University of Massachusetts Amherst

From the SelectedWorks of Ana Lucia Caicedo

2007

Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice

Ana Lucia Caicedo, *University of Massachusetts - Amherst*
S. H Williamson
R, D Hernandez
A. Boyko
A. Fleded-Alon, et al.

Available at: https://works.bepress.com/ana_caicedo/15/

Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice

Ana L. Caicedo[1☯¤a], Scott H. Williamson[2☯], Ryan D. Hernandez[2], Adam Boyko[2], Adi Fledel-Alon[2¤b], Thomas L. York[2], Nicholas R. Polato[3], Kenneth M. Olsen[1¤c], Rasmus Nielsen[2¤d], Susan R. McCouch[3], Carlos D. Bustamante[2*], Michael D. Purugganan[1,4,5*]

1 Department of Genetics, North Carolina State University, Raleigh, North Carolina, United States of America, **2** Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, **3** Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York, United States of America, **4** Department of Biology, New York University, New York, New York, United States of America, **5** Center for Comparative Functional Genomics, New York University, New York, New York, United States of America

Domesticated Asian rice (*Oryza sativa*) is one of the oldest domesticated crop species in the world, having fed more people than any other plant in human history. We report the patterns of DNA sequence variation in rice and its wild ancestor, *O. rufipogon*, across 111 randomly chosen gene fragments, and use these to infer the evolutionary dynamics that led to the origins of rice. There is a genome-wide excess of high-frequency derived single nucleotide polymorphisms (SNPs) in *O. sativa* varieties, a pattern that has not been reported for other crop species. We developed several alternative models to explain contemporary patterns of polymorphisms in rice, including a (i) selectively neutral population bottleneck model, (ii) bottleneck plus migration model, (iii) multiple selective sweeps model, and (iv) bottleneck plus selective sweeps model. We find that a simple bottleneck model, which has been the dominant demographic model for domesticated species, cannot explain the derived nucleotide polymorphism site frequency spectrum in rice. Instead, a bottleneck model that incorporates selective sweeps, or a more complex demographic model that includes subdivision and gene flow, are more plausible explanations for patterns of variation in domesticated rice varieties. If selective sweeps are indeed the explanation for the observed nucleotide data of domesticated rice, it suggests that strong selection can leave its imprint on genome-wide polymorphism patterns, contrary to expectations that selection results only in a local signature of variation.

Citation: Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, et al. (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet 3(9): e163. doi:10.1371/journal.pgen.0030163

Introduction

Domestication is a complex, cumulative evolutionary process in which human use of organisms leads to morphological and/or behavioral changes distinguishing domesticated species from their wild ancestors [1,2]. Beginning with Charles Darwin [3,4], there has been strong interest in the study of domestication of crop species as a means of understanding the nature of selection. Moreover, domestication and the development of agriculture are arguably the most important technological innovations in human history [5]. Crop plant domestication was the linchpin of the Neolithic Revolution 10,000–12,000 years ago, in which hunter-gatherer groups transitioned into sedentary agricultural societies that gave rise to current human cultures [6]. With domestication came the availability of food surpluses, and this agricultural development led to craft specializations, art, religious and social hierarchies, writing, urbanization, and the origin of the state [5].

One of the earliest domesticated crop species is cultivated Asian rice, *Oryza sativa* L., which has become the world's most widely grown crop and has also assumed the stature of a key model system in plant biology. Rice consumption constitutes about 20% of the world's caloric intake, and in Asian countries, where over half of the world's population lives, rice often represents over 50% of the calories consumed [7]. Because of its small genome size, rice has been the first crop plant to have its whole genome sequenced [8–10].

A wealth of morphological, physiological, and ecological variation exists within cultivated Asian rice, reflected in the large number of recognized cultivars or strains [11,12]. Two main rice varietal groups, *O. sativa indica* and *O. sativa japonica*, have been recognized since ancient China [13]. Although phenotypic distinctions between these groups is not always straightforward, *indica* varieties tend to be found throughout

Editor: Gil McVean, University of Oxford, United Kingdom

Received February 20, 2007; **Accepted** August 6, 2007; **Published** September 28, 2007

A previous version of this article appeared as an Early Online Release on August 6, 2007 (doi:10.1371/journal.pgen.0030163.eor).

Copyright: © 2007 Caicedo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: AIC, Akaike information criterion; GOF, goodness-of-fit; SNP, single nucleotide polymorphism; STS, sequence-tagged site(s)

* To whom correspondence should be addressed. E-mail: cdb28@cornell.edu (CDB); mp132@nyu.edu (MDP)

☯ These authors contributed equally to this work.

¤a Current address: Department of Biology, University of Massachusetts, Amherst, Massachusetts, United States of America

¤b Current address: Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America

¤c Current address: Department of Biology, Washington University, St. Louis, Missouri, United States of America

¤d Current address: Centre for Bioinformatics, University of Copenhagen, Denmark

Author Summary

Domesticated Asian rice is one of the oldest and most important crops in the world. Two main rice evolutionary lineages have been identified, and are thought to have been independently domesticated in Asia. We have examined patterns of DNA sequence variation in the genomes of rice and its wild ancestor to make inferences about the origin of domesticated rice. Population bottlenecks (a reduction in the size of the founding population) in the evolutionary transition from wild to cultivated species has long been thought to be the dominant force shaping patterns of molecular evolution during domestication. We find that the nucleotide variation patterns in rice are inconsistent with a simple bottleneck model. Rice genetic variation, however, can be explained by either a model that incorporates both a bottleneck and migration among rice variety groups, or a model that incorporates a bottleneck and multiple rounds of artificial selection on rice. Selection by humans is believed to have played an important role during crop domestication, and these results may suggest that strong, recurrent selection can leave a signal that can be observed throughout the genomes of domesticated species.

the tropical regions of Asia and are primarily grown in lowland conditions, while *japonica* types are differentiated into *tropical japonica*, distributed in upland tropical regions, and *temperate japonica*, a recently derived group cultivated in temperate regions [11,13,14]. Additional variety groups include *aus*, drought-tolerant rice from Bangladesh and West Bengal, and *aromatic*, fragrant rice from the Himalayan range [14,15]. All rice varieties have a predominantly self-fertilizing mating system [13]. Both morphological and isozyme data have established that *O. rufipogon* Griff., a partially outcrossing species native to southern Asia, is the wild ancestor of domesticated rice [13].

In this paper, we describe the levels and patterns of DNA sequence polymorphism across the rice genome and that of its wild ancestor, *O. rufipogon*. To our knowledge this is the first genome-wide characterization of sequence variation in domesticated Asian rice, and we show that rice contains a unique pattern of excess high-frequency derived single nucleotide polymorphisms (SNPs) that has not been reported in other species. We develop four models to explain patterns of genetic variation in *O. sativa* and *O. rufipogon*, including a simple selectively neutral bottleneck model that has been previously thought to be the dominant demographic force shaping levels of nucleotide variation in crop species. We demonstrate that this simple bottleneck model is inadequate to explain the origin of domesticated rice. We conclude that either positive selection has made a significant impact on genomic polymorphism patterns, or that domestication involved an extremely severe bottleneck (~99.5% reduction) coupled with gene flow among modern varieties and between domesticated rice and its wild ancestor.

Results/Discussion

Nucleotide Variation in the Rice Genome

To assess levels and patterns of polymorphism in the rice genome, we sequenced one hundred eleven randomly chosen gene fragments (sequence-tagged sites or STS) in a diverse panel of *Oryza* accessions, including 72 from *O. sativa* and 21 from *O. rufipogon* (Tables S1 and S2). Average silent (synonymous and noncoding) site nucleotide diversity (θ_π) across all sampled loci in *O. sativa* is approximately 3.20×10^{-3} (Table 1). Levels of polymorphism in the wild ancestral species, *O. rufipogon*, are predictably higher than rice, with a mean silent θ_π of 5.19×10^{-3} (Table 1). These levels of polymorphism are lower than those observed for maize, a domesticated outcrossing species [16], and *Arabidopsis thaliana*, a selfing, wild species [17,18].

To determine if any genetic differentiation due to population structure among rice groups is evident in these STS sequences, we used the Bayesian clustering program STRUCTURE [19]. The highest likelihood obtained was with a model specifying $K = 7$ groups (Figure 1; Table S1). Five groups occur within *O. sativa* and correspond to the traditional variety designations, as described previously [14]. Evidence of some limited geographical population structure is also observed in *O. rufipogon* (Figure 1; Table S1). Neighbor-joining analysis of the concatenated STS sequences (Figure S1) revealed two distinct clusters within cultivated rice; one comprises a *tropical japonica*, *temperate japonica*, and *aromatic* rice lineage, and another consists of *aus* and *indica* rice. The apparent monophyly of these major groups is consistent with at least two domestication events in rice [14,20–24]. The nesting of the *aromatic* and the *temperate japonica* variety groups within *tropical japonica* suggests the first two groups originated from secondary divergence events from the latter, although

Table 1. Average Diversity Measures in *Oryza* spp. across 111 STS Regions

Statistic	Category	*O. sativa*							*O. rufipogon*
		Combined	*aromatic*[a]	*aus*	*indica*	*japonica*	*temperate japonica*	*tropical japonica*	
θ_π[b] per Kb	Total sites	2.29	1.26	1.20	1.35	1.11	0.418	1.15	3.57
	Silent sites[c]	3.20	1.86	1.69	1.91	1.47	0.510	1.57	5.19
θ_W[d] per Kb	Total sites	2.11	1.32	1.12	1.58	1.23	0.693	1.27	3.70
	Silent sites	2.92	1.97	1.57	2.21	1.67	0.948	1.70	5.42
Tajima's D[e]	Total sites	0.2784	−0.2225	0.0148	−0.3382	−0.2642	−0.8459	−0.4137	−0.2710
	Silent sites	0.2223	−0.2768	0.0588	−0.2177	−0.3619	−1.1115	−0.4677	−0.3040

