Deleterious variants in Asian rice and the potential cost of domestication

Qingpo Liu[1], Yongfeng Zhou[2], Peter L. Morrell[3] and Brandon S. Gaut[2]

1. College of Agriculture and Food Science, Zhejiang A&F University, Lin'an, Hangzhou 311300, People's Republic of China
2. Dept. of Ecology and Evolutionary Biology, UC Irvine, Irvine, CA 92697-2525
3. Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108 -6026

Addresses for correspondence:

Brandon S. Gaut
Dept. of Ecology and Evolutionary Biology
321 Steinhaus Hall
UC Irvine,
Irvine, CA 92697-2525
Phone: 949 824-2564
Email: bgaut@uci.edu

Qingpo Liu
College of Agriculture and Food Science
Zhejiang A&F University,
Lin'an, Hangzhou 311300
People's Republic of China
Phone: 086-571-63741276
Email: liuqp@zafu.edu.cn

Running title: Deleterious variants in rice

ABSTRACT

SNPs that are predicted to encode deleterious amino acid variants provide unique insights into population history, the dynamics of selection, and the genetic bases of phenotypes. This may be especially true for domesticated species, where a history of bottlenecks and selection can contribute to the accumulation of deleterious SNPs (dSNPs). Here we investigate the numbers and frequencies of deleterious variants in Asian rice (*O. sativa*), focusing on two separate varieties (*japonica* and *indica*) that may have been domesticated independently. Comparative analyses in two population datasets each for *japonica* and *indica* rice -- using SNPs identified in separate variant calling pipelines and applying two distinct tools for the prediction of deleterious variants -- were consistent in indicating that the transition to domesticated rice has shifted site frequency spectra for all derived variants but particularly for dSNPs. This potential "cost of domestication" is higher in genomic regions of low recombination and within regions of putative selective sweeps. A characteristic feature of rice domestication was a shift in mating system from outcrossing to predominantly selfing. Using forward simulations, we show that this shift in mating system likely ameliorates the cost of domestication through purging of deleterious variants.

INTRODUCTION

Several studies have suggested that there is a "cost of domestication" (Schubert et al., 2014), because domesticated species may accumulate deleterious mutations that reduce their relative fitness (Lu et al., 2006). Under this hypothesis, the small effective population size (N_e) during a domestication bottleneck reduces the efficacy of genome-wide selection (Charlesworth and Willis, 2009), leading to the accumulation of deleterious variants (Lohmueller et al., 2008; Casals et al., 2013). The fate of deleterious variants also relies on linkage, because selection is less effective in genomic regions of low recombination (Hill and Robertson, 1966; Felsenstein and Yokoyama, 1976) and because deleterious variants may hitchhike with alleles that are positively selected for agronomic traits (Fay and Wu, 2000; Hartfield and Otto, 2011; Campos et al., 2014). Overall, the cost of domestication is expected to increase the prevalence of deleterious variants in small relative to large populations, in regions of low recombination, and near sites of positive selection.

This hypothesis about the cost of domestication closely parallels the debate regarding the genetic effects of migration-related bottlenecks (Lohmueller et al., 2008; Casals et al., 2013) and demographic expansion (Peischl et al., 2013) in human populations. The debate regarding human population is contentious, perhaps in part because it suggests that some human populations may, on average, carry a greater load of deleterious variants than others. Studies in humans also suggest that subtlety of interpretation is required

when considering the relative frequency of deleterious variants in populations; both the effect size and relative dominance (Henn et al., 2016) of deleterious variants likely play a role in how mutations impact the fitness of populations. Moreover, deleterious variants in non-equilibrium populations, such as those that have experienced a recent bottleneck, may also return to pre-bottleneck frequencies more rapidly than neutral variants (Brandvain and Wright, 2016). It nonetheless remains an important task to identify the frequency and genomic distribution of deleterious variants in humans, for the purposes of disentangling evolutionary history and for understanding the association between deleterious variants and disease (Kryukov et al., 2007; Eyre-Walker, 2010).

