

Vaughan DA, Lu BR, Tomooka N (2008) The evolving story of rice evolution. *Plant Science*, **174**, 394–408.

Wang RL, Hey J (1996) The speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the period locus. *Genetics*, **144**, 1113–1126.

Wang RL, Stec A, Hey J, Lukens L, Doebley J (1998) The limits of selection during maize domestication. *Nature*, **398**, 236–239.

Wang MX, Zhang HL, Zhang DL *et al.* (2008) Genetic structure of *Oryza rufipogon* Griff. in China. *Heredity*, **101**, 527–535.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Xu X, Lu BR, Chen YH *et al.* (2006) Inferring population history from fine-scale spatial genetic analysis in *Oryza rufipogon* (Poaceae). *Molecular Ecology*, **15**, 1535–1544.

Yuan LP, Virmani SS, Mao CX (1989) Hybrid rice—achievement and outlook. In: *Progress in Irrigated Rice Research*, pp. 219–235. International Rice Research Institute, P. O. Box 933, Manila, Philippines.

Zeng K, Fu YX, Shi S, Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, **174**, 1431–1439.

Zhang LB, Zhu Q, Wu ZQ *et al.* (2009) Selection on grain shattering genes and rates of rice domestication. *New Phytologist*, **184**, 708–720.

Zhou HF, Xie ZW, Ge S (2003) Microsatellite analysis of genetic diversity and population genetic structure of a wild rice (*Oryza rufipogon* Griff.) in China. *Theoretical and Applied Genetics*, **107**, 332–339.

Zhu Q, Ge S (2005) Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. *New Phytologist*, **167**, 249–265.

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Molecular Biology and Evolution*, **24**, 875–888.

Zou XH, Zhang FM, Zhang JG *et al.* (2008) Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biology*, **9**, R49.

This research is conducted as part of P.H.'s PhD dissertation research project at Washington University in St Louis. His main research interests include population genetics, biogeography and molecular evolution in plants. J.M., J.M.F., S.R. and M.D.P. are interested in plant evolutionary and ecological genomics, including the origin and spread of domesticated species. S.A.J. is interested in genomics of crop plants. B.A.S.'s research interests center on the evolutionary genetics of plants.

Data accessibility

We made the original data (sequence alignments in FASTA format of 42 STS loci in 96 *O. rufipogon* samples) available as a zipped file at Dryad entry doi:10.5061/dryad.rq6874m6.

Supporting information

Additional supporting information may be found in the online version of this article.

Additional data, including accession information, STS loci information and STRUCTURE clustering results, can be found online at ME website.

Table S1 Sample information of the *O. rufipogon* and *O. sativa* accessions.

Table S2 Loci information of the 42 STS.

Table S3 Fisher's exact test for association between genetic clusters and geographic regions.

Fig. S1 Genome location of 42 STS loci. Each star sign shows the genomic location of one STS locus.

Fig. S2 Delimitation of the four geographic regions.

Fig. S3 Bayesian clustering using STRUCTURE in comparison with BAPS. Clustering method and the assumed numbers of clusters (K) used are shown on the left side of the figure. The variety/species affiliation of each individual is indicated on top of the figure.

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