[a]Values for the aromatic group are based on 110 STS.
[b]Average nucleotide diversity across all sequenced STS fragments.
[c]Silent site estimates include synonymous and noncoding sites.
[d]Average values for Watterson's estimate of theta across all sequenced STS fragments.
[e]Average Tajima's D across all sequenced STS fragments.
doi:10.1371/journal.pgen.0030163.t001

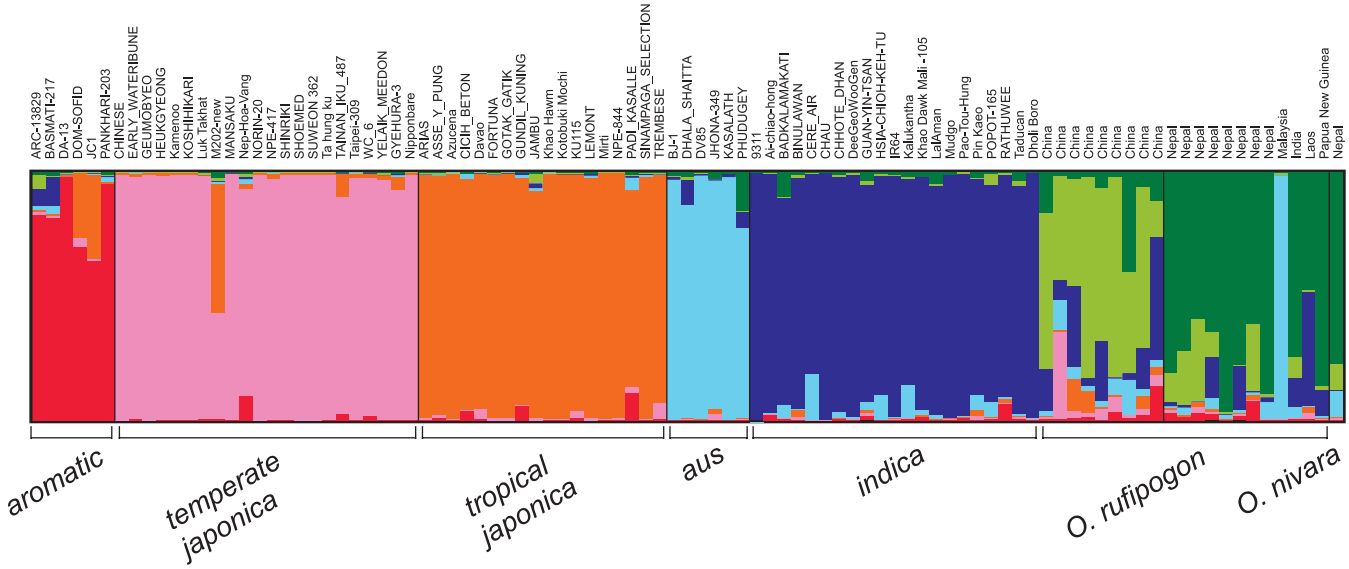


Figure 1. Estimated Population Structure for 97 Accessions of *O. sativa* and *O. rufipogon* from 111 STS Loci

Vertical bars along the horizontal axis represent each *Oryza* accession; for all accessions, the proportion of ancestry under $K = 7$ clusters that can be attributed to each cluster is given by the length of each colored segment in a bar.

doi:10.1371/journal.pgen.0030163.g001

the lack of support for *tropical japonica* branches does not exclude other possible divergence scenarios (Figure S1). *Indica* and *aus* relationships, on the other hand, are consistent with rapid divergence after domestication or separate domestication events from the same ancestral gene pool. Within-group SNP levels of cultivated rice are lower than those of the whole species (Table 1), with subpopulations harboring between 19% (*temperate japonica*) and 43% (*indica*) of the polymorphism of *O. rufipogon*. Assuming separate domestication events, the *japonica* clade contains 42% and the *indica* clade contains 48% of the diversity levels found in *O. rufipogon*.

The Derived Site-Frequency Spectrum is U-Shaped in *O. sativa*

Because of the strong population structure evident in our rice sample, it is necessary to assess patterns of variation separately for each group when making inferences about the evolutionary dynamics of domestication. *Indica* and *tropical japonica* represent the most widely grown cultivars for each of the separate domestication events, and we limited our characterization of polymorphism patterns to these two groups. We examined the frequency spectrum of segregating sites within loci using Tajima's D [25], and found that *O. rufipogon* and the two main rice subspecies show an excess of rare alleles, as evidenced by the biased distribution of Tajima's D toward negative values (Figure S2; Table 1). Crops are expected to have gone through a population bottleneck during domestication, as only a limited number of founding individuals were brought into cultivation. The distribution of Tajima's D in the domesticated rice varieties is inconsistent with a recent bottleneck, however, as these should reduce levels of low-frequency variants and bias measures of Tajima's D toward positive values. It is possible that subsequent population expansion, due to the spread of rice agriculture, could be responsible for the over-representation of rare alleles segregating in domesticated rice varieties, or selection may have played a role.

We further examined the derived site-frequency spectrum across SNPs (i.e., the fraction of derived polymorphisms present at various frequencies within a group) in *indica* and *tropical japonica*. To infer ancestral alleles for each SNP, we used as an outgroup *O. meridionalis*, a species believed to have diverged from *O. sativa* ~2 million years ago [21]. In each *O. sativa* variety we observed a large number of high-frequency derived mutations (i.e., derived SNPs above 70% frequency in the population) leading to a U-shaped frequency distribution (Figure 2); this type of pattern has not been reported at the genomic level in any other species.

Possible explanations for the excess of high-frequency derived SNPs in *O. sativa* include the misidentification of ancestral states due to shared polymorphism with *O. meridionalis*, or the occurrence of multiple mutations at given sites since divergence from *O. meridionalis*. However, both misidentification of derived alleles and multiple hits would be expected to also affect the site-frequency spectrum of *O. rufipogon*, which is not observed (Figure 2). This suggests that the *O. sativa* derived site-frequency distribution is a result of the domestication process. Furthermore, derived alleles at high frequency in the *O. sativa* varieties occur primarily at low to intermediate frequency in *O. rufipogon*, suggesting that such alleles have only recently increased in frequency (Figure S3).

We also checked the ancestral state calls in *O. sativa* using the African wild rice *O. barthii*. Although *O. barthii* is more closely related to *O. sativa* than is *O. meridionalis*, if we assume that both wild species share ancestral polymorphisms with domesticated rice, the possibility that we always identified the same alternative allele as derived in our sample should be low. Using this approach, we find that 88% of our ancestral SNP calls in *indica* and 86% in *tropical japonica* matched in *O. barthii* and *O. meridionalis*. Even when using only the matched calls (which is a very conservative criterion, since it does not take into account drift and/or fixation processes in *O. barthii*), the site frequency spectrum in *O. sativa* varieties remains U-shaped.

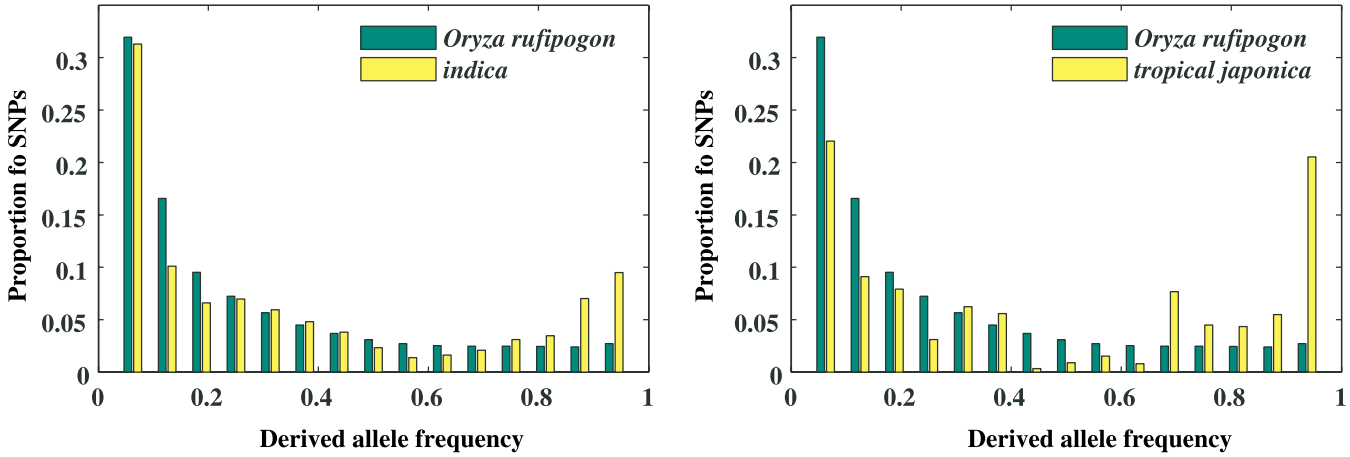


Figure 2. The Observed Marginal Derived Site-Frequency Spectra of Noncoding and Synonymous SNPs for Two Population Pairs: *indica* and *O. rufipogon* and *tropical japonica* and *O. rufipogon*

To accommodate SNPs with missing data, all spectra are plotted as the expected site frequency spectrum in a subsample of the data of size $n = 16$.

doi:10.1371/journal.pgen.0030163.g002

An excess of high-frequency derived SNPs is often interpreted as a result of genetic hitchhiking during recent selective sweeps [26]. Because the site-frequency spectrum in rice varieties is observed from randomly selected loci, and the loci contributing high-frequency derived SNPs are distributed across the genome (Figure S4), this pattern suggests that strong linkage to positively selected mutations occurred within most of the genome. However, demographic forces may have also played a role in shaping the rice genomes. We developed several demographic models and a multiple selective sweeps model to test which evolutionary processes may best explain the observed patterns of polymorphism in rice.

