Natural Selection in Gene-Dense Regions Shapes the Genomic Pattern of Polymorphism in Wild and Domesticated Rice

Jonathan M. Flowers,[1] Jeanmaire Molina,[1] Samara Rubinstein,[1] Pu Huang,[2] Barbara A. Schaal,[2] and Michael D. Purugganan[1],*

[1]Department of Biology and Center for Genomics and Systems Biology, New York University
[2]Department of Biology, Washington University
*Corresponding author: E-mail: mp132@nyu.edu.
Associate editor: Michael Nachman

Abstract

Levels of nucleotide variability are frequently positively correlated with recombination rate and negatively associated with gene density due to the effects of selection on linked variation. These relationships are determined by properties that frequently differ among species, including the mating system, and aspects of genome organization such as how genes are distributed along chromosomes. In rice, genes are found at highest density in regions with frequent crossing-over. This association between gene density and recombination rate provides an opportunity to evaluate the effects of selection in a genomic context that differs from other model organisms. Using single-nucleotide polymorphism data from Asian domesticated rice *Oryza sativa* ssp. *japonica* and ssp. *indica* and their progenitor species *O. rufipogon*, we observe a significant negative association between levels of polymorphism and both gene and coding site density, but either no association, or a negative correlation, between nucleotide variability and recombination rate. We establish that these patterns are unlikely to be explained by neutral mutation rate biases and demonstrate that a model of background selection with variable rates of deleterious mutation is sufficient to account for the gene density effect in *O. rufipogon*. In *O. sativa* ssp. *japonica*, we report a strong negative correlation between polymorphism and recombination rate and greater losses of variation during domestication in the euchromatic chromosome arms than heterochromatin. This is consistent with Hill–Robertson interference in low-recombination regions, which may limit the efficacy of selection for domestication traits. Our results suggest that the physical distribution of selected mutations is a primary factor that determines the genomic pattern of polymorphism in wild and domesticated rice species.

Key words: background selection, deleterious mutation, selfing, hitchhiking, interference, Hill–Robertson, domestication.

Introduction

How nucleotide variation is distributed across the genome is a central question in population genetics that has been a subject of considerable interest since the earliest surveys of molecular polymorphism. Interest in this question arises from the insights it provides into the frequency and strength of natural selection and the extent to which it shapes genomic patterns of polymorphism (Andolfatto 2001; Wright and Andolfatto 2008). A significant breakthrough came in 1992, when Begun and Aquadro reported a positive correlation between nucleotide polymorphism and recombination rate in *Drosophila melanogaster* (Begun and Aquadro 1992). This association between crossover frequency and the level of nucleotide polymorphism, but not divergence, indicated that natural selection is a primary force operating over a large fraction of the Drosophila genome and that nonselective mechanisms such as neutral mutation rate variation cannot fully account for differences in nucleotide polymorphism in different genomic regions (Andolfatto 2001; Wright and Andolfatto 2008).

Since the initial discovery in *D. melanogaster*, associations between polymorphism and nucleotide diversity attributable to linked selection have been reported in humans (Nachman 1998; Przeworski et al. 2000; Cai et al. 2009; see

also Lercher and Hurst 2002; Hellman et al. 2003), *Caenorhabditis elegans* (Koch et al. 2000; Cutter and Payseur 2003), mice (Nachman 1997; Takahashi et al. 2004), and other species (e.g., Begun et al. 2007; Cutter and Choi 2010; Cutter and Moses 2011). This correlation has been interpreted in the context of population genetic theory, which predicts that the fate of neutral polymorphism is coupled to that of selected mutations through the effects of linkage. In regions of low recombination, natural selection reduces polymorphism at linked sites due to a reduction in the local effective population size and a concurrent increase in the probability of fixation of neutral variants. This loss of nucleotide variation in regions of suppressed recombination is predicted by theoretical models that incorporate either deleterious ("background selection"; Charlesworth et al. 1993) or advantageous mutations ("genetic hitchhiking"; Maynard-Smith and Haigh 1974; Kaplan et al. 1989), and both models predict a positive association between polymorphism and the rate of crossing-over when the frequency and intensity of selection is uniform across the genome (Wiehe and Stephan 1993; Hudson and Kaplan 1995; Nordborg et al. 1996).

Although this correlation is consistent with both background selection and hitchhiking models, an association between polymorphism and recombination rate may

not be expected when rates of deleterious or advantageous mutation are not uniform. A more general prediction is that the level of variation will be proportional to the density of selected mutations per genetic map unit (Barton 1995; Hudson and Kaplan 1995; Nordborg et al. 1996). Therefore, reductions in nucleotide variability associated with selection may not be restricted to regions of suppressed recombination but may be apparent in any region where the density of selected mutations is high relative to the local rate of recombination. Thus, gene density or the density of other genomic features that are common targets of selection may better predict the level polymorphism than the recombination rate. This may explain why some species including *Arabidopsis thaliana* show a negative association between polymorphism and gene density but no relationship between polymorphism and recombination rate (Nordborg et al. 2005).

How standing variation responds to selection therefore depends on properties of the genome including the distribution of genes along recombining chromosomes, aspects of the species demography and life history such as the degree of population substructure and the outcrossing rate, and the history of selection. Characterizing how nucleotide variation is distributed across the genome in species that differ in these respects may provide insight into how selection influences genome evolution and the nature of selection itself. Here, we report on resequencing data from two domesticated subspecies of Asian cultivated rice, *Oryza sativa* ssp. *indica* and *japonica*, and their wild progenitor *O. rufipogon*. *Oryza sativa* has experienced intense population bottlenecking (Rakshit et al. 2007), recurrent episodes of artificial selection (Caicedo et al. 2007; He et al. 2011), and a recent (i.e., <9,000 years) transition to self-fertilization during the process of domestication. In contrast, *O. rufipogon* is a wild, self-compatible, primarily outcrossing species with historically large effective size ($N_e > 250,000$; Molina et al. 2011) and an uncertain history of selection. An interesting feature of the rice genome is that genes are organized along recombination gradients with genes found in highest densities in regions with frequent crossing-over (Tian et al. 2009). This organizational property differs notably from species with holocentric chromosomes such as *C. elegans* and could provide insight into the effects of selection in organisms with different chromosome structures.

Materials and Methods

Resequencing Data and Population Genetic Statistics

The population samples and resequencing data for the 552 gene fragments (referred to as sequence-tagged sites or STS) are reported in Molina et al. (2011) and the data available at http://puruggananlab.bio.nyu.edu/Rice_data/. These sequenced regions consist almost exclusively of small intron sequences and the immediately flanking exons. Population summary statistics including measures of nucleotide diversity (i.e., θ_π, Tajima 1983 and θ_W, Watterson 1975) were computed for silent sites (i.e., intergenic, intron,

5′ and 3′ UTRs, and synonymous sites) for each population using the libsequence package (Thornton 2003). Linkage disequilibrium was calculated using a composite likelihood method with estimates of r^2 calculated after removing doubly heterozygous genotypes with unknown phase following the methodology in Mather et al. (2007). Haplotype estimates used in estimating divergence between *O. rufipogon* and *O. meridionalis* (see below) were constructed with PHASE v. 2.1.1 (Stephens et al. 2001; Stephens and Scheet 2005).

Nucleotide polymorphism estimates accounted for missing data by weighting each site by the proportion of sequences with unambiguous base calls using the following equations:

$$\pi = \sum_{i=1}^{i=S} \left(1 - \sum_{j=1}^{j=4} \frac{k_{j,i} \times (k_{j,i} - 1)}{n_i \times (n_i - 1)} \right); \quad k_{j,i} > 0, \qquad (1)$$

where S is the number of segregating sites, $k_{j,i}$ is the number of occurrences of the jth character state at site i, and n_i is the sample size at site i (Thornton 2003).

$$\theta_W = \sum_{i=1}^{i=S} \frac{S}{\sum_{j=1}^{j=n_i-1} \frac{1}{j}}, \qquad (2)$$

where S is the number of segregating sites and n_i is the sample size at site i (Thornton 2003).