In plant crops, the potential for the accumulation of deleterious variants was first examined in Asian rice (*O. sativa*) (Lu et al., 2006). At the time, few resequencing data were available, so Lu et al (2006) compared two *O. sativa* reference genomes to that of a related wild species (*O. brachyntha*) (Lu et al., 2006). They found that radical, presumably deleterious amino acid variants were more common within *O. sativa* genomes, suggesting a cost of domestication. A handful of studies have since analyzed deleterious variants in crops based on resequencing data (Gunther and Schmid, 2010; Nabholz et al., 2014; Renaut and Rieseberg, 2015; Kono et al., 2016), and together they suggest that the accumulation of deleterious variants is a general outcome of domestication. More limited analyses have also shown that deleterious variants are enriched within genes associated with phenotypic traits (Mezmouk and Ross-Ibarra, 2014; Kono et al., 2016), suggesting that the study of deleterious variants is crucial for

understanding potentials for crop improvement (Morrell et al., 2011). While a general picture is thus beginning to emerge, most of these studies have suffered from substantial shortcomings, such as small numbers of genes, low numbers of individuals, or the lack of an outgroup to infer ancestral states. Moreover, no study of crops has yet investigated the prevalence of deleterious variants in putative selective sweep regions.

In this study, we reanalyze genomic data from hundreds of accessions of Asian rice and its wild relative *O. rufipogon.* Asian rice feeds more than half of the global population (Project, 2005), but the domestication of the two main varieties of Asian rice (ssp. *japonica* and ssp. *indica*) remains enigmatic. It is unclear whether the two varieties represent independent domestication events (Londo et al., 2006; Civian et al., 2015), a single domestication event with subsequent divergence (Gao and Innan, 2008; Molina et al., 2011), or separate events coupled with substantial homogenizing gene flow of beneficial domestication alleles (Caicedo et al., 2007; Sang and Ge, 2007; Zhang et al., 2009; Huang et al., 2012a; Huang et al., 2012b). It is clear, however, that domestication has included a shift in mating system: from predominantly outcrossing *O. rufipogon* [which has outcrossing rates between 5% and 60%, depending on the population of origin and other factors (Oka and Miroshima, 1967)] to predominantly selfing rice [which has outcrossing rates of ~1% (Oka, 1988)]. This shift in mating system has the potential to affect the population dynamics of deleterious variants, because inbreeding exposes partially recessive variants to selection (Lande and Schemske, 1985), which may in turn facilitate purging of deleterious alleles (Arunkumar et al., 2015).

Commensurate with its agronomic importance, the population genetics of Asian rice have been studied in great detail. Comparative resequencing studies have estimated that nucleotide sequence diversity is ~2 to 3-fold higher in *indica* varieties than in *japonica* varieties (Zhu et al., 2007; Huang et al., 2012b), which are often separated into temperate and tropical germplasm. Sequence polymorphism data have also shown that the derived site frequency spectrum (SFS) of both varieties exhibits a distinct U-shaped distribution relative to *O. rufipogon*, due either to the genome-wide effects of selection or migration (Caicedo et al., 2007). Surprisingly, however, the population genetics of putatively deleterious variants have not been studied across *O. sativa* genomes.

In this study, we reanalyze genomic data from hundreds of *indica*, *japonica,* and *O. rufipogon* accessions to focus on the population frequencies of putatively deleterious genetic variants. To assess the robustness of our results, we have utilized two *O. sativa* datasets: one with many accessions (n = 776) but low sequencing coverage (1-2x), the other with fewer individuals (n = 45) but enhanced (>12x) coverage. For both datasets, we have re-mapped raw reads and then applied independent computational pipelines for SNP variant detection. We have also used two different approaches – PROVEAN (Choi et al., 2012) and SIFT (Kumar et al., 2009)- to predict deleterious variants from nonsynonymous SNPs. Armed with consistent results from multiple datasets and different methodological approaches, we address four questions. First, does the number, proportion, or frequency of deleterious mutations reflect a 'cost of domestication' in Asian rice, despite the potential for purging associated with a shift to inbreeding? Second,

does the number of deleterious variants vary with recombination rate, suggesting a pervasive effect of linkage? Third, is the accumulation of deleterious variants exacerbated in regions of the genome that may have experienced a selective sweep? Finally, can we garner insights into the relative contributions of demography, linkage, positive selection, and inbreeding for the accumulation of deleterious variants?