Table 2. Maximum Likelihood Estimates for Demographic Parameters of the Bottleneck and Bottleneck plus Migration Models in *indica*, *tropical japonica*, and *O. rufipogon*

Model	Rice Group	τ_1[a]	ν_2[b]	ν_b[c]	$0.5 \times M \times \nu_2$[d]
Bottleneck[e]	*indica*	0.1044	0.70	0.0246	0
	tropical japonica	0.0508	0.397	0.0113	0
Bottleneck + migration[f]	*indica*	0.04[g]	0.27	0.0055[g]	0.945
	tropical japonica	0.04[g]	0.12	0.0055[g]	0.42
	rufipogon	—	1	—	3.5

[a] τ_1 is time back to the start of domestication period (~12,000 y before present) scaled by $2N_{rufi}$.

[b] ν_2 is the ratio of the effective population size of current population to ancestral populations.

[c] ν_b is the ratio of the effective population size of bottleneck population to ancestral population (i.e., N/N_{rufi}).

[d] $0.5 \times M \times \nu_2 = 2Nm$ is the number of migrants arriving per generation into the *indica*, *tropical japonica*, and *O. rufipogon* populations after the bottleneck ends (maximum composite likelihood estimate of migration rate, $M = 4N_{rufi} = 7.0$).

[e] Bottleneck model was evaluated by Poisson random fields analysis of site frequency spectrum.

[f] Bottleneck + migration model was evaluated by composite likelihood analysis of marginal site-frequency spectrum using coalescent simulations.

[g] Within ν_2 and ν_b columns, these parameters are set to be equal to one another in optimization.

doi:10.1371/journal.pgen.0030163.t002

Demographic Models for Rice Domestication: A Neutral Population Bottleneck Model

The most widely accepted demographic model for crop domestication is a neutral bottleneck model [27–29]. In this model, rice domestication is assumed to be a result of recent population divergence, with one of the two daughter populations experiencing a reduction in population size at divergence associated with the founder effect at the time of domestication, followed by population growth as cultivation of the crop increases. To fit this model to our data, we used a diffusion-based approach [30–32] to predict the pattern of allele frequencies in domestic and ancestral populations under selective neutrality.

Details of the inference procedure can be found in the Materials and Methods section. The composite-likelihood function we employed uses the reduction in diversity observed in either of the domesticated rice subspecies and the shift in allele frequency distribution to estimate four parameters: the time back until the start of domestication (τ_1), duration of the bottleneck (τ_2), ratio of current population to ancestral population size (ν_2), and relative size of the bottleneck population to the ancestral population (ν_b). The duration of the bottleneck was assumed to be 25% of the time back until domestication ($\tau_2 = 0.25 \times \tau_1$), which is consistent with archeological data suggesting it took ~3,000 y from the time of initial cultivation (~12,000 y ago) until the appearance of domesticated rice grains [33,34].

Bottleneck parameter estimates for *indica* and *tropical japonica* are broadly comparable, with a slightly more severe bottleneck in *tropical japonica* (Table 2). Assuming the time back to the beginning of domestication for both variety groups was ~12,000 y [35], we can independently derive estimates of the current *O. rufipogon* effective population size, N_{rufi}, using the relationship $\tau_1 \times 2N_{rufi} = 12{,}000$ (because τ_1 is scaled by $2N_{rufi}$). From the *indica* analyses, N_{rufi} is equal to $12{,}000/(2 \times 0.1044) = 57{,}471$, and from the *tropical japonica* analyses is equal to $12{,}000/(2 \times 0.0508) = 118{,}110$ (this exact

value of N_{rufi} is important in scaling all of the estimated parameters into years and number of individuals). The *indica*-derived N_{rufi} estimate implies bottleneck and current estimated population size (N_e) for *indica* of ($\nu_b \times N_{rufi}$) = 1,413 and ($\nu_2 \times N_{rufi}$) = 40,229 respectively. The second estimate suggests a bottleneck and current N_e sizes for *tropical japonica* of ($\nu_b \times N_{rufi}$) = 1,334 and ($\nu_2 \times N_{rufi}$) = 46,889, respectively.

The differences in estimates of N_{rufi} from each analysis could be attributable to differences in the founding population of each variety group or differences in the timing of each domestication event. We note, however, that a bottleneck model conditioned on coincident domestication for *indica* and *tropical japonica* (equal τ_1 values) differs only by 1.8 log likelihood units (unpublished data), suggesting that equal timing of domestication is likely to have occurred. An independent estimate of N_{rufi} can be found by using the estimated scaled population silent mutation rates ($\theta_W = 4N_{rufi}\mu = 5.42 \times 10^{-3}$ per bp; Table 1) and the observation that the *O. rufipogon* site-frequency spectrum is consistent with that of a population of long-term constant size (Figure 2). Assuming a neutral mutation rate of 10^{-8} per bp, yields a point estimate of $N_{rufi} = 135{,}500$, which is slightly higher, but close to the estimates found by conditioning on the start of domestication.

Demographic Models for Rice Domestication: A Complex Model Incorporating Subdivision, Bottlenecks, and Migration

It is important to note that population bottlenecks alone would not generate the strong excess of high-frequency derived alleles and strong U-shaped site-frequency spectrum observed in *O. sativa* (Figure 2) [36]. In order to explain this aspect of the data, we considered several demographic models that included ancient subdivision in the ancestor of rice, a bottleneck at the time of domestication for each domesticated varietal group, and limited gene flow between the independently domesticated rice groups *indica* and *tropical japonica*. Ancient, strong subdivision is not evident in our *O. rufipogon* sample (Figure 1); F_{st} between Chinese and non-Chinese *O. rufipogon* is low, about 0.16, and no interior modes are evident in the site-frequency spectrum of *O. rufipogon*, as expected under subdivision. However, it is possible for limited gene flow in *O. rufipogon* to lead to some differentiation of allele frequency between groups, but not so much that it would have a strong effect on a combined *O. rufipogon* sample. Furthermore, the population bottlenecks induced by independent domestication events could amplify any allele frequency differentiation between *indica* and *tropical japonica*, and limited gene flow between these two groups could introduce ancestral alleles into each population, causing mutations previously fixed in one group to be observed as high-frequency derived alleles in the other.

To test the effect of ancestral population substructure within *O. rufipogon* prior to the domestication of the two *O. sativa* groups, we fit the parameters of a complex demographic model to our data using a composite likelihood technique (see Materials and Methods). We began by exploring a model with seven demographic parameters, which consists of *O. rufipogon* being subdivided into two demes of equal size, sharing on average M_R migrants per generation. Current-day *indica* varieties are descended from one of these demes, while *tropical japonica* varieties descend from the other. During the domestication process, each population underwent a bottleneck that began τ_1 generations ago (in units $2N_{rufi}$) and had severity ν_b (the ratio of the reduced population size to the ancestral size). After $\tau_2 = 0.25 \times \tau_1$ generations (~3,000 y), both *indica* and *tropical japonica* partially recovered, instantaneously reaching a fraction υ_I and υ_J of the ancestral size, respectively. Contemporary gene flow (since domestication) between *tropical japonica*, *O. rufipogon*, and *indica* is captured by the last parameter, the average number of migrants per generation between these demes (M_S). This model was conceived because it incorporates key demographic features of rice or crop domestication (e.g., bottlenecks, two domestication events) and could conceptually generate the observed derived SNP site frequency spectrum.

In preliminary analyses, we found that the migration rate (M_R) between the two ancestral *O. rufipogon* demes was very large, with the marginal likelihood surface for this parameter near its maximum value whenever $M_R > 7$. This is consistent with our observations of limited population structure in *O. rufipogon* (above), and we therefore discarded ancestral population structure as a main contributor to the patterns observed in our dataset, and simplified the demographic model to consider only a single ancestral population from with both *indica* and *tropical japonica* derive (with migration rates among the three remaining demes, $M_S = 4N_{rufi}m$). This assumption reduced the computational complexity, so that the remaining parameters could be estimated via a grid search using an initial size of over 2,000 points with 1,000,000 coalescent simulations per point. The resulting model (which we refer to as the bottleneck plus migration model) has five free parameters with composite maximum likelihood estimates of $M_S = 7.0$ (migration between demes), $\nu_b = 0.0055$ (domestication bottleneck size), $\nu_I = 0.27$ (ratio of *indica* to *O. rufipogon* N_e), $\nu_J = 0.12$ (ratio of *tropical japonica* to *O. rufipogon* N_e), and $\tau_1 = 0.04$ (start of domestication in units of $2N_{rufi}$) (Table 2). It is important to note that coalescent simulations scale the migration based on population size, so the number of migrants entering into the *tropical japonica* population is smaller ($0.5 \times M \times \nu_J = 0.42$), than into *indica* ($0.5 \times M \times \nu_I = 0.945$), and *O. rufipogon* ($0.5 \times M = 3.5$).