The substitution rate, K_{sil}, was estimated between pairs of sequences by treating each noncoding site as a full site and synonymous sites as fractions of sites using the method of Nei and Gojobori (1986) followed by a Jukes–Cantor correction (Jukes and Cantor 1969). Estimates were obtained from the *O. sativa* Nipponbare–*O. meridionalis* comparison as well as the average pairwise difference between gametic phase-resolved *O. rufipogon* sequences and *O. meridionalis*. Our analyses are based on the gene predictions in the Rice Genome Sequence Annotation v. 6.0 (Ouyang et al. 2006) of the *O. sativa* spp. *japonica* cv. Nipponbare genome (International Rice Genome Sequencing Project 2005). These measures of divergence are highly correlated (Spearman's $r = 0.93$, $P < 1.0 \times 10^{-15}$) and yielded comparable results; we present only the results for the Nipponbare–*O. meridionalis* comparison. All analyses of divergence are based on a reduced data set ($n = 439$) consisting of loci from which we were able to obtain an *O. meridionalis* sequence. The average divergence between Nipponbare and *O. meridionalis* was 2.6%, which on average is approximately 3.5-fold higher than diversity (π_{sil}) within *O. rufipogon* (mean = 0.0071). Given this moderate level of divergence, it is plausible that some estimates of K_{sil} will be affected by shared ancestral polymorphism.

Genomic Features

Gene and codon densities were determined based on gene predictions in the Rice Genome Sequence Annotation v. 6.0 of the *O. sativa* spp. *japonica* cv. Nipponbare genome sequence (Ouyang et al. 2006). Gene density was determined by counting the number of protein-coding genes

(excluding transposable elements and pseudogenes), or parts of genes, in windows centered on the STS. Because the appropriate physical widths of these windows are unclear, we report statistical results for window sizes ranging from 100 kb to 2 Mb. Incomplete windows at the ends of chromosomes were excluded from these analyses. Estimates of codon density were obtained by counting the number of codons in each window. Intergenic GC content was determined for nonoverlapping windows in the MSU v. 6.0 pseudomolecules (Ouyang et al. 2006).

Recombination rates were obtained using markers from the Rice Genome Project (RGP) (Harushima et al. 1998) that have been anchored to the rice pseudomolecules used in this study (Ouyang et al. 2006). Markers and their genetic and physical positions were downloaded from the MSU website (http://rice.plantbiology.msu.edu) and then filtered to remove those that mapped to multiple physical positions and to remove those whose physical anchor and genetic location were on different chromosomes. Estimates of recombination were then obtained with either a sliding-window or a polynomial approach. The sliding-window approach was implemented with the MareyMap R package (Rezvoy et al. 2007) with window size of 4 Mb and stepsize of 200 kb. Local recombination rates were then estimated from the slope of locally fitted linear equations. The polynomial approach was implemented by splicing three cubic equations, which were each fit separately to the arms and pericentromeric region of the three chromosomes. These different measures of recombination rate were highly correlated, and results were broadly similar using either method. Results are presented for the sliding-window method unless otherwise mentioned.

Analysis of the CpG Effect

Noncoding single-nucleotide polymorphisms (SNPs) in *O. rufipogon* were polarized using parsimony with *O. meridionalis* and *O. barthii* as outgroups and assigned to one of the six transition:transversion (ts:tv) classes of mutation. SNPs were assigned to each of the six ts:tv categories by inferring the ancestral state at each polymorphic site with parsimony. The ancestral state was inferred and an SNP included in the analyses if both *O. meridionalis* and *O. barthii* sequences were homozygous for the same allele and if the allele was also segregating in *O. rufipogon*. The Poisson regression analyses were conducted by either assuming independence of CpG and CpH site classes sampled from the same STS or by including only one class of sites from each STS. In all analyses, SNP counts were treated as the response variable and the number of available sites incorporated as an offset in the model. Overdispersion of residual deviances was a consistent feature of our analyses, so each model was refitted with quasipoisson errors and tests of nested models conducted with *F*-tests (Crawley 2007, p. 542).

Specification of Chromosomal Regions

The locations of the centromeres were based on the physical positions of genetic markers that flank a region with no recombinants in the 186 F_2 individuals screened in the Nipponbare × Kasalath cross (Harushima et al. 1998). The centromeres in our analysis are located at 54.3, 15.7, and 51.5 cM on the genetic maps of chromosomes 8, 10, and 12 and correspond to the *Cen8*, *Cen10*, and *Cen12* designations of Yan et al. (2005, 2008). The physical boundaries of these regions on the MSU v. 6.0 pseudomolecules are provided in the footnote to supplementary table S1 (Supplementary Material online). To explore the causes of regional differences in polymorphism and divergence, chromosomes were also divided broadly into pericentromeric regions (i.e., areas of low recombination including the centromere: 8, 9,413,442–15,766,428; 10, 6,125,059–9,704,963; 12, 7,551,430–14,887,432) and the chromosome arms.

Statistical Tests

Tests of differences between regions (e.g., pericentromeres vs. arms) were performed with two-tailed Wilcoxon rank sum tests. Correlations were assessed with Spearman's tests and significance determined using the asymptotic *t* approximation in R (http://www.r-project.org). Partial correlations between two variables, *X* and *Y*, controlling for a third variable, *Z*, were performed with partial Spearman's correlations in R using the pcor.R protocol obtained from Soojin Yi at http://www.yilab.gatech.edu/pcor.R. Tests within the cultivated variety groups should be considered with some caution because possible spatial autocorrelation in *japonica* and *indica* will artificially increase the false positive rate. Spatial nonindependence of observations is unlikely to be problematic in *O. rufipogon* because linkage disequilibrium in wild rice decays to background levels by approximately 5 kb (Mather et al. 2007) (see below) and the median distance between adjacent STS fragments in our data set is 106 kb (interquartile range 88.5–179 kb). Moreover, all but one pair of STS fragments in our data set is separated by greater than 50 kb. Poisson regression tests of mutation biases were interpreted assuming independence of observations in *O. rufipogon*.

In addition to the Spearman's correlations, a nonparametric approach that accounts for possible autocorrelation among observations in *O. sativa* was applied to test for correlations in the domesticated cultivars. This approach is based on a method that maintains the original autocorrelation structure in the data by preserving the physical order of the observations on each chromosome arm while randomizing two variables (e.g., nucleotide diversity and gene density) with respect to each other (Nordborg et al. 2005). For each set of ordered observations (i.e., each chromosome arm), the procedure randomly sets the polarity of the arm by orienting the observations as a unit in either the forward or the reverse direction. The arms are then joined in random order, while maintaining the order of observations within each arm, and then circularized. Two circles, one for each set of observations whose correlation is to be tested, are then randomly oriented with respect to each other. Spearman's rank correlations were

then calculated for the data set and the procedure repeated 10,000 times to generate a null distribution.

Background Selection Model

Models of background selection with recombination frequently assume a uniform rate of deleterious mutation across the genome and therefore predict a monotonic increase in polymorphism with increasing recombination rate (Hudson and Kaplan 1995; Nordborg et al. 1996). Since we found no association between the level of polymorphism and the frequency of crossing-over in *O. rufipogon* (see below), but instead observed a negative correlation with gene density, we modeled the effect of background selection on nucleotide variation by incorporating variable rates of deleterious mutation. Our aim was to determine if background selection is sufficient to account for the gene density effect in the wild progenitor of domesticated rice.

Our approach was to use densities of phylogenetically conserved sites (i.e., sites most susceptible to mutations that reduce fitness) to obtain coarse interval-specific rates of deleterious mutation (Rockman et al. 2010). Conserved sites were determined using phastCons predictions (Siepel et al. 2005; available for rice chromosomes at http://www.ricemap.org; Wang et al. 2011), and interval-specific rates of deleterious mutation obtained assuming a diploid genomic rate of deleterious mutation (U; see below). Levels of neutral variation at each surveyed position on the chromosomes in our analysis were then modeled following Hudson and Kaplan (1995) in which the proportion of neutral variation remaining under background is defined as $\pi_{BGS}/\pi_o = e^{-G}$, where π_o is the diversity of a population free of selection and π_{BGS} is the level of variation remaining under background selection. G is the mean of the distribution of the number of deleterious mutations per chromosome (Charlesworth et al. 1993) and is defined according to equation (15) in Hudson and Kaplan (1995) with modifications to accommodate partial outcrossing (Rockman et al. 2010) as

$$G_k = \sum_i \frac{u_i sh}{2(sh + P|M_k - M_i|)(sh + P|M_k - M_{i+1}|)}.$$
(3)

In this equation, sh is a compound parameter representing the strength of selection against deleterious mutations and dominance, which is assumed to be constant across loci. P is a panmixis index equal to $1 - (1 - r_{out})/(1 + r_{out})$, where r_{out} is the outcrossing rate, derived from the equilibrium inbreeding coefficient (Charlesworth et al. 1997; Rockman et al. 2010). P is incorporated here to rescale the genetic distances between sites to account for partial outcrossing and population substructure in *O. rufipogon* (e.g., Gao et al. 2001). M_k and M_i are the map positions in Morgans at the focal site where neutral polymorphism is measured and the flanking intervals. The diploid deleterious mutation rate for each region, u_i, was approximated by multiplying U by the fraction of conserved sites in the genome that are found in

each interval. π_{BGS}/π_o is then calculated separately for each locus, k, which correspond to the genomic positions where we obtained an estimate of nucleotide diversity for *O. rufipogon* in our resequencing study.