RESULTS

Data sets

To investigate the population dynamics of deleterious variants, we collated two rice datasets. The first was based on the genomic data of 1,212 accessions reported in Huang et al. (2012b) (Table S1). This first dataset, which we call the 'BH' data after the senior author, contains raw reads from 766 individuals of Asian rice, including 436 *indica* accessions and 330 *japonica* accessions. The BH dataset also included 446 accessions representing three populations of *O. rufipogon,* the wild ancestor of cultivated rice (Table 1). For these data, we remapped sequencing reads to the *japonica* reference sequence (Goff et al., 2002), then used ANGSD (Korneliussen et al., 2014) to apply cut-offs for quality and coverage and to estimate SFS (see Materials and Methods).

The second dataset, which we call the '3K' data (Li et al., 2014), consisted of 15 cultivated, high-coverage (>12X) accessions for each of *indica*, tropical *japonica,* and temperate *japonica* (Table S2). This dataset was used primarily to assess the robustness of results based on the larger, lower coverage dataset, but these data were also employed

to examine regions of selective sweeps. For this dataset, reads were again mapped to the *japonica* reference, but SNPs were called using tools from GATK and SAMtools (see Materials and Methods).

Huang et al (2012b) determined that their *O rufipogon* accessions represented three different wild populations, which we denote W_I, W_{II} and W_{III}. They also inferred that W_I was ancestral to *indica* rice and that W_{III} was ancestral to *japonica* rice. Accordingly, we base our cultivated-to-wild comparisons on *indica* vs. W_I and *japonica* vs. W_{III} for the BH data; when appropriate we also include genetic comparisons to the complete set of wild accessions (W_{all}).

The number and frequency of deleterious variants

Based on reads in the BH dataset, we identified between 243,964 and 388,615 SNPs from the cultivated samples and >1.8M SNPs from each of W_I, W_{II} and W_{III} wild populations (Table 1). Despite fewer accessions, we identified more SNPs within the 3K data, owing to higher sequence coverage (Table 1). Once identified, we annotated SNPs as either non-coding (ncSNPs), synonymous (sSNPs), Loss of Function (LoF) or nonsynonymous. LoF SNPs were those that contribute to apparent splicing variation, the gain of a stop codon or the loss of a stop codon. Nonsynonymous SNPs were predicted to be tolerant (tSNPs) or deleterious (dSNPs) based on either PROVEAN (Choi et al., 2012) or SIFT (Choi et al., 2012).

In the *japonica* and *indica* BH samples, we identified hundreds of LoF mutations and

predicted 4,640 and 7,579 dSNPs using PROVEAN (Table 1). Interestingly, the proportion of the number of detected dSNPs to sSNPs did not increase markedly in domesticated vs. wild germplasm. For example, in the BH dataset, where wild and domesticated accessions had similar levels of coverage, the *indica* and *japonica* accession had a dSNP to sSNP ratio of 0.26 and 0.29, while the *O. rufipogon* populations had similar but slightly lower ratios of ~0.25 (Table 1). Thus, domesticated rice contains only a modestly higher proportion of *total* dSNPs compared to *O. rufipogon* germplasm.

To test whether the frequency distribution of dSNPs shifted during domestication, we defined SNPs as either ancestral or derived based on comparison to 93 *O. barthii* accessions (Table S3) and plotted the SFS for different SNP categories. For the BH data, we reduced the sample size to 70 for each population, based on sampling and coverage criteria (Materials and Methods). The resulting SFS had a U-shape for all SNP categories in cultivated rice, but not for ancestral *O. rufipogon* (Figure 1). The SFS differed significantly between wild and domesticated samples for all SNP categories (Kolmogorov-Smirnoff tests; Figure 1).