In Figure 3, we report the profile composite-likelihood contours for the three key demographic parameters in the bottleneck plus migration model: migration rate, start of the bottleneck, and severity. The figure is constructed by holding two parameters fixed at a given point in the (x,y) plane, optimizing over the third parameter, and reporting the maximum likelihood attained for the (x,y) point (due to computational limitations the figure was constructed holding the ratio of current-day *indica* and *tropical japonica* populations at their maximum composite-likelihood estimates). We note that the three parameters are moderately to strongly correlated, but only a restricted set of values in high dimensional space is consistent with the data. These solutions all include: a very strong bottleneck (>99% reduction), high rates of migration within and between domesticated and wild populations of Asian rice ($M > 5$), and current-day effective population sizes for cultivated rice that are substantially smaller than those seen in the ancestral population. We also note that the model solutions show a positive correlation between size of bottleneck population and timing of the

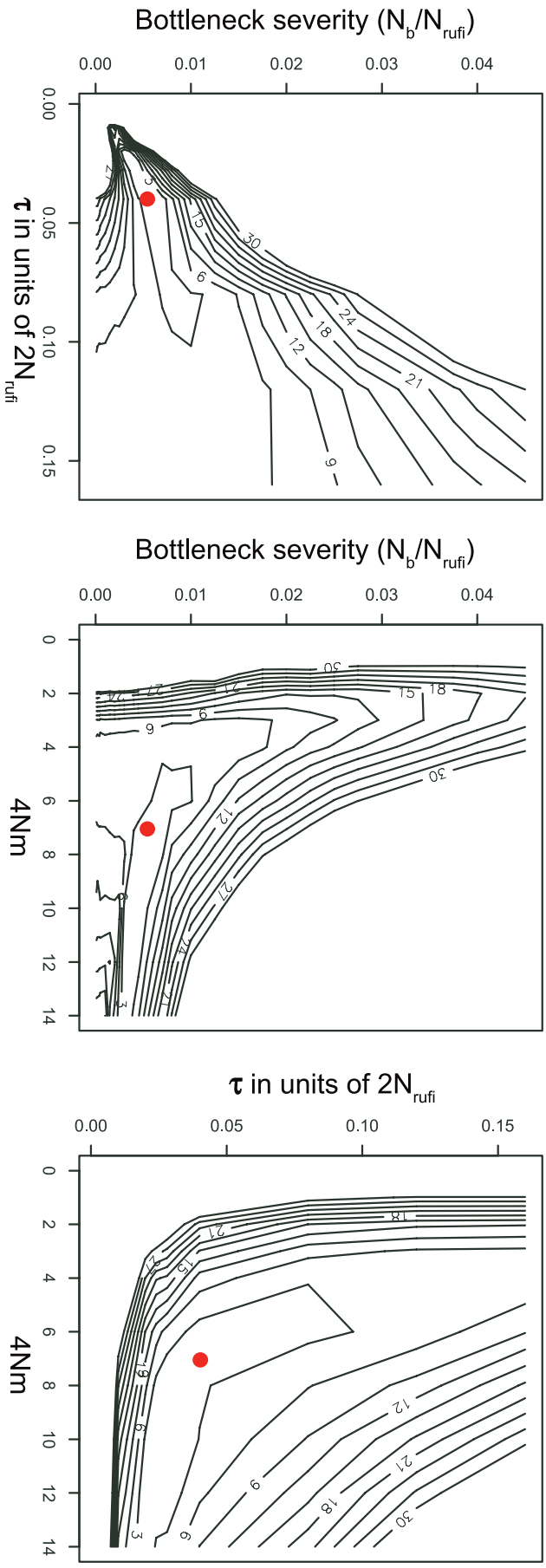


Figure 3. Contours of Composite Profile Log-Likelihood Surface under the Bottleneck and Migration (i.e., "Complex Demography") Model for Three Key Demographic Parameters

Parameters include bottleneck severity, migration rate among demes ($4Nm$), and τ_1 (time back until start of domestication scaled in units of $2N_{rufi}$). The maximum composite-likelihood estimate of the parameters is denoted by a red filled circle.

doi:10.1371/journal.pgen.0030163.g003

bottleneck, a negative correlation between size of the bottleneck and migration, and a negative correlation between migration and timing (consistent with the ~2-fold difference in the estimated time of the bottleneck between the model with migration and the model without).

As can be seen in Figure 4, the expected site-frequency spectrum under the best fitting bottleneck plus migration model matches the observed frequency distributions fairly well for both *O. rufipogon* as well as *indica*, but not as well for *tropical japonica*. As expected, the total number of SNPs in each of the three populations is predicted quite well by the model. We quantified the fit of the model to the observed data using a modified Pearson Chi-square goodness-of-fit (GOF) statistic, and found that the best-fitting complex demographic model is an excellent fit to the marginal *indica* ($GOF_I = 20.26$, $p = 0.72$) and *O. rufipogon* site-frequency spectra ($GOF_R = 7.57$; $p = 0.99$), and an adequate fit to the *tropical japonica* site-frequency spectrum ($GOF_T = 37.83$, $p = 0.22$). One interesting observation is that the demographic model underpredicts the excess of high-frequency derived alleles observed in *tropical japonica*—a potential indication of recent positive selection. Given that artificial selection was probably quite strong and frequent during and after domestication, we further explored models that incorporate selection during the domestication process of *O. sativa*.

Selection Models for Rice Domestication

Since strong selection is known to accompany crop domestication, we developed two alternative models incorporating multiple selective sweeps to explain the unusual polymorphism patterns in *indica* and *tropical japonica*. In a neutral locus linked to a single, recent selective sweep, let f_i be the probability of observing a neutral mutation segregating at frequency i in a sample of size n, conditional on the locus being variable. An expression for f_i has been derived [26] and further extended [37,38], and includes the genomic distance d (measured in bp) between neutral and selected loci, a compound parameter α, which represents the combined contributions of recombination, selection, and population size, and the "background" allele frequency distribution (i.e., the expected site-frequency spectrum for loci unlinked to a selected site).

These results for a single sweep can be used to predict the site-frequency spectrum at randomly chosen loci if multiple sweeps have recently occurred. Assuming that selective sweeps occur at random positions in the genome at a density of κ sweeps per bp, the distance between a random neutral locus and the nearest sweep will be approximately exponentially distributed with mean $1/(2\kappa)$. Define the function $\phi_i(d, \alpha, \kappa)$ to be the probability of observing i copies of a neutral mutation in a sample of n chromosomes, given that a sweep occurred at a distance d bp away with compound parameter α [38], and background site-frequency spectrum $\mathbf{q}$. By integrat-

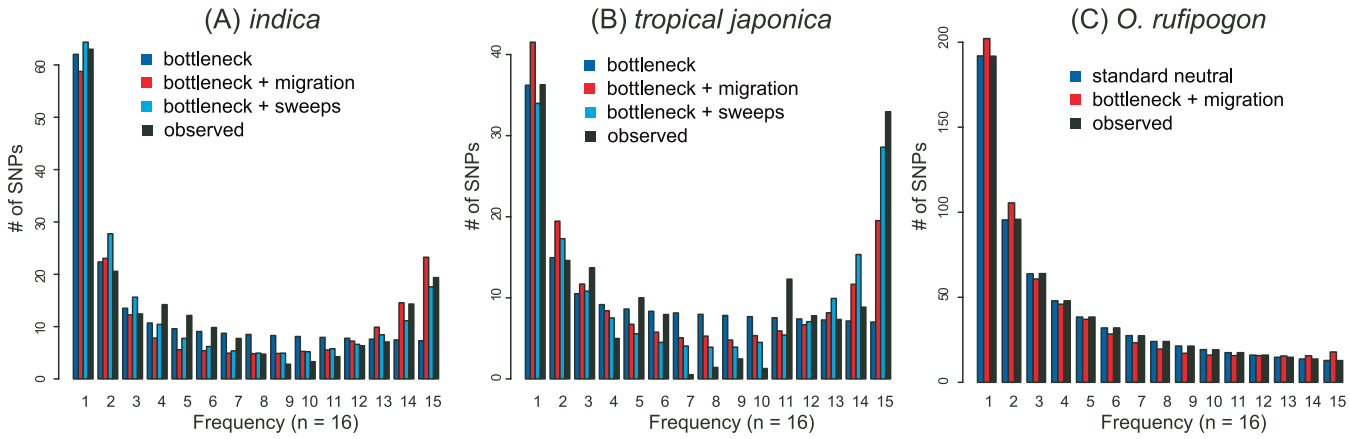


Figure 4. Observed and Expected Derived Site-Frequency Spectra under Various Models

The observed derived site-frequency spectrum for (A) *indica* and (B) *tropical japonica,* along with the expected site-frequency spectrum under the simple bottleneck, bottleneck plus migration demography, and bottleneck plus sweeps models. (C) Observed site-frequency spectrum for *O. rufipogon* and expected frequencies using a standard neutral model and a bottleneck plus migration model.

doi:10.1371/journal.pgen.0030163.g004

ing over the distance between the sampled locus and the unknown target of the sweep, the marginal probability, P_i, of observing a randomly chosen SNP at frequency i in a sample of n chromosomes is a function of κ, α, and $\mathbf{q}$ [38]:

$$P_i(\kappa, \alpha, \mathbf{q}) = \frac{\int_0^\infty \phi_i(d, \alpha, \mathbf{q}) e^{-2\kappa d} dd}{\sum_{j=1}^{n-1} \int_0^\infty \phi_j(d, \alpha, \mathbf{q}) e^{-2\kappa d} dd} \quad (1)$$

This probability can be used to calculate the composite likelihood of the data and estimate the parameters κ and α (see Materials and Methods). It should be noted that this equation assumes that the neutral locus is affected only by the nearest selective sweep.

We considered two distinct models. The first is a model in which strong selection is the only force that has acted in domesticated rice populations, and uses the normalized *O. rufipogon* site-frequency spectrum as the background frequency distribution. The second, a bottleneck plus sweeps model, allows multiple selective sweeps to affect patterns of variation immediately following a population size change. The background site-frequency spectrum in the latter case can be approximated using the predictions of a simplified neutral bottleneck model. The bottleneck plus sweeps model incorporates the sweep density κ, the compound parameter α (the combined contributions of recombination, selection, and population size), and a bottleneck severity parameter ν.