Since U, P, and sh are unknown, we surveyed a 25 × 25 grid of P and sh values evenly distributed in log space. A coarse estimate of U can be obtained by assuming a neutral mutation rate of 6.5×10^{-9} per site per generation (Gaut et al. 1996) and a genome size of 370 Mb (Ouyang et al. 2006). Given roughly 25 Mb of conserved sites in the rice genome from phastCons scores derived from alignments between three grass species (Wang et al. 2011), this yields an estimate of $U = 0.3$. This may underestimate the true rate since it assumes that only sites conserved within the grasses are susceptible to deleterious mutation, and the contribution of other classes of mutation, notably transposable elements (Charlesworth 1996), are ignored. An extreme upper bound can be placed at $U \sim 5$, given that $2 \cdot (3.7 \times 10^8 \text{ bp} \cdot 6.5 \times 10^{-9}$ mutations per base pair per generation) total point mutations are expected per diploid genome per generation. Given the uncertainty in these values, we aim to assess whether a set of plausible values can recover the observed negative correlation between polymorphism and gene density under background selection alone.

Results

Nucleotide Polymorphism and Divergence in Wild and Domesticated Rice

Approximately 268 kb of sequence data was collected from 552 protein-coding genes on chromosomes 8, 10, and 12 in regions that span the full range of recombination (0–8 cM/Mb) including the recombination-free centromeres, the pericentromeres, and the chromosome arms (Molina et al. 2011). The sequenced individuals included 36 mostly landrace accessions of the Asian domesticated rice, *O. sativa*, 20 accessions of Asian wild rice, *O. rufipogon*, and single outgroup accessions of *O. barthii* and *O. meridionalis*. Our panel of domesticated rice accessions consisted of representatives of two genetically distinct groups, *O. sativa* ssp. *indica* (20 accessions) and *O. sativa* ssp. *tropical japonica* (16 accessions), which have well-documented differences in levels of diversity (Caicedo et al. 2007) that are also reflected in our data (*O. rufipogon*, 5,906 SNPs > *indica*, 2,242 SNPs > *japonica*, 1,557 SNPs).

To assess the influence of variation in mutation rate and selective constraints on our data, we tested for associations between silent divergence, K_{sil}, measured from *O. sativa* Nipponbare and *O. meridionalis* and various genomic properties. Considering all genes in our analysis, K_{sil} was positively correlated with π_{sil} in *O. rufipogon* (Spearman's $r = 0.20$, $P < 0.00005$), recombination rate ($r = 0.14$, $P < 0.005$), and gene density (0.1 Mb windows, $r = 0.11$, $P < 0.02$; 1 Mb windows, $r = 0.22$, $P < 0.00001$) but not intergenic GC content ($r = -0.06$, $P > 0.05$). These associations appear to be primarily driven by lower divergence rates in the pericentromeres (pericentromeres vs. arms;

Wilcoxon rank sum test, $P < 0.01$), and only the correlation between π_{sil} and K_{sil} ($r = 0.19$, $P < 0.0005$) and an association between K_{sil} and gene density measured at larger window sizes (1 Mb windows, $r = 0.17$, $P < 0.005$) remain significant when pericentromeric genes are excluded from the analysis. Given this evidence for mutation rate effects on polymorphism, we present tests for associations between polymorphism and genomic features with a normalized measure of polymorphism (π_{sil}/K_{sil}), partial correlations ($\pi_{sil}|K_{sil}$), and unadjusted estimates of diversity below.

Absence of a Positive Correlation between Nucleotide Diversity and Recombination Rate

To estimate local recombination rates in rice, we used genetic markers from the RGP(Harushima et al. 1998) that have been anchored to the TIGR-MSU rice pseudomolecules used in this study (Ouyang et al. 2006). Estimates of recombination rate were then obtained with either a sliding-window or polynomial approach (see Materials and Methods); both methods gave similar estimates and both measures were strongly and positively correlated with both gene and coding site density (fig. 1), a property that may differ between cereals and Arabidopsis (table 1).

To characterize the factors influencing polymorphism, we partitioned the chromosomes broadly into the centromeres, pericentromeres, and chromosome arms (fig. 2). We first consider patterns of polymorphism on the mostly euchromatic arms and return to the low-recombination heterochromatic regions below. In *O. rufipogon*, bivariate tests for correlations between recombination rate and polymorphism found either no correlation ($r = -0.09$, $P = 0.050$; sliding-window method) or a negative association ($r = -0.16$, $P < 0.0005$; polynomial) depending on the method used to estimate the local recombination rate (see table 2 and fig. 3). A similar result was obtained for the chromosome arms of domesticated cultivars, where negative associations between recombination rate and levels of polymorphism were observed in both *indica* ($r = -0.18$, $P < 0.0005$) and *japonica* ($r = -0.31$, $P < 1.0 \times 10^{-10}$; table 2 and fig. 3). Significantly negative associations between nucleotide diversity and recombination remained in all three groups after controlling for substitution rate (table 2).

To better characterize the relationship between recombination and polymorphism in wild rice, *O. rufipogon*, we focused on very low-recombination regions in the centromeres and pericentromeres of chromosomes 8, 10, and 12. In rice, the centromeres are well characterized (Yan and Jiang 2007) and consist of unique centromeric repeat units, transposable elements, and actively expressed protein-coding genes at moderate densities (Yan et al. 2005, 2008). Recombination rates are extremely low in the centromeres (i.e., no recombinants in 186 F_2 hybrids in the Kasalath × Nipponbare cross; Harushima et al. 1998), thus providing an opportunity to test for reduced polymorphism when meiotic crossing-over is rare.

Seven of the genes in our data set are located in the "recombination-free" (Harushima et al. 1998; Yan et al. 2005) domain encompassing *Cen8* and five additional loci are in
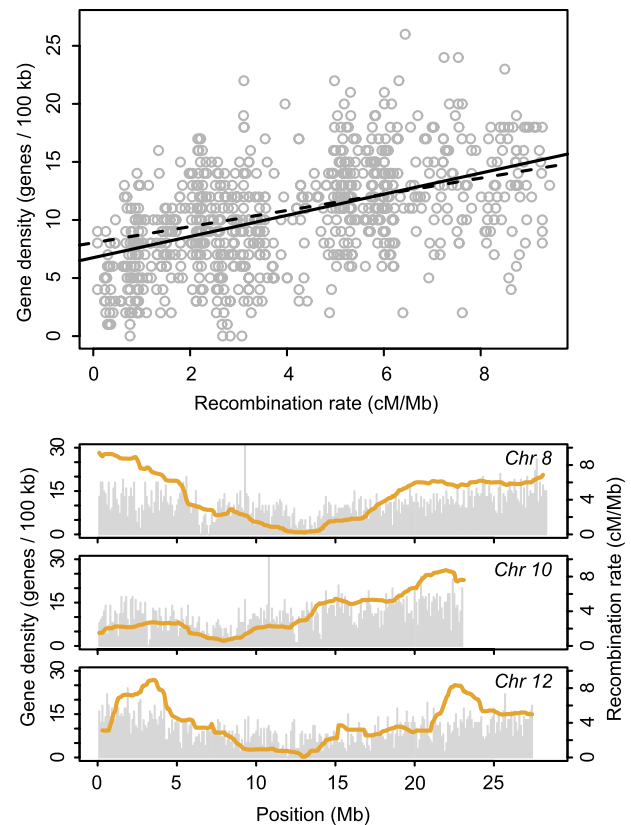