The SFS shifts between cultivated and wild germplasm were robust to: *i*) dataset, because the 3K dataset yielded similar results (Figure S1), *ii*) SNP calling approaches, because different methods were applied to the 3K and BH datasets, *iii*) the composition of the wild sample, because similar patterns were observed when the BH *japonica* and *indica* samples were compared to W_{all} ($p \leq 1.93e^{-08}$ for all comparisons in both varieties) (Figure S2) and *iv*) the prediction approach used to identify dSNPs (i.e., PROVEAN or

SIFT; Figures S1 and S3). Thus, there was a consistent signal of high-frequency derived SNPs in domesticated rice relative to its progenitor (Caicedo et al., 2007).

The primary question with regard to a cost of domestication is whether frequency shifts affect SNP types differentially. To investigate this question, we plotted the ratio of derived dSNPs vs. derived sSNPs for each frequency category of the SFS. Figure 2 shows that both *indica* and *japonica* have enhanced frequencies of derived dSNPs to sSNPs across the entire frequency range compared to W_I (Wilcoxon rank sum $p = 4.98e^{-16}$; Fig 2a) and W_{III} (Wilcoxon rank sum $p < 2.20e^{-16}$; Fig 2b), respectively. The 3K dataset exhibited similar properties (Figure S4).

In addition, we calculated $R_{(A/B)}$, a measure that compares the frequency and abundance of dSNPs vs. sSNPs in one population (A) relative to another (B) (Xue et al., 2015). When $R_{(A/B)}$ is > 1.0, it reflects an overabundance of derived dSNPs (or LoF variants) relative to sSNPs in one population over another across the entire frequency range. As expected from SFS analyses, we found that $R_{(A/B)}$ was > 1.0 for LoF variants and for dSNPs in *indica* relative to the W_I population ($p \leq 2.30e^{-139}$ for all three comparisons; Fig. 2c) and in *japonica* relative to W_{III} ($p \sim 0.000$ for the three comparisons; Fig. 2c). The 3K dataset yielded similar results (Figure S4). Hence, all of our cultivated samples illustrate increased proportions of derived dSNPs relative to sSNPs compared to wild samples, consistent with a "cost of domestication".

Deleterious variants as a function of recombination rate

Theory predicts that diversity should be lower in low recombination regions (Begun and Aquadro, 1992; Charlesworth, 1994) and also that the ratio of dSNPs to sSNPs should be higher in low recombination regions due to interference (Felsenstein, 1974b). To test these predictions, we used a genetic map to calculate recombination rate in windows across rice chromosomes, and then estimated the ratio of the number of dSNPs to sSNPs for each window. Owing to different numbers of SNPs, we used larger (3MB) windows for the BH data than the 3K data (2MB). We found that the correlation between recombination rate and dSNPs to sSNPs was negative in *indica* and *japonica* for the BH data, but not significant for either taxon (Figure 3a and Table 2). However, these same results were significant for the 3K data (Figure S5 and Table 2).

We also examined the density of sSNPs and dSNPs relative to map-based recombination rate (Figure 3bc). For both the BH and 3K datasets (Figure S5), the density of sSNPs and dSNPs were significantly positively correlated with recombination rate, indicating reduced diversity in low recombination regions (Begun and Aquadro, 1992; Charlesworth, 1994). Overall, our results indicate that lower recombination regions contain less diversity but higher proportions of dSNPs.

dSNPs in regions of putative selective sweeps

Regions linked to selective sweeps (SS) should have increased frequencies of derived mutations (Fay and Wu, 2000), including dSNPs (Hartfield and Otto, 2011). We tested two hypotheses concerning SNPs in putative SS regions. The first was that SS regions

have increased frequencies of derived SNPs relative to the remainder of the genome. The second is that SS regions can alone explain the accumulation of high frequency derived dSNPs in Asian rice.