The likelihood surfaces for both the pure selection and the bottleneck plus sweeps model in rice each contains a long ridge where different parameter combinations have almost equally high likelihoods, implying that a model with high sweep density and relatively weak selection is just as likely as a model with low sweep density and strong selection (Figure 5). For both models, the ridge of maximum likelihood is shifted to the right in *tropical japonica*, indicating that for a given value of the selection severity parameter α, the sweep density in *tropical japonica* is estimated to be twice that in *indica*.

Sweep density is confounded with selection strength due to the effect of a mating system change on recombination rate. In domesticated rice, the transition to selfing likely occurred simultaneously with the sweeps, making it difficult to disentangle the recombination rate and selfing parameters. Under a recent selective sweep in a randomly mating population, the compound parameter $\alpha \approx rs^{-1} \ln(2N)$, where r is the per-basepair recombination rate, s is the selection coefficient and N is the population size [39]. In a partially selfing population such as domesticated rice, however, both effective recombination rate and population size are affected by selfing rate. While the rate of coalescence (and hence the effective population size) is at most doubled by the rate of selfing, the rate of recombination can be radically altered. An expression for effective recombination rate is $r(1-\sigma/[2-\sigma])$, where σ is the selfing rate [40]. For domesticated rice, estimates of selfing rates are typically $\sim$0.99 [13], resulting in a reduced recombination rate by approximately 10^{-3}. If we assume 400 selective sweeps occurred in the rice genome since domestication ($\kappa = 10^{-6}$), we estimate that $\alpha = 2 \times 10^{-12}$ for *indica*. With $r = 10^{-9}$ recombination events per generation per base pair and $\ln(2N) \approx 10$, this estimate of α corresponds to an unreasonably high estimate of a 5,000-fold fitness advantage. Substituting an effective recombination rate of 10^{-12} (corresponding to a reduced effective rate due to selfing), we find more reasonable values for the strength of selection for the selective sweeps, with $s \approx 5$. This example illustrates how high selfing rates can amplify the signal of selection and contribute to the pattern of polymorphism in the rice genome.

Comparing Models to Explain Patterns of Nucleotide Polymorphism in Rice

Visually, it appears both the bottleneck plus sweeps model and the bottleneck plus migration model predict the site-frequency spectrum of domesticated rice better than the bottleneck model alone (Figure 4) or the pure selection model (unpublished data). To compare likelihoods and determine which model best fits the data, we used the Akaike information criterion (AIC) [41]. Since SNPs in our dataset are linked, we used a composite likelihood function and

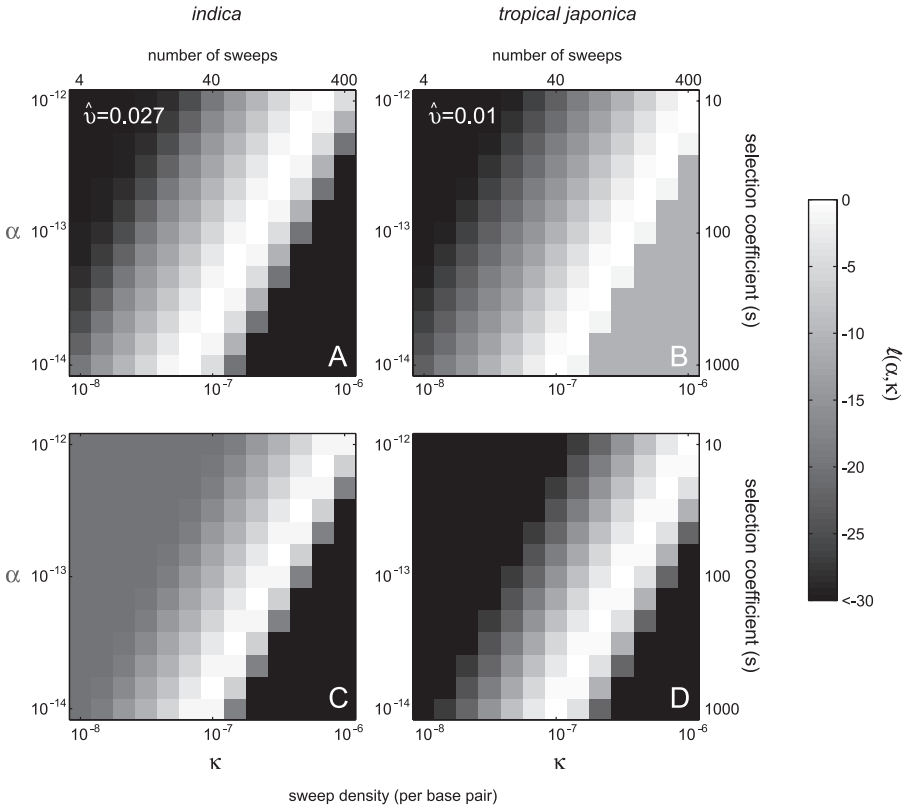


Figure 5. Composite Likelihood Surfaces in *indica* and *tropical japonica* under Models Incorporating Selection

A density plot of the marginal composite log-likelihood surface of the parameters α and κ, with the bottleneck severity υ fixed to its estimate, under the bottleneck plus sweeps model for (A) *indica* and (B) *tropical japonica*. The composite log-likelihood surface of the parameters α and κ under the pure selection model for (C) *indica* and (D) *tropical japonica*. The composite log-likelihood is represented as a deviation from the maximum log-likelihood, with lighter values representing higher composite likelihoods. Numbers above (A) and (B) indicate the total number of sweeps in the rice genome corresponding to each value of κ, and numbers to the right of (B) and (D) represent the selection coefficient, *s*, corresponding to each value of α, substituting an effective recombination rate of $r = 10^{-12}$ and $\ln(2N) = 10$ into the expression: $\alpha \approx rs^{-1} \ln(2N)$, then solving for *s*.

doi:10.1371/journal.pgen.0030163.g005

simulations to assign *p*-values to the observed AIC statistic (see Materials and Methods).

For *indica*, the bottleneck plus sweeps model is significantly better than the neutral bottleneck model ($\Lambda = -17.18$, $p < 0.05$) as is the bottleneck plus migration model ($\Lambda = -14.19$, $p < 0.05$). For *tropical japonica*, we also reject the neutral bottleneck model in favor of both the bottleneck plus sweeps model ($\Lambda = -56.88$, $p < 0.01$) and the bottleneck plus migration model ($\Lambda = -53.60$, $p < 0.01$). For both rice variety groups, the AIC for the bottleneck plus sweeps model was slightly lower than for the bottleneck plus migration models ($\Lambda = -2.26$, *indica*; $\Lambda = -3.28$, *japonica*), but this difference is likely not statistically meaningful given the various assumptions made. A separate (but not independent) assessment is comparing the fit of the predictions of each model to the data. The bottleneck plus sweeps model fits the marginal site-frequency spectrum of *indica* quite well (GOF = 13.86; $p = 0.92$), and does a slightly better job explaining the site-frequency spectrum of *tropical japonica* than does the complex demographic model incorporating bottlenecks plus migration ($GOF_{\text{sweeps + bottleneck}} = 31.21$, $p = 0.33$; $GOF_{\text{bottlenecks + migration}} = 37.83$; $p = 0.22$). These results underscore the importance of jointly modeling demographic and selective effects when considering the evolution of domesticated crop species.

Domestication and the Shaping of Genome-Wide Polymorphism Patterns in Rice

Population bottlenecks are believed to be the primary demographic event associated with crop species origins, and are the accepted mechanism to explain observed genome-wide polymorphism levels among these taxa. There have been concerted efforts to model the impact of population bottlenecks on domesticated species genomes [27–29,42–44]. It appears from our results, however, that a population bottleneck alone is inadequate to explain the observed nucleotide polymorphism patterns in rice, one of the oldest and the most predominant food crop species in the world.

A more complex demographic scenario involving very strong bottlenecks that led to the fixation of alternate alleles during the two rice domestication events, with concurrent gene flow between variety groups, can explain the site-frequency spectrum of *indica* and *O. rufipogon*. However, this pure demography model requires a bottleneck 4-fold stronger in *indica* and twice as strong in *tropical japonica* relative to the model that incorporates selection (Figure 5; Table 2), and a relatively high migration rate between domesticated rice and wild *O. rufipogon* populations. It is also important to note that the model is a poor fit to the observed frequency distribution of alleles in *tropical japonica*.

Domestication, however, is characterized by strong directional selection on a suite of traits that lead to the establishment of cultivated species as distinct entities from their wild progenitors within agricultural settings. We show that, in contrast to the complex demographic model, a simple bottleneck with sweeps model fits data from both *tropical japonica* and *indica* well without requiring an extremely strong domestication bottleneck. Since domesticated Asian rice has been subject to artificial selection, the selection plus demography model is a very plausible explanation for the observed strong excess of high-frequency derived alleles in domesticated rice varieties, and is consistent with recent reports about domestication genes in rice [45,46].

Positive selection on specific genes results in reductions in variation within a genome through selective sweeps [47,48]. Unlike bottlenecks, however, selection is thought to have largely localized effects on genome variation. Our results suggest that a model that incorporates selection can explain patterns of nucleotide variation in a set of genome-wide markers. We suggest two reasons why selective sweeps during domestication could cause a genome-wide effect in *O. sativa* and not in other cereal crop species such as maize. First, the origin of domesticated Asian rice is associated with a transition to self-fertilization, which results in a low effective recombination rate and greatly increases the genomic distance affected by selection. Second, *O. sativa* possesses such a small genome (<400 Mb) that it is likely that a few dozen to hundreds of selective sweeps could leave a genome-wide imprint.