FIG. 1. Positive association between gene density and recombination rate on rice chromosomes 8, 10, and 12. (*A*) Estimates of recombination rate were made at 100-kb intervals along each chromosome with gene densities determined in 100-kb windows flanking each position. The solid line is the least squares fit to all data points. The broken line represents a fit to points from the chromosome arms only. Both associations are significant by Spearman's rank correlation (all regions, $r = 0.52$, $P < 1 \times 10^{-15}$; arms, $r = 0.36$, $P < 1 \times 10^{-15}$). (*B*) Gene density (vertical gray bars) and recombination rate (orange line) on each chromosome.

zero-recombination regions of *Cen10* and *Cen12* (see supplementary table S1, Supplementary Material online). All 12 centromeric loci are polymorphic with silent SNP densities being approximately 2-fold lower in the centromere (49 SNPs/3,147 silent sites = 0.016) compared with the chromosome arms (4,494/155,027 = 0.029). Estimates of diversity based on the number of segregating sites suggest that polymorphism in the centromere is reduced (0.0036 ± 0.0007, mean $\pm$ standard error of the mean [SEM]; supplementary table S1, Supplementary Material online) compared with noncentromeric regions (0.0075 ± 0.0003; Wilcoxon rank sum test; $P = 0.02$), although estimates based on π_{sil} were not ($P = 0.16$).

If we extend the very low-recombination regions to include the entire pericentromere, we do not find evidence of lower polymorphism relative to the chromosome arms using either measure of polymorphism (Wilcoxon rank sum test, $P > 0.05$). A test for reduced divergence with *O. meridionalis* in the centromere based on eight genes for which we could obtain an *O. meridionalis* sequence was not significant (Wilcoxon rank sum test, $P = 0.12$),

Table 1. Relationship between Recombination Rate and Gene (functional site) Density.

Species	Chromosomal Regions	Correlation (recombination vs. gene density)	Reference
Oryza ssp *japonica*[a]	Arms	Positive	Tian et al. (2009)
O. ssp *japonica*[b]	Arms	Positive	Tian et al. (2009)
O. ssp *japonica*[c]	All regions	Positive	Wu et al. (2002)
Maize[a]	All regions	Positive	Fengler et al. (2007)
Maize[c]	All regions	Positive	Anderson et al. (2006)
Maize[ac]	All regions	Positive	Liu et al. (2009)
Wheat[b]	3.1 Mb on chromosome 3	Positive	Saintenac et al. (2011)
Arabidosis thaliana[b]	All regions	Weakly positive	Wright et al. (2003)
A. thaliana[b]	Arms	Weakly negative	Wright et al. (2003)
A. thaliana[a]	Chromosome 4	No association	Drouaud et al. (2006)
Drosophila melanogaster[a]	All regions	Weakly positive	Hey and Kliman (2002)
D. melanogaster[a]	n.a.	No association	Rizzon et al. (2002)
Human[b]	n.a.	Weakly positive	Cai et al. (2009)
Human[a]	n.a.	Weakly positive	Kong et al. (2002)
Caenorhabditis elegans[ab]	n.a.	Negative	*C. elegans* Sequencing Consortium (1998)

NOTE.—n.a., not applicable.
[a] Gene density.
[b] Coding site density.
[c] EST density.

although reduced rates of substitution in *Cen8* has previously been reported (Ma and Bennetzen 2006). A more powerful test of divergences between the pericentromere (0.020 ± 0.002, mean $\pm$ SEM) and arms (0.029 ± 0.001) indicates that substitution rates are lower in the pericentromere (Wilcox rank sum test, $P = 0.006$). This observation coupled with the positive correlation between recombination and divergence suggests that recombination-associated, or possibly transcription-associated, mutation occurs in rice, but it is unclear if the reduction in polymorphism observed in centromeric genes can be explained solely by mutation effects or selection.

Significant Negative Correlation between Nucleotide Diversity and Gene Density

Although there was not a positive relationship between levels of diversity and recombination rate, tests for associations between π_{sil} and gene density on the chromosome arms of wild rice revealed significantly negative relationships at all scales over which we measured gene density (table 2, fig. 4, and supplementary fig. S2, Supplementary Material online). This negative correlation is driven by gene-dense regions of the chromosome arms that have

approximately two-thirds the level of nucleotide diversity as those in the low-density regions (Wilcoxon rank sum test, upper vs. lower gene density quartile, $P < 0.05$; fig. 4). The negative association between nucleotide diversity and gene density in *O. rufipogon* is stronger across all window sizes when we adopt a normalized measure of polymorphism (e.g., π_{sil}/K_{sil}; $r = -0.26$, 1 Mb windows, $P < 0.00001$; fig. 4) and remains significant in partial correlations that control for K_{sil} (all intervals, $P < 0.01$, table 2). This gene density effect is also apparent when analysis is restricted to regions of the arms with inferred recombination rates > 2 cM/Mb ($r = -0.14$, 0.1 Mb windows, $P < 0.005$; $r = -0.19$, 1 Mb windows, $P < 0.0001$), indicating that the pattern we observe is not driven by low-recombination rate regions. These statistical associations on the chromosome arms are qualitatively the same when data from the pericentromeres and arms are combined (supplementary fig. S2, Supplementary Material online).

In the domesticated cultivars, the association between polymorphism and gene density is stronger than in *O. rufipogon* in bivariate tests on the chromosome arms whether we normalize nucleotide diversity by the substitution rate or not (*japonica*, $P < 1 \times 10^{-5}$, *indica* $P < 1 \times 10^{-7}$, all window sizes; table 2). This relationship is also

Table 2. Spearman's Rank Correlations (r) between π_{sil} and Genomic Features on Rice Chromosome Arms.

	Oryza rufipogon		*O. sativa* ssp. *indica*		*O. sativa* ssp. *japonica*	
	π_{sil}	$\pi_{sil}\|K_{sil}$	π_{sil}	$\pi_{sil}\|K_{sil}$	π_{sil}	$\pi_{sil}\|K_{sil}$
Recombination[a]	−0.092	−0.105*	−0.183***	−0.186***	−0.314***	−0.261***
Gene density (100 kb)[b]	−0.14**	−0.134**	−0.262***	−0.277***	−0.219***	−0.232***
Gene density (1 Mb)	−0.195***	−0.191***	−0.308***	−0.322***	−0.288***	−0.249***
Codon density (100 kb)	−0.113*	−0.074	−0.178***	−0.194***	−0.253***	−0.228***
Codon density (1 Mb)	−0.18***	−0.173**	−0.297***	−0.315***	−0.328***	−0.282***
GC content (100 kb)	0.003	−0.019	−0.059	−0.066	0.031	−0.004
GC content (1 Mb)	−0.029	−0.027	−0.074	−0.092	0.09	0.07

[a] Recombination rates estimated using the sliding-window method.
[b] Window sizes are indicated in parentheses.
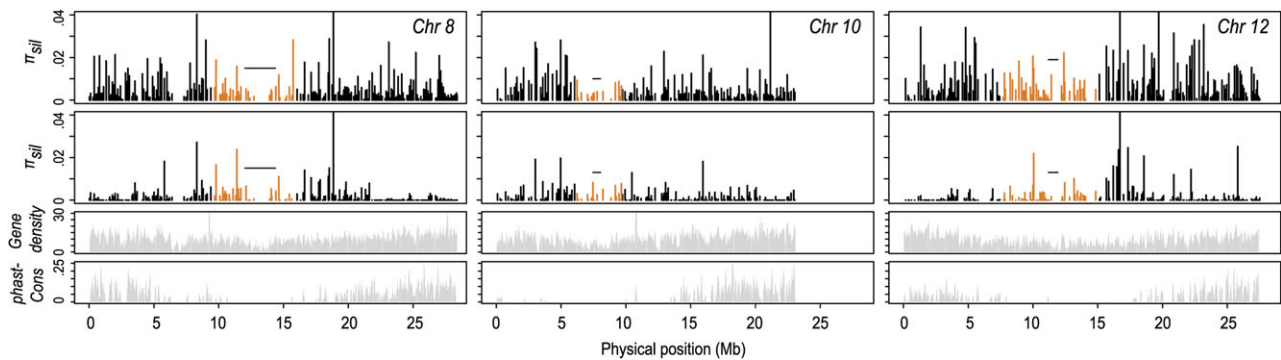*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

Fig. 2. The distribution of polymorphism in *Oryza rufipogon* and *O. sativa* ssp. *japonica* on chromosomes 8, 10, and 12. The top two rows of panels represent polymorphism estimates for *O. rufipogon* and *O. sativa* ssp. *japonica*, respectively. Vertical orange bars indicate polymorphism in pericentromeric regions. Black bars represent polymorphism on the chromosome arms. Horizontal bars indicate the approximate location of the centromeres. Gene density and the proportion of conserved sites with phastCons conservation probabilities of 0.95 or greater are presented in the two lower rows of panels in sliding windows with window size of 100 kb and stepsize of 25 kb. Data for *O. sativa* ssp. *indica* are presented in supplementary figure S1 (Supplementary Material online).

significant in a conservative test that accounts for spatial autocorrelation that may exist among observations because of long-range linkage disequilibrium in the domesticated cultivars (supplementary table S2 and fig. S3, Supplementary Material online). Reduced variation in gene-dense regions is most striking on the euchromatic arms of *japonica* chromosomes (e.g., chromosome 8, long arm), where we observed megabase-scale losses of diversity (fig. 2).

The severely reduced diversity in euchromatic regions translates into a large difference in the levels of polymorphism between the pericentromeres and the chromosome arms in *japonica* (Wilcoxon rank sum test; $P < 0.0005$), a pattern that is not observed in either *O. rufipogon* ($P = 0.30$) or *indica* ($P = 0.82$). In fact, the loss of variation during domestication in *japonica* appears to be disproportionately associated with the arms where nucleotide polymorphism is on average reduced by 69% compared with *O. rufipogon* versus a 47% loss in the pericentromeric regions. Given that all selective sweeps inferred by Molina et al. (2011) fall outside the boundaries of the pericentromeres on chromosomes 8, 10, and 12, we conclude that artificial selection has eliminated variation across large portions of the highly recombining and gene-rich portion of the *japonica* genome above and beyond the extensive losses of variation associated with population bottlenecking.

The CpG Effect Does Not Account for the Pattern of Polymorphism

A major factor contributing to mutation rate variation is an increased rate of transitions at CpG dinucleotides due to spontaneous deamination of methylated cytosines (Duncan and Miller 1980). Evidence that methylated CpG sites are vulnerable to this effect has been documented in plants (SanMiguel et al. 1998; Morton et al. 2006; Ossowski et al. 2010), and increased mutation rates at CpG dinucleotides could influence the regional patterns of polymorphism we observe because cytosine residues are methylated at greater frequency in gene-poor regions of rice chromosomes (Yan et al. 2010). In principle, the

CpG effect could elevate rates of transition mutation preferentially in regions of low gene density and contribute to the negative correlation we observe between polymorphism and gene density in *O. rufipogon*.

To evaluate this hypothesis, noncoding SNPs in *O. rufipogon* were polarized using parsimony with *O. meridionalis* and *O. barthii* as outgroups and assigned to one of the six transition:transversion classes of mutation. Pooling all noncoding SNPs from plus and minus strands, the number of transition mutations per available site was higher at CpG sites (213 C:G→T:A SNPs/2,687 CpG sites) compared with other cytosine contexts (400/25,746; G-test, $G = 292.57$, degrees of freedom $= 1$, $P = 0.00$). This is evidence that spontaneous deamination of methylated cytosines increases the mutation rate at CpG sites in rice (fig. 5).

However, this CpG effect does not appear to contribute to regional differences in polymorphism in wild and domesticated rice. In a regression analysis of C:G→T:A SNP counts, a model with cytosine context (i.e, CpG or CpH), region (i.e., pericentromeres or arms), and their interaction did not fit the data significantly better than a nested model without the interaction ($F_{1,285} = 0.0725$, $P = 0.788$). Removal of region from a model with cytosine context and region also did not lead to a significant change in the deviance ($F_{1,286} = 0.1644$, $P = 0.6854$). These tests do not support an effect of chromosomal region on the density of C:G→T:A SNPs (fig. 5), despite evidence of a greater frequency methylated cytosines in rice heterochromatin (Yan et al. 2010). In a second set of tests between nested models, a model including cytosine context, gene density, and their interaction did not fit the data significantly better than a model without the interaction effect ($F_{1,285} = 0.0006$, $P = 0.981$). Although we did find an effect of gene density on C:G→T:A SNP counts ($F_{1,286} = 6.44$, $P = 0.012$) that mirrors the pattern observed in bivariate correlations across all mutational classes (π_{sil} vs. gene density; table 2), the absence of an interaction between cytosine context and gene density argues against a significant contribution of an elevated rate of transition mutation at CpG
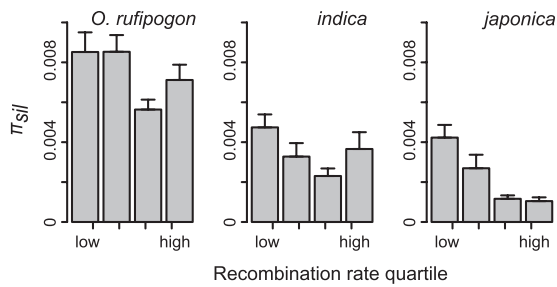
681

Fig. 3. Polymorphism for quartiles of recombination rate in wild and domesticated rice. Average silent site nucleotide diversity (π_{sil}) in recombination quartiles in *Oryza rufipogon* and the *indica* and *tropical japonica* variety groups of *O. sativa*. Error bars represent mean +SEM. Data are based on genes located on the chromosomal arms.

dinucleotides to the higher levels of polymorphism observed in gene-poor regions.

Background Selection

Using phastCons estimates of phylogenetically conserved sites (Wang et al. 2011) and a genomic deleterious rate, U, to obtain coarse estimates of interval-specific deleterious rates, a model of background selection predicts a negative association between gene density and the proportion of variation remaining under background selection under a restrictive set of conditions (fig. 6). With a genomic deleterious rate of $U = 1$, the gene density effect is weak but apparent when the strength of selection is approximately 0.005 or less and the index of panmixia on the order of 0.7. In the situation where this index is determined by outcrossing alone, this is equivalent to an outcrossing rate of 56% (fig. 6 and supplementary fig. S4, Supplementary Material online). The effect becomes notably stronger, however, with smaller selection coefficients and greater frequencies of self-fertilization. When $U = 0.3$, the range of conditions where selection against deleterious mutations can introduce a gene density effect is reduced with readily apparent effects not observed until the index of panmixia is less than 0.1 or an outcrossing rate of approximately 5% (supplementary fig. S5, Supplementary Material online).

Discussion

Examination of molecular variation in the wild rice *O. rufipogon* and domesticated subspecies of *O. sativa* reveals a negative correlation between nucleotide variation and gene density and either a negative correlation or no correlation with recombination rate. If mutation rate or variation in selective constraints explains the gene density effect, we would also expect gene-dense regions to have lower K_{sil}. Divergence, however, is significantly positively correlated with gene density and recombination rate and controlling for divergence in our analysis largely has no effect on the correlations we observe. An alternative mutation-related explanation is that spontaneous mutation at methylated CpG sites elevates the rate of mutation in gene-poor regions owing to preferential methylation of cytosines in rice heterochromatin (Yan et al. 2010). This CpG effect does not
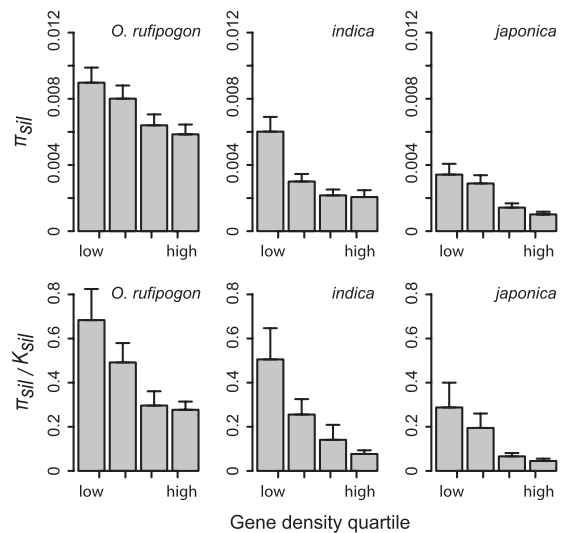

Fig. 4. Polymorphism for quartiles of gene density in wild and domesticated rice. Average nucleotide diversity in gene density quartiles in *Oryza rufipogon* and the *indica* and *tropical japonica* variety groups of *O. sativa* are shown. The top plots are for silent site nucleotide diversity (π_{sil}), whereas the bottom plots are silent site nucleotide diversity normalized against the nucleotide divergence between *O. sativa* and *O. meridionalis* (K_{sil}). Error bars represent mean +SEM. Data are based on genes located on the chromosomal arms.

appear to contribute to regional differences in polymorphism in wild and domesticated rice. Together, these observations suggest that the relationship between nucleotide diversity and gene density in rice does not arise from mutation rate effects or variation in selective constraints but from selection at linked loci.