Interestingly, under the bottleneck plus selective sweeps model, the dynamics of domestication appear to differ in significant ways between *indica* and *tropical japonica*. Despite the fact that these two variety groups were domesticated from the same species and both have contributed significantly to Asian agriculture, it appears that the number of selective events and/or the bottleneck severity differs between them. It is possible that the two subspecies would diverge from each other in the demographic patterns associated with domestication, given that they were established by different cultures. If this is correct, then *tropical japonica* appears to have undergone a more severe bottleneck associated with domestication. Alternatively, it may be that the establishment of *tropical japonica*, which includes landraces that expanded to upland growing areas, may be associated with stronger selection pressures on a larger number of traits.

The process of domestication is one of recent, rapid species evolution, and studies on the dynamics of this process inform our understanding of the origins and diversification of new species. Simple demographic scenarios that have been employed in the past may not fully capture the domestication process of some crop species such as Asian rice. Our models indicate that selection and population bottlenecks together, or more complex scenarios that invoke very strong bottlenecks and current gene flow, could be responsible for determining genome-wide variation in the rice genome, a finding that has not been described in other domesticated species. Domesticated crop species are particularly suitable subjects in which to study the interaction between demographic events and selection in shaping species characteristics, and exploring the relative contributions of these forces require developing predictions for patterns of DNA polymorphism using models that allow selection to vary in timing (i.e., both during and after population bottlenecks) and strength. Nevertheless, our findings do underscore the possible role that selection may play in shaping genomic variation in domesticated species, reinforcing our appreciation of the foresight showed by Charles Darwin nearly a century-and-a-half ago [3] when he sought to illustrate the power of selection by drawing on the lessons learned from the evolution of domesticated species.

Materials and Methods

Samples. A panel of 72 *O. sativa* accessions was chosen to represent the diversity found within the species. These include representatives of five major subpopulations identified in a previous study [14], including 21 *indica*, 18 *tropical japonica*, 21 *temperate japonica*, six *aus*, and six *aromatic* accessions (Table S1). Most accessions are landraces, but five accessions studied correspond to modern cultivars. Also included in the panel were 21 accessions of the wild progenitor of rice, *O. rufipogon*, along with one sample each of *O. nivara* (a close relative of *O. rufipogon* not believed to have contributed to the ancestry of cultivated rice) and the outgroup species *O. barthii* and *O. meridionalis* (Table S1).

DNA was extracted from single plants as described in [49] with minor modifications. All *O. sativa* and one *O. rufipogon* accession (International Rice Germplasm Collection [http://www.irri.org/grc/] #105491) were self-fertilized for two generations prior to initiating the study. Seeds from *O. rufipogon* from Nepal were collected in the field by H. J. Koh and colleagues (Seoul National University); all other seeds were obtained from germplasm repositories as summarized in Table S1.

PCR and DNA sequencing. A total of 121 approximately 400–600 bp gene regions across the rice genome were chosen at random for sequencing from a set of 6,591 ESTs [50]. Four fragments were also selected from genes coding for well-known allozymes, including: catalase, acid phosphatase, *pgi-a*, and *Adh*. Primers were designed from the Nipponbare genomic sequence available from Gramene using Primer3 [51]. Primers were designed in exons, and attempts were made to include both exon and intron sequence within each fragment. DNA sequencing was carried out in Genaissance's sequencing facilities (New Haven, Connecticut, United States) as described in [52]. Amplification and sequencing were successful for 111 fragments referred to as STS (Table S2). Approximately 54 kbp per accession were sequenced, composed of, on average, 55% coding and 45% noncoding sequence.

Base-pair calls, quality score assignment, and construction of contigs were carried out using the Phred and Phrap programs (Codon Code). Sequence alignment and editing were carried out with BioLign Version 2.09.1 (Tom Hall, North Carolina State University, Raleigh, North Carolina, United States). Heterozygous sites were identified with Polyphred (Deborah Nickerson, University of Washington, Seattle, Washington, United States) and by visually inspecting chromatograms for double peaks. Heterozygous sites were rare for *O. sativa*. For heterozygous *O. sativa* and *O. rufipogon* sequences, heterozygous sites were labeled with ambiguity codes. For all analyses, the published sequence of Nipponbare was included.

To assess the sequencing error rate, 18 randomly chosen STS fragments were resequenced in a single direction for four *Oryza* accessions. Only three discordant base pairs within a single individual in a single fragment sequence were observed. This corresponds to three errors in 33,193 resequenced bp, or a sequencing error rate of less than 0.01%.

Diversity analyses. Population structure among *O. sativa* and *O. rufipogon* accessions was evaluated with STRUCTURE 2.1 [19] using an admixture model with no linkage. To limit the effect of correlation between SNPs due to linkage, one SNP per fragment (the SNP with the highest minor allele frequency across the entire accession set) was used in the analysis. *O. sativa* is primarily selfing, and most accessions exist as homozygotes; thus, SNP data were considered haploid for this species. *O. rufipogon* is partially outcrossing, a condition that cannot be adequately represented by considering each locus as diploid; thus, SNP data for *O. rufipogon* were also considered haploid. Because alternate alleles could occur at a given site in heterozygous *O. rufipogon* accessions, ten datasets were created with randomly chosen alternative base pairs in heterozygous individuals. Analyses were carried out for all ten datasets. All analyses had a burn-in length of 50,000 iterations and a run length of 100,000 iterations. Three replicates at each value of K (population number) were carried out.

Simulations were run with uncorrelated allele frequencies. Results were entirely consistent among replicate runs within datasets and among datasets; the results from one run are presented in Figure 1 and Table S1.

To assess relationships among *Oryza* accessions, all STS fragment alignments were concatenated to form a single dataset. Relationships were estimated with a neighbor-joining analysis as implemented in PAUP* version 4.0 b3 [53]. Distances were calculated using the Kimura two-parameter model. Branch bootstrap estimates were obtained from 1,000 replicates.

Perl scripts were written to assess levels of nucleotide variation (θ_W) and nucleotide diversity (θ_π) and Tajima's D across rice groups for all STS fragments, and to calculate the frequency distributions of derived SNPs across the genome. For *O. sativa* accessions, where heterozygotes were rare, all measures were calculated considering each accession as contributing a single haplotype; for *O. rufipogon* population measures, each accession was considered to contribute two haplotypes, except for one accession (International Rice Germplasm Collection [http://www.irri.org/grc/] #105491) from Malaysia, which had been selfed for several generations prior to this study.

Analysis of the neutral bottleneck model. Under a neutral bottleneck model, the history of rice domestication is represented by recent population divergence, with one of the two daughter populations experiencing a size bottleneck at divergence associated with the founder effect at the time of domestication. We use the sample frequencies of variable noncoding and synonymous nucleotides in the STS alignments (i.e., the site-frequency spectrum of putatively neutral SNPs) to infer the parameters of the bottleneck model. Our analytical approach makes use of standard Wright-Fisher population genetic theory within a Poisson random field setting [54–57]. The assumptions of this model include independence among SNPs, no selection, an underlying Poisson process governing mutations, and a piecewise constant population of large size amenable to modeling using diffusion approximations.

The model we employ is an extension of Williamson et al. [58], where we present the relevant population and statistical inference theory for modeling a population experiencing a recent size change. The key addition to our previous model is a second size change event, corresponding to the post-bottleneck growth phase. This amounts to modeling the components of the site-frequency spectrum (X_1, X_2, . . ., X_n) as independent Poisson random variables with mean:

$$E(X_i) = \frac{\theta}{2}\int_0^1 \binom{n}{i} x^i (1-x)^{n-i} f(x;\Theta)dx \qquad (2)$$

where θ is the genome-wide mutation rate, x represents the (unknown) population frequencies of mutations, and $f(x;\Theta)$ is the distribution of mutation frequencies given demographic history parameters $\Theta = \{\nu,\tau_1,\tau_2\}$. These parameters are: the time back until the start of domestication (τ_1), duration of the bottleneck (τ_2), ratio of current population to ancestral population size (ν_2), and relative size of the bottleneck population to the ancestral population (ν_b). The duration of the bottleneck was assumed to be 25% of the time back until domestication ($\tau_2 = 0.25 \times \tau_1$), which is consistent with archaeological data suggesting domestication took 3,000 y and began 12,000 y ago. The mutation rate, θ, was estimated from the number of synonymous and noncoding segregating SNPs assuming *O. rufipogon* represented a population of constant size. This assumption is quite reasonable given the excellent concordance between the *O. rufipogon* and the predictions of the standard neutral model (Figure 4), and is equivalent to using Watterson's (1975) estimator of θ. In order to account for missing data, we fitted the population bottleneck model using the projected site-frequency spectrum for a sample of $n = 16$ chromosomes.