The negative association between polymorphism and gene density is likely caused by two interacting factors. First, self-fertilization and population substructure increase linkage disequilibrium and enhance the effects of selection at linked loci (Charlesworth et al. 1993; Charlesworth et al. 1997; Nordborg 2000). Without these or other factors that cause departures from panmixia, background selection would be expected to have only a weak effect on rice chromosomes. For example, the proportion of variation remaining after background selection is a negative exponential function of the ratio of the chromosomal deleterious rate and the map length in Morgans (Hudson and Kaplan 1995; Nordborg et al. 1996). Rice chromosome 12 has a map length of 1.1 Morgans and a chromosomal deleterious rate of $\sim U = 0.3/12$, and we therefore would expect less than a 3% reduction in variation due to background selection. We observe, however, that the highest gene density quartile on this chromosome has $\sim$ 60% of the variation of the lowest. Therefore, if background selection explains the reduction in polymorphism, a departure from panmixia in the form of reduced outcrossing and/or population substructure is necessary to account for the gene density effect. *Oryza rufipogon* has a mixed mating system with moderate outcrossing rates (e.g., 20–83%; Gao et al. 2001; Song et al. 2003) and populations that show intermediate levels
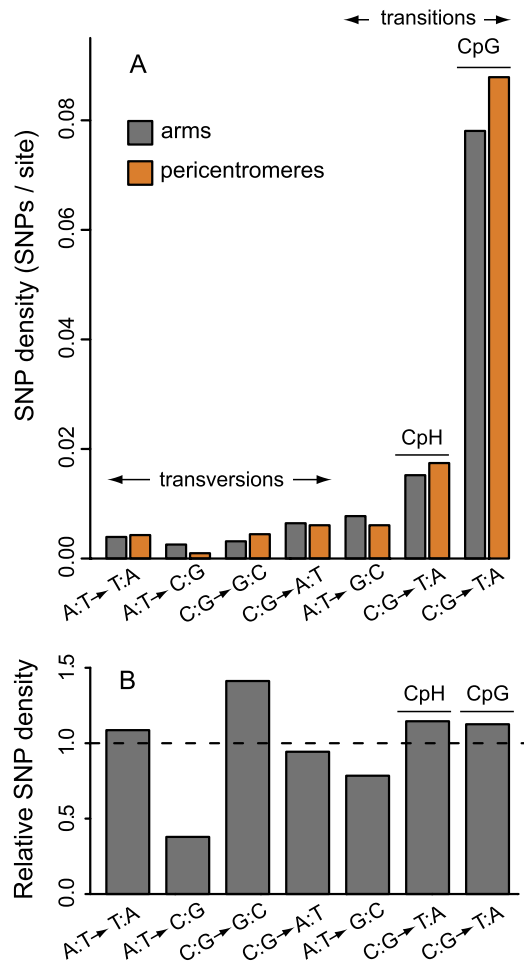
FIG. 5. The mutation spectrum in wild rice. (*A*) SNP densities refer to the number of SNPs in each mutational class divided by the number of available sites in the pericentromeres or arms. Complementary mutation types such as G→A and C→T are pooled. CpH refers to combined CpA, CpT, and CpC contexts. Derived SNPs were inferred with parsimony using *O. meridionalis* and *O. barthii* as outgroups. Data are based on noncoding SNPs. (*B*) Relative SNP densities were calculated separately for each class of mutation as the SNPs per site in pericentromere divided by SNPs per site in the arms. This shows that the nominally higher density of C:G→T:A transitions at CpG sites in pericentromeres is not unusual relative to other mutational classes.

of differentiation (e.g., F_{st} = 0.2–0.3; Gao et al. 2001 and references therein). Our modeling of background selection incorporates these effects through a "panmixis index" (Rockman et al. 2010) and suggests that departures from panmixia are necessary to explain the gene density effect. For the case where this index is governed solely by out-crossing (i.e., no population substructure), the gene density effect becomes apparent when outcrossing rates are less than 50% when $U = 1$ and at rates close to an order of magnitude lower when $U = 0.3$.

Second, genes are organized along recombination gradients in rice, which introduces a positive association between the frequency of crossing-over and gene density (fig. 1 and table 1; Tian et al. 2009). This feature and the location of phylogenetically conserved sites primarily in the euchromatic regions of rice chromosomes (fig. 2;

Wang et al. 2011) suggest that that the density of selected mutations may be higher in regions with normal rates of crossing-over potentially because of a higher rate of deleterious mutation. Selection models with uniform rates of deleterious (or advantageous) mutation predict a positive correlation between polymorphism and recombination rate. However, by introducing interval-specific rates of deleterious mutation proportional to the number of conserved sites in our model, we observed that background selection can effectively reverse the relationship between polymorphism and recombination rate, reduce polymorphism on the chromosome arms relative to regions of low recombination, and recover the negative correlation between polymorphism and gene density we observe. Nevertheless, given the uncertainty in parameter values and assumptions made by the model (Hudson and Kaplan 1995; Nordborg et al. 1996), we place little emphasis on the parameter values themselves and conclude that background selection appears to be sufficient to explain the gene density effect in wild rice provided that deleterious rates are significantly higher in gene-rich regions and that *O. rufipogon* populations conform to a somewhat restrictive but plausible set of conditions.

Although background selection should still operate in the domesticated *O. sativa*, evidence of recurrent selective sweeps associated with artificial selection suggests a clear role for positive selection in shaping patterns of nucleotide variability during rice domestication (Caicedo et al. 2007; Molina et al. 2011; Xie et al. 2011). In our analysis, the negative correlations between nucleotide variation and gene density and recombination frequency are stronger in both *indica* and *japonica* than *O. rufipogon* and appear to be driven by large regions of the chromosome arms that are depauperate of nucleotide variation, particularly in *japonica*. This could reflect greater linkage effects associated with a shift to self-fertilization, recurrent artificial selection, or both. An interesting observation is that large regions depauperate of variation in *japonica* are found exclusively on the chromosome arms. For example, we detected little or no variation in a region spanning nearly half of the long arm of chromosome 8 despite higher frequencies of crossing-over in this region. Evidence of a strong negative correlation between polymorphism and recombination rate and greater hitchhiking in high-recombination regions may support a model of selection during domestication where interference between beneficial alleles in repulsion phase effectively prevents completion of selective sweeps in regions of low recombination. In regions where crossing-over is more frequent, complete selective sweeps are facilitated by more frequent generation of favorable allele combinations. A similar mechanism has been proposed to account for the maintenance of residual heterozygosity preferentially in low-recombination regions of maize recombinant inbred line populations (Gore et al. 2009; McMullen et al. 2009). These observations may indicate a significant role for Hill–Robertson interference (Hill and Robertson 1966) in domestication.