Alternative demographic scenarios for rice domestication. We considered alternative demographic scenarios, in which ancestral population subdivision, followed by gene flow between *rufipogon*, *indica*, and *tropical japonica*, led to an excess of high-frequency derived alleles in domesticated rice groups, as well as a simpler model that has no ancestral substructure. For these models, the composite likelihood function was based on the marginal site-frequency spectrum of each of the three groups analyzed. For ease of notation, let S_{ind}, S_{jap}, and S_{ruf} be the number of SNPs for which we could distinguish ancestral from derived alleles using the outgroup (223, 172, and 636, respectively). Let $\boldsymbol{y}$ denote the set of derived allele counts for each SNP, with $\boldsymbol{y}_\bullet^{ind}$, $\boldsymbol{y}_\bullet^{jap}$, and $\boldsymbol{y}_\bullet^{ruf}$ referring to set of SNPs for *indica*, *tropical japonica*, and *O. rufipogon* (with lengths S_{ind}, S_{jap}, and S_{ruf}, respectively). To account for missing data, let $\boldsymbol{n}$ refer to the number of chromosomes sequenced at each SNP, with $\boldsymbol{n}_\bullet^{ind}$, $\boldsymbol{n}_\bullet^{jap}$, and $\boldsymbol{n}_\bullet^{ruf}$ the vector for each group (again with lengths S_{ind}, S_{jap}, and S_{ruf}, respectively). For a given demographic model discussed above (the parameters of which we collectively denote Θ), the composite likelihood function is written as

$$\begin{aligned} L(\mathbf{y}|\Theta) = &\{\Pr(S_{ind}|\Theta)\prod_{k=1}^{S_{ind}}\Pr(y_k^{ind}, n_k^{ind}|\Theta)\} \\ &\times\{\Pr(S_{jap}|\Theta)\prod_{k=1}^{S_{jap}}\Pr(y_k^{jap}, n_k^{jap}|\Theta)\} \\ &\times\{\Pr(S_{ruf\,i}|\Theta)\prod_{k=1}^{S_{ruf}}\Pr(y_k^{rui}, n_k^{ruf}|\Theta)\} \end{aligned} \qquad (3)$$

where $\Pr(S_\bullet|\Theta)$ is assumed to follow a Poisson probability of observing $S_\bullet$ SNPs in a given population under the demographic model Θ assuming the population scaled mutation rate $\theta = 148.6$ (estimated using the observed number of SNPs in *O. rufipogon*), and $\Pr(y^\cdot_\cdot, n^\cdot_\cdot|\Theta)$ is the probability of observing a SNP configuration in a given population under the demographic model. It is important to note that the inference scheme assumes the allele frequency distributions, conditional on the observed number of segregating sites and demographic parameters, are independent among populations. This composite-likelihood function (like all composite-likelihood functions) must, therefore, be taken as an approximation of the true likelihood function since it ignores dependencies among SNPs due to linkage and among populations due to shared variation. To account for missing data at an arbitrary SNP k in population x, we set

$$\Pr(y_k^x, n_k^x|\Theta) = \begin{cases} P_{y_k^x}(\Theta, N_x) & \text{if no missing data} \\ \sum_{j=y_k^x}^{N_X-(n_k^x-y_k^x)} \left[\frac{\binom{j}{y_k^x}\binom{N_x-j}{n_k^x-y_k^x}}{\binom{N_x}{n_k^x}} P_j(\Theta, N_x)\right] & \text{otherwise} \end{cases} \qquad (4)$$

where $P_z(\Theta,N_x)$ is the expected proportion of SNPs at a frequency z in a sample of N_x chromosomes under the demographic model Θ, and the fraction within the summation represents the hypergeometric probability of sampling y_k^x derived alleles in a subsample of n_k^x chromosomes if the unknown frequency of the SNP were j out of N_x (summed over all possible underlying SNP frequencies, j). Details on calculating the expected number of SNPs in each population as well as $P_z(\Theta,N_x)$ are described below.

Optimizing complex neutral demographic models. For a given set of parameters, Θ, we determine the expected site-frequency spectra for all three populations (*O. rufipogon*, *indica*, and *tropical japonica*) using 100,000 iterations of the coalescent simulation program *ms* [1] conditional on the observed genome-wide estimate of θ for *O. rufipogon*. To generate data under this model, we used the following code:

ms 80 200000 –t 148.6487 –r 148.6487 111 –I 3 21 18 41 *M* –en 0.5*0.75*τ_1 1 ν_B –en 0.5*0.75*$\tau_1$2ν_B –ej 0.5*τ_1 1 3 –ej τ_1 2 3 –em 0.5*τ_1 3 1 0 –em 0.5*τ_1 3 2 0 –n 1 ν_I –n 2 ν_J.

Note that the factor 0.75 enters from the assumption that the bottleneck lasted 3,000 y of the 12,000 y time since domestication began, and 0.5 enters since *ms* scales time in units of $4N$ generations.

To optimize the three- and five-dimensional likelihood surface, we used an iterative technique, whereby a very coarse grid is initially chosen for each parameter, followed by successively tighter intervals containing the previous iteration's maximum likelihood estimates. Because we were pooling data across 111 STS loci, we generated our expected site-frequency spectrum accordingly. Although recombination within or between STS loci will not affect the expected number of segregating sites or the expected site frequency spectrum under a neutral demographic model, it does impact the rate at which simulations will approach them. We therefore assumed 111 mostly independent loci of equal size when generating our expectations.

Modified Pearson Goodness-of-Fit test. In order to compare the fit of the demographic model to the observed data accounting for missing genotypes and partial selfing, we considered a projection of the observed and predicted site-frequency spectra into a sample of size $n = 16$ chromosomes from each of the three populations using the hypergeometric distribution. The "observed" data can be thought of as the predicted SFS in a subsample of $n = 16$ based on the actual SNP data assuming each of the *O. sativa* accessions contributes one chromosome to the observed allele frequency spectrum, and each of

the *O. rufipogon* accessions contributes two, with the exception of one accession that was known to have been purified. The "expected" data are the predicted marginal site-frequency spectrum at the maximum composite-likelihood estimates of the parameters from the complex demographic model that includes bottlenecks in the two domesticated populations, migration within domesticated populations, and migration between domesticated and ancestral populations. There were 45 observed data points (15 segregating site-frequency spectrum components multiplied by three populations), and the GOF statistic for a given population was tabulated as $GOF = \sum_{i=1}^{15} \frac{(Obs_i - Exp_i)^2}{Exp_i}$. In order to assign a *p*-value, we simulated 10,000 datasets each containing 111 independent loci with no recombination within loci under the best-fitting demographic model. For each dataset, we then calculated the GOF test statistic using the expected site-frequency spectrum from Figure 4 scaled to the observed number of segregating sites within each of the subpopulations. Ideally, one would re-estimate the demographic parameters in order to fully mimic the inference procedure we used. Unfortunately, estimation of the demographic parameters was extremely computationally intensive for each dataset; the single observed STS data point analyzed here, for example, took over a week of computer time on a dedicated 100-node computing cluster.

Composite likelihood under multiple sweeps models. Conditioning on the observed number of segregating sites in the dataset, the site-frequency spectrum is multinomially distributed with frequency probabilities according to Equation 1. For the pure selection model, the composite likelihood is:

$$\ell_s(\kappa, \alpha|\mathbf{x}) = \sum_{i=1}^{n-1} x_i \log(P_i(\kappa, \alpha, \mathbf{q}_r)) \quad (5)$$

where $\mathbf{q}_r$ is the normalized site-frequency spectrum of *O. rufipogon*. For the bottleneck plus multiple sweeps model, the composite likelihood is:

$$\ell_s(\kappa, \alpha, \nu|\mathbf{x}) = \sum_{i=1}^{n-1} x_{\bullet i} \log(P_i(\kappa, \alpha, \mathbf{q}_\nu)) \quad (6)$$

where $\mathbf{q}_\nu$ is the predicted spectrum from a neutral bottleneck model with severity ν. Equations 5 and 6 can be maximized to quantify the number and strength of selective sweeps in domestic rice, and the optimization of Equation 5 provides an estimate of the severity of the population bottleneck that preceded the selective sweep.

The background site-frequency spectrum for the bottleneck plus multiple sweeps model. The bottleneck plus sweeps model assumes that a short bottleneck (representing to the founding of domestic populations) precedes the selective sweeps. To calculate the background site-frequency spectrum at the end of the bottleneck and the beginning of the selective sweeps, we again used numerical methods to solve the one-population diffusion equation with population size changes:

$$\frac{\partial}{\partial t} f(q,t) \frac{1}{2} \frac{\partial^2}{\partial q^2} \left\{ \frac{q(1-q)}{\nu(t)} f(q,t) \right\} \quad (7)$$

In this case, the recovery time, τ_1, was set to 0, corresponding to the assumption that new mutations since the bottleneck do not make a strong contribution to the observed SFS. Because the bottleneck duration, τ_2, and the severity, ν, are confounded parameters, we set $\tau_2 = 0.01$ and allow ν to vary. With $f(q,\tau_2)$ as the numerical solution to Equation 7 evaluated at time τ_2, we calculate the background site-frequency spectrum $q\nu$ as:

$$q_\nu[i] = \int_0^1 \binom{n}{i} q^i (1-q)^{n-i} f(q, \tau_2) dq \quad (8)$$

AIC as a test statistic for comparing non-nested models. To properly interpret differences in AIC between models, we simulated 10,000 datasets of 111 nonrecombining loci under the null hypothesis of the best-fitting neutral bottleneck model using the ms coalescent simulation program [59]. Because we did not allow recombination within loci, these simulations conservatively account for the effects of linkage. For each simulated dataset, we found the maximum composite likelihoods under each model (bottleneck, bottleneck plus migration, multiple sweeps, and bottleneck plus sweeps) and calculated the AIC value. The AIC statistic of model *i* is defined as: $AIC_i = -2(lmax_i - k_i)$ where $lmax_i$ is the maximum likelihood under model *i* and k_i is the number of free parameters in model *i*. We used $\Lambda = AIC_1 - AIC_2$ as a test statistic for comparing the bottleneck and alternative models using a one-tailed test: the *p*-value was estimated as the proportion of simulations under the null distribution with $\Lambda > \Lambda_{obs}$.

Supporting Information

Figure S1. Clustering of *Oryza* Accessions Based on Neighbor-Joining Analysis of Concatenated STS Sequences

Numbers by branches are bootstraps of 1,000 replicates. Only branches with a bootstrap value higher than 60% for major clades (five or more accessions) are labeled. The monophyly of each rice variety group is well supported, with the exception of *tropical japonica*. The accession M202-new is an elite *temperate japonica* line that has been subjected to possible crosses with other groups, perhaps explaining its inclusion within the *tropical japonica*.

Found at doi:10.1371/journal.pgen.0030163.sg001 (409 KB EPS).

Figure S2. Frequency Distribution of Tajima's D Values for All STS Sampled in (A) *indica*, (B) *tropical japonica*, and (C) *O. rufipogon*

Found at doi:10.1371/journal.pgen.0030163.sg002 (373 KB EPS).

Figure S3. The Distribution of Allele Frequency in *O. rufipogon* for Derived Alleles That Are at High Frequency in *indica* or *tropical japonica*

Notably, most alleles are at low to intermediate frequency in *O. rufipogon*, consistent with multiple selective sweeps in *O. sativa*, and discounting the possibility of misidentification of ancestral alleles or interspecific introgression being responsible for the pattern observed in rice. HFD, high-frequency derived SNPs.

Found at doi:10.1371/journal.pgen.0030163.sg003 (390 KB EPS).

Figure S4. The Genomic Distribution of STS Fragments Contributing High-Frequency Derived SNPs in *indica* and *tropical japonica*

In each group, high-frequency derived SNPs occur in ten of 12 rice chromosomes. Fragments containing high-frequency derived SNPs comprise a large portion of fragments containing any variation at all in each *O. sativa* group. In both rice groups, the sample of STS fragments used to construct the site-frequency spectrum is slightly lower than 111 due to missing data in *O. meridionalis*.

Found at doi:10.1371/journal.pgen.0030163.sg004 (433 KB EPS).

Table S1. *Oryza* Accessions Used in the Study and Inferred Ancestry Coefficients

Found at doi:10.1371/journal.pgen.0030163.st001 (45 KB XLS).

Table S2. STS Fragment Information and Silent Sites Diversity Measures for the Various *Oryza* Groups

Found at doi:10.1371/journal.pgen.0030163.st002 (133 KB XLS).

Accession Numbers

The National Center for Biotechnology Information GenBank (http://www.ncbi.nlm.nih.gov/Genbank) ID numbers for the sequences and alignments discussed in this article are EF000002–EF010509.

Acknowledgments

We are grateful to two anonymous reviewers for suggestions that much improved the manuscript.

Author contributions. RN, SRM, CDB, and MDP conceived the experiments. ALC, KMO, and MDP designed the experiments. ALC collected the data. ALC, SHW, RDH, and CDB analyzed the data. AB, AFA, NRP, TLY, and SRM and contributed materials/analysis tools. ALC, SHW, CDB, and MDP wrote the paper.

Funding. This work was funded by the US National Science Foundation Plant Genome Research Program..

Competing interests. The authors have declared that no competing interests exist.

References

1. Hancock JF (2004) Plant evolution and the origin of crop species. Cambridge (Massachusetts): CABI Publishing. 313 p.
2. Mannion AM (1999) Domestication and the origins of agriculture: an appraisal. Prog Phys Geogr 23: 37–56.
3. Darwin C (1859) On the origin of species. London: John Murray.
4. Darwin C (1868) The variation of animals and plants under domestication. London: John Murray.
5. Armelagos GJ, Harper KN (2005) Genomics at the origins of agriculture, part one. Evol Anthropol 14: 68–77.
6. Diamond J (2002) Evolution, consequences and future of plant and animal domestication. Nature 418: 700–707.
7. FAO (2005) FAOSTAT data, in last update 2003. http://faostat.fao.org/site/346/DesktopDefault.aspx?PageID=346. Accessed 1 August 2007.
8. Yu J, Hu SN, Wang J, Wong GKS, Li SG, et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp indica). Science 296: 79–92.
9. Goff SA, Ricke D, Lan TH, Presting G, Wang RL, et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp japonica). Science 296: 92–100.
10. IRGSP (2005) The map-based sequence of the rice genome. Nature 436: 793–800.
11. Takahashi N, Hamamura K, Tsunoda S, Sakamoto S, Sato Y (1997) Differentiation of ecotypes in cultivated rice. In: Matsuo T, Futsuhara Y, Kikuchi F, Yamaguchi H, editors. Science of the rice plant. Tokyo: Food and agriculture policy research center. pp. 112–160.
12. Jackson MT (1997) Conservation of rice genetic resources: the role of the International Rice Genebank at IRRI. Plant Mol Biol 35: 61–67.
13. Oka HI (1988) Origin of cultivated rice. Tokyo: Japan Scientific Societies Press and Elsevier Science Publishers. 254 p.
14. Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in Oryza sativa L. Genetics 169: 1631–1638.
15. Glaszmann JC (1987) Isozymes and classification of Asian rice varieties. Theor Appl Genet 74: 21–30.
16. Tenaillon MI, Sawkins MC, Anderson LK, Stack SM, Doebley J, et al. (2002) Patterns of diversity and recombination along chromosome 1 of maize (Zea mays ssp mays L.). Genetics 162: 1401–1413.
17. Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005) A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 169: 1601–1615.
18. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, et al. (2005) The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3: 1289–1299.
19. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
20. Cheng CY, Motohashi R, Tsuchimoto S, Fukuta Y, Ohtsubo H, et al. (2003) Polyphyletic origin of cultivated rice: based on the interspersion pattern of SINEs. Mol Biol Evol 20: 67–75.
21. Zhu QH, Ge S (2005) Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. New Phytol 167: 249–265.
22. Vitte C, Ishii T, Lamy F, Brar D, Panaud O (2004) Genomic paleontology provides evidence for two distinct origins of Asian rice (Oryza sativa L.). Mol Genet Genomics 272: 504–511.
23. Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA (2006) Phylogeography of Asian wild rice, Oryza rufipogon, reveals multiple independent domestications of cultivated rice, Oryza sativa. Proc Natl Acad Sci U S A 103: 9578–9583.
24. Second G (1982) Origin of the genic diversity of cultivated rice (Oryza-Spp)—study of the polymorphism scored at 40 isoenzyme loci. Jpn J Genet 57: 25–57.
25. Tajima F (1989) Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
26. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.
27. Wright SI (2005) The effects of artificial selection on the maize genome. Science 310: 54–54.
28. Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol 21: 1214–1225.
29. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS (1998) Investigation of the bottleneck leading to the domestication of maize. Proc Natl Acad Sci U S A 95: 4441–4446.
30. Takahata N (1991) Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. Genetics 129: 585–595.
31. Tachida H, Iizuka M (1991) Fixation probability in spatially changing environments. Genet Res 58: 243–251.
32. Maruyama T (1970) On fixation probability of mutant genes in a subdivided population. Genet Res 15: 221-&.
33. Lu HY, Liu ZX, Wu NQ, Berne S, Saito Y, et al. (2002) Rice domestication and climatic change: phytolith evidence from East China. Boreas 31: 378–385.
34. Zhao ZJ (1998) The middle Yangtze region in China is one place where rice was domesticated: phytolith evidence from the Diaotonghuan cave, northern Jiangxi. Antiquity 72: 885–897.
35. Normile D (1997) Archaeology—Yangtze seen as earliest rice site. Science 275: 309–309.
36. Wakeley J, Hey J (1997) Estimating ancestral population parameters. Genetics 145: 847–855.
37. Kim Y, Stephan W (2003) Selective sweeps in the presence of interference among partially linked loci. Genetics 164: 389–398.
38. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566–1575.
39. Durrett R, Schweinsberg J (2004) Approximating selective sweeps. Theor Popul Biol 66: 129–138.
40. Nordborg M (2000) Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154: 923–929.
41. Burnham KP, Anderson DR (1998) Model selection and inference: a practical information-theoretic approach. New York: Springer-Verlag. 353 p.
42. Vigouroux Y, Mitchell S, Matsuoka Y, Hamblin M, Kresovich S, et al. (2005) An analysis of genetic diversity across the maize genome using microsatellites. Genetics 169: 1617–1630.
43. Thuillet AC, Bataillon T, Poirier S, Santoni S, David JL (2005) Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. Genetics 169: 1589–1599.
44. Muller MH, Poncet C, Prosperi JM, Santoni S, Ronfort J (2006) Domestication history in the Medicago sativa species complex: inferences from nuclear sequence polymorphism. Mol Ecol 15: 1589–1602.
45. Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. Science 311: 1936–1939.
46. Sweeney MT, Thomson MJ, Cho Y, Park YJ, Williamson SH (2007) Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genet 3: e133. doi:10.1371/journal.pgen.0030133
47. Maynard Smith J, Haigh J (2004) The hitch-hiking effect of a favourable gene. Genet Res 23: 23–35.
48. Kaplan NL, Hudson RR, Langley CH (1989) The hitchhiking effect revisited. Genetics 123: 887–899.
49. McCouch SR, Kochert G, Yu ZH, Wang ZY, Khush GS, et al. (1988) Molecular mapping of rice chromosomes. Theor Appl Genet 76: 815–829.
50. Wu JZ, Maehara T, Shimokawa T, Yamamoto S, Harada C, et al. (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. Plant Cell 14: 525–535.
51. Rozen S, Skaletsky HJ (2000) In: Krawetz S, Misener S, editors. Bioinformatics methods and protocols: methods in molecular biology. Totowa (New Jersey): Humana Press. pp. 365–386.
52. Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, et al. (2006) Selection under domestication: evidence for a sweep in the rice Waxy genomic region. Genetics 173: 975–983.
53. Swofford DL (2000) PAUP* Phylogenetic analysis using parsimony (* and other methods). Sunderland, Massachusetts: Sinauer Associates.
54. Sawyer SA, Hartl DL (1992) Population-genetics of polymorphism and divergence. Genetics 132: 1161–1176.
55. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159: 1779–1788.
56. Bustamante CD, Nielsen R, Hartl DL (2003) Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. Theor Popul Biol 63: 91–103.
57. Williamson S, Fledel-Alon A, Bustamante CD (2004) Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. Genetics 168: 463–475.
58. Williamson S, Perry SM, Bustamante CD, Orive ME, Stearns MN, et al. (2005) A statistical characterization of consistent patterns of human immunodeficiency virus evolution within infected patients. Mol Biol Evol 22: 456–468.
59. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.