A limitation of our analysis is the coarseness of the rice genetic map and our application of an *O. sativa* genetic
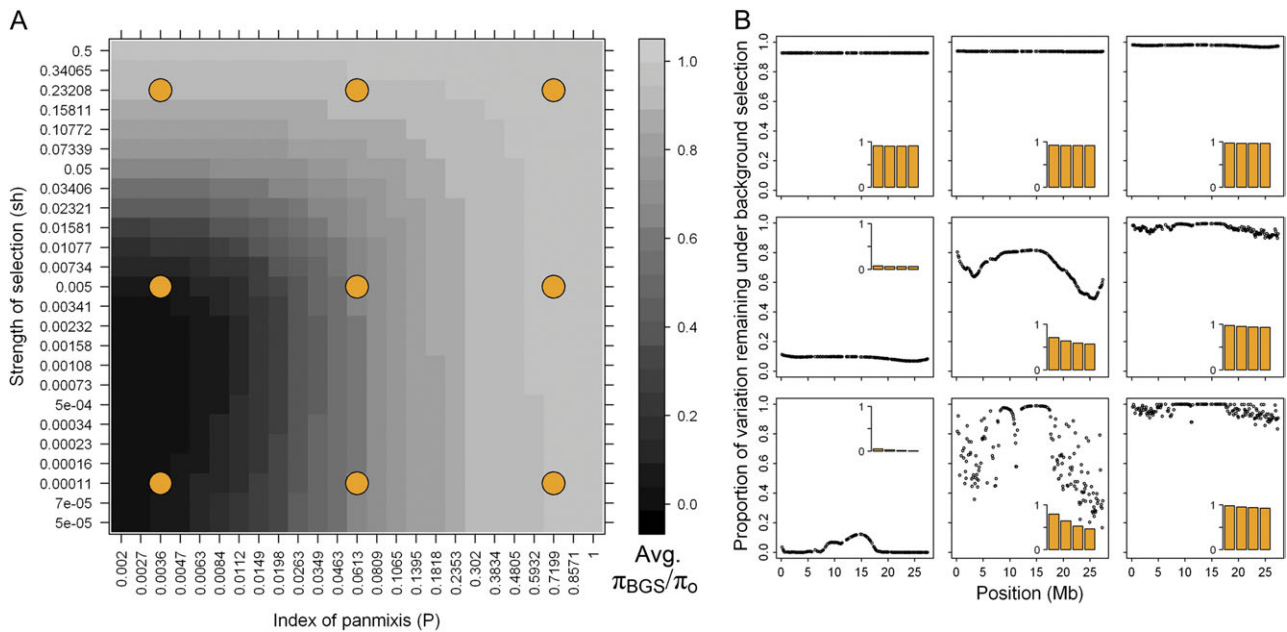
FIG. 6. Background selection predicts a negative correlation between polymorphism and gene density. (A) The predicted average proportion of neutral variation remaining under background selection with regional-specific rates of deleterious mutation plotted as function of the strength of selection and the index of panmixis for $U = 1$. (B) Predicted proportion of neutral variation remaining under background selection (π_{BGS}/π_o) at positions on chromosome 12 for nine parameter combinations (corresponding to the orange dots in A). In the situation where the index of panmixia is governed solely by outcrossing, the values correspond to outcrossing rates of 0.2%, 3%, and 56%. Insets represent average π_{BGS}/π_o in gene density quartiles (analogous to fig. 4) and illustrate the range the conditions where selection against deleterious mutations introduces a negative association between polymorphism and gene density. Average π_{BGS}/π_o in gene density quartiles represent pooled data from chromosomes 8, 10, and 12.

map to *O. rufipogon*. The map data were originally collected from an *indica* × *japonica* cross (Harushima et al. 1998) and the low marker density compared with maps constructed from large-scale genotyping projects means that our estimates of local recombination rates are relatively coarse. However, the negative, but statistically insignificant, relationship between recombination frequency and π_{sil} in *O. rufipogon* and the significantly negative association with π_{sil}/K_{sil} (table 2) are unlikely to be altered significantly by finer scale estimates of the frequency of crossing-over. Moreover, the stronger negative correlations in *indica* and *japonica* are similarly unlikely to be reversed by a more detailed linkage map. Our results may, however, be influenced by applying an *O. sativa* map to *O. rufipogon*. There is evidence that recombination rates have increased in response to selection during domestication across diverse plant groups (Ross-Ibarra 2004). If this is the case in rice, the genetic map in *O. rufipogon* is shorter than the map used in our analysis, which should enhance the effect of background selection. This would effectively increase the range of conditions under which selection against deleterious mutations could explain the negative association between polymorphism and gene density that we observe.

The genomic pattern of polymorphism in *O. rufipogon* is qualitatively similar to observations made in other population genetic model organisms. In *A. thaliana*, polymorphism is also negatively correlated with gene density likely due to background selection and high rates of self-fertilization (Nordborg et al. 2005). Similarly, the self-incompatible,

but geographically structured species, *A. lyrata*, has reduced polymorphism on the chromosome arms relative to the pericentromeres (Wright et al. 2006; Kawabe et al. 2008). If selection is in fact responsible for these patterns, they suggest that selection reduces polymorphism in rice and Arabidopsis over relatively large genomic intervals including genomic regions with frequent crossing-over.

Comparisons with *Caenorhabditis* are also of interest because of the unusual organization of genes in this genus. Genes in *Caenorhabditis* are located primarily in broad domains of suppressed recombination at the center of each chromosome (Barnes et al. 1995; *C. elegans* Sequencing Consortium 1998). Frequent selection in these gene-rich and low-recombination regions should have a pronounced effect on neutral polymorphism (Cutter and Choi 2010). This expectation is met in hermaphroditic species *C. elegans* (Koch et al. 2000; Cutter and Payseur 2003) and *C. briggsae* (Cutter and Choi 2010), and nucleotide variation is both positively correlated with recombination rate and negatively associated with gene density (Cutter and Payseur 2003). Although the gene density effect is comparable with that in rice, the primary cause in *Caenorhabditis* may be selection in or near genes whose effect is enhanced by self-fertilization. This may differ from rice where the reduction in polymorphism in gene-dense regions may be strongly dependent on departures from panmixia. Further clarity on the relative roles of genome organization and mating system in shaping these patterns may be obtained through surveys of polymorphism in obligately outcrossing

species with holocentric chromosomes including gono-choristic species of *Caenorhabditis* (Cutter 2008; Wang et al. 2010) and *Bombyx mori*.

The positive correlation between recombination rate and gene density is found not only in rice (see fig. 1) but in other grass species as well (see table 1). In rice, maize, and other cereal grass species, recombination has been observed to initiate in gene regions (Gaut et al. 2007; Li et al. 2007), which explains the strong correlation of recombination rate and gene density in these species (see table 1). Whether this property has significant impact on patterns of polymorphism in other species as it appears to in rice will be an interesting avenue for future work. It is apparent that more genome-level studies of nucleotide variation need to be undertaken to dissect how the specific relationships between genomic structural elements (e.g., gene density) and other genetic factors (e.g., mutation, recombination, out-crossing rates) interact to pattern the distribution of molecular diversity. The challenge remains to understand, at both theoretical and experimental levels, how the interplay between these factors, and the diversity in genomic architectures of species, can determine how molecular variation is distributed within and between genomes.

Supplementary Material

Supplementary figures S1 – S5 and tables S1and S2 are available online at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

References

Anderson LK, Lai A, Stack SM, Rizzon C, Gaut BS. 2006. Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. *Genome Res.* 16:115–122.

Andolfatto P. 2001. Adaptive hitchhiking effects on genome variability. *Curr Opin Genet Dev.* 11:635–641.

Barnes TM, Kohara Y, Coulson A, Hekimi S. 1995. Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* 141:159–179.

Barton NH. 1995. Linkage and the limits to natural selection. *Genetics* 140:821–841.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519–520.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* 5:2534–2559.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012–2018.

Cai JJ, Macpherson M, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* 5:e1000336.

Caicedo AL, Williamson SH, Hernandez RD, et al. (12 co-authors). 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* 3:1745–1756.

Charlesworth B. 1996. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet Res.* 68:131–149.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular evolution. *Genetics* 134:1289–1303.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res.* 70:155–174.

Crawley MJ. 2007. The R book. West Sussex (United Kingdom): Johnson Wiley & Sons, Ltd.

Cutter AD. 2008. Multilocus patterns of polymorphism and selection across the X chromosome of *Caenorhabditis remanei*. *Genetics* 178:1661–1672.

Cutter AD, Choi JY. 2010. Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome Res.* 20:1103–1111.

Cutter AD, Moses AM. 2011. Polymorphism, divergence, and the role of recombination in *Saccharomyces cerevisiae* genome evolution. *Mol Biol Evol.* 28:1745–1754.

Cutter AD, Payseur BA. 2003. Selection at linked sites in the partial selfer *C. elegans*. *Mol Biol Evol.* 20:665–673.

Drouaud J, Camilleri C, Bourguignon PY, et al. (12 co-authors). 2006. Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". *Genome Res.* 16:106–114.

Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature* 287:560–561.

Fengler K, Allen SM, Li BL, Rafalski A. 2007. Distribution of genes, recombination, and repetitive elements in the maize genome. *Crop Sci.* 47:S83–S95.

Gao L, Wei C, Yang Q, Hong D, Ge S. 2001. Intra-population genetic structure of *Oryza rufipogon* Griff. in Yunnan, China. *J Plant Res.* 114:107–113.

Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc Natl Acad Sci U S A.* 93:10274–10279.

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat Rev Genet.* 8:77–84.

Gore MA, Chia JM, Elshire RJ, et al. (12 co-authors). 2009. A first-generation haplotype map of maize. *Science* 326:1115–1117.

Harushima Y, Yano M, Shomura P, et al. (17 co-authors). 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* 148:479–494.

He Z, Zha W, Wen H, Tang T, Wang Y, Lu X, Greenberg AJ, Hudson RR, Wu C-I, Shi S. 2011. Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genet.* 7:e1002100.

Hellman I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Human Genet.* 72:1527–1535.

Hey J, Kliman RM. 2002. Interactions between natural selection, recombination and gene density in the genes of Drosophila. *Genetics* 160:595–608.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res.* 8:269–294.

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* 141:1605–1617.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* 436:793–800.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. New York: Academic Press.

Kaplan NL, Hudson RR, Langley CH. 1989. The hitchhiking effect revisited. *Genetics* 123:887–899.

Kawabe A, Forrest A, Wright SI, Charlesworth D. 2008. High DNA sequence diversity in pericentromeric genes of the plant *Arabidopsis lyrata*. *Genetics* 179:985–995.

Koch R, van Luenen HGAM, van der Horst M, Thijssen KL, Plasterk RHA. 2000. Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*. *Genome Res.* 10:1690–1696.

Kong A, Gudbjartsson DF, Sainz J, et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. *Nat Genet.* 31:241–247.

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* 18:337–340.

Li J, Hsia A-P, Schnable PS. 2007. Recent advances in plant recombination. *Curr Opin Plant Biol.* 10:131–135.

Liu SZ, Yeh CT, Ji TM, Ying K, Wu HY, Tang HM, Fu Y, Nettleton D, Schnable PS. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.* 5:e1000733.

Ma JX, Bennetzen JL. 2006. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc Natl Acad Sci U S A.* 103:383–388.

Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD. 2007. The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* 177:2223–2232.

Maynard-Smith J, Haigh J. 1974. The hitch-hiking effect of a favorable gene. *Genetical Res.* 23:23–35.

McMullen MD, Kresovich S, Villeda HS, Bradbury PJ, Li H, et al. (32 co-authors). 2009. Genetic properties of the maize nested association mapping population. *Science* 325:737–740.

Molina J, Sikora M, Garud N, et al. (12 co-authors). 2011. Molecular evidence for a single evolutionary origin of domesticated rice. *Proc Natl Acad Sci U S A.* 108:8351–8356.

Morton BR, Bi IV, McMullen MD, Gaut BS. 2006. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* 172:569–577.

Nachman MW. 1997. Patterns of DNA variability at X-linked loci in *Mus domesticus*. *Genetics* 147:1303–1316.

Nachman MW. 1998. DNA variability and recombination rates at X-linked loci in humans. *Genetics* 150:1133–1141.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3:418–426.

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graphic with partial self-fertilization. *Genetics* 154:923–929.

Nordborg M, Charlesworth C, Charlesworth D. 1996. The effect of recombination on background selection. *Genet Res.* 67:159–174.

Nordborg M, Hu TT, Ishino Y, et al. (24 co-authors). 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* 3:e196.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92–94.

Ouyang S, Zhu W, Hamilton J, et al. (14 co-authors). 2006. The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.* 35:D883–D887.

Przeworski M, Hudson RR, Di Rienzo A. 2000. Adjusting the focus on human variation. *Trends Genet.* 16:296–302.

Rakshit S, Rakshit A, Matsumura H, Takahashi Y, Hasegawa Y, Ito A, Ishii T, Miyahita NT, Terauchi R. 2007. Large-scale DNA polymorphism study of *Oryza sativa* and *O. rufipogon* reveals the origin and divergence of Asian rice. *Theor Appl Genet.* 114:731–743.

Rezvoy C, Charif D, Gueguen L, Marais GAB. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* 23:2188–2189.

Rizzon C, Marais G, Gouy M, Biemont C. 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Res.* 12:400–407.

Rockman MV, Skrovanek SS, Kruglyak L. 2010. Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* 330:372–376.

Ross-Ibarra J. 2004. The evolution of recombination under domestication: a test of two hypotheses. *Am Nat.* 163:105–112.

Saintenac C, Faure S, Remay A, Choulet F, Ravel C, Paux E, Balfourier F, Feuillet C, Sourdille P. 2011. Variation in crossover rates across a 3-Mb contig of bread wheat (*Triticum aestivum*) reveals the presence of a meiotic recombination hotspot. *Chromosoma* 120:185–198.

SanMiguel PJ, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet.* 20:43–45.

Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionary conserved elements in vertebrate, insect, worm and yeast genomes. *Genome Res.* 15:1034–1050.

Song ZP, Xu X, Wang B, Chen JK, Lu B-R. 2003. Genetic diversity in the northernmost *Oryza rufipogon* populations estimated by SSR markers. *Theor Appl Genet.* 107:1492–1499.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Human Genet.* 76:449–462.

Stephens M, Smith N, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Human Genet.* 68:978–989.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Takahashi A, Liu Y-H, Saitou N. 2004. Genetic variation versus recombination rate in a structured population of mice. *Mol Biol Evol.* 21:404–409.

Thornton KR. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.

Tian Z, Rizzon C, Du J. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* 19:2221–2230.

Wang GX, Ren SA, Ren YA, Ai H, Cutter AD. 2010. Extremely high molecular diversity within the East Asian nematode *Caenorhabditis* sp. 5. *Mol Ecol.* 19:5022–5029.

Wang J, Kong L, Zhao S, Zhang H, Tang L, Li Z, Gu X, Luo J, Gao G. 2011. Rice-Map: a new-generation rice genome browser. *BMC Genomics.* 12:165.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7:256–278.

Wiehe THE, Stephan W. 1993. Analysis of a genetic hitchhiking model and its application to DNA polymorphism data from Drosophila melanogaster. *Mol Biol Evol.* 10:842–854.

Wright SI, Agrawal N, Bureau TE. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Res.* 13:1897–1903.

Wright SI, Andolfatto P. 2008. The impact of natural selection on the genome: emerging patterns in Drosophila and Arabidopsis. *Annu Rev Ecol Evol Syst.* 39:193–213.

Wright SI, Foxe JP, DeRose-Wilson L, Kawabe A, Looseley M, Gaut BS, Charlesworth D. 2006. Testing for effects of recombination rate on nucleotide diversity natural populations of *Arabidopsis lyrata*. *Genetics* 174:1421–1430.

Wu J, Maehara T, Shimokawa T, et al. (19 co-authors). 2002. A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* 14:525–535.

Xie X, Molina J, Hernandez R, Reynolds A, Boyko AR, Bustamante CD, Purugganan MD. 2011. Levels and patterns of nucleotide variation in domestication QTL regions on rice chromosome 3 suggest lineage-specific selection. *PLoS One.* 6:e20670.

Yan H, Jiang J. 2007. Rice as a model for centromere and heterochromatin research. *Chromosome Res.* 15:77–84.

Yan H, Jin W, Nagaki K, Tian S, Ouyang S, Buel CR, Talbert PB, Henikoff S, Jiang J. 2005. Transcription and histone modification in the recombination-free region spanning a rice centromere. *Plant Cell* 17:3227–3238.

Yan H, Kikuchi S, Neumann P, Zhang W, Wu Y, Chen F, Jiang J. 2010. Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice. *Plant J.* 63:353–365.

Yan H, Talbert PB, Lee HR, Jett J, Henikoff S, Chen F, Jiang J. 2008. Intergenic locations of rice centromeric chromatin. *PLoS Biol.* 6:e286.