To test our hypotheses, we made use of previously identified SS regions. Huang et al. (2012c) defined SS regions based on the relative difference in average pairwise nucleotide diversity (π) between wild and domesticated populations (Huang et al., 2012c). That is, the regions were based on π_d/π_w, where π is measured per base pair and the subscripts refer to domesticated and wild samples. We also inferred selective sweeps using two additional approaches: SweeD (Pavlidis et al., 2013) and XP-CLR (Chen et al., 2010). SweeD identifies regions of skewed selective SFS relative to background levels for a single population (i.e., the rice sample). In contrast, XP-CLR searches for genomic regions for which the change in allele frequency between two populations (cultivated vs. wild samples) occurred too quickly at a locus, relative to the size of the region, to be caused by genetic drift. Both SweeD and XP-CLR were applied with a 5% cutoff. Because XP-CLR requires explicit genotypes, we used the 3K datasets for all of the SS analyses, along with a subset of 29 *O. rufipogon* genomes that had higher (> 4x) coverage than the full *O. rufipogon* dataset (Table S1).

Focusing first on the *indica* 3K dataset, the three approaches identified different numbers, locations and sizes of selective sweeps (Table 3). For example, Huang et al. (2012b) defined 84 SS regions that encompassed 9.98% of the genome. In contrast, SweeD identified 485 SS regions, and XP-CLR distinguished an intermediate number of

161 SS regions. Consistent with the 5% cutoff, SweeD and XP-CLR identified 4.61% and 5.02% of the genome, respectively, as having been under selection (Table 3).

The locations of putative SS regions are of interest because they may correspond to genes of agronomic significance (Wright et al., 2005) and provide unique insights into domestication events (He et al., 2011). To see if the same genes were identified with different SS identification methods, we calculated the degree of overlap across methods, focusing on the percentage of genes that two methods identified in common (see Methods). The overlap was surprisingly low (Figure 4 and Figures S6-16). Across the entire genome, the putative SS regions defined by SweeD and Huang et al (2012c) shared 6.24% of genes. Similarly, the regions defined by XP-CLR shared 8.51% and 8.69% of genes with Huang et al. (2012c) and SweeD, respectively.

To test our first hypothesis, we contrasted the SFS between SS and non-SS regions for derived sSNPs and dSNPs (Marsden et al., 2016). The SFS for derived sSNPs and dSNPs were skewed for SS regions relative to non-SS regions, independent of the method used to detect selective sweeps (Figure 5). These SS regions also contained higher proportions of derived allele counts (Figure 5). We note that these results were not completely unexpected, because all of the methods used to define SS regions rely, in part, on identifying a skewed SFS relative to the genomic background (see Discussion).

These observations raise the intriguing possibility that selective sweeps alone explain the shifted SFS of *indica* rice relative to *O. rufipogon*. To examine this second hypothesis, we removed all SS regions (as defined by SweeD, XP-CLR and p_d/p_w) from the *indica* 3K

dataset and recalculated the SFS for non-SS regions. Even with SS regions removed, the SFS for wild and cultivated samples remained significantly different for sSNPs and dSNPs ($p \leq 0.0067$). These results imply either that positive selection is not the only cause of the U-shaped SFS in *indica* rice (Caicedo et al., 2007) or, alternatively that positive selection has affected more of the genome than is encompassed within the identified SS regions.

We performed these same analyses for the 3K temperate and tropical *japonica* datasets (Table 3), with similar results. First, although a greater extent of the genome tended to be identified as SS regions in *japonica* (Table 3), the overlap among SS regions identified by different methods was again low (< 9%). Second, for both *japonica* datasets, derived sSNPs and dSNPS were generally at higher frequencies and at higher counts in putative SS regions, although the effect was not as apparent for sweeps identified with SweeD (Figure S17). Third, like *indica* rice, the SS regions alone did not account for the difference in SFS between wild and either tropical or temperate *japonica* germplasm ($p \leq 0.0049$ for both comparisons).

The potential effect of mating system

Our empirical analyses indicate that domestication has increased the frequency of derived dSNPs relative to sSNPs for both *indica* and *japonica* rice (Figure 2), and that these increased frequencies are exacerbated in putative SS regions (Figure 5). However, we also find that the proportion of the total number of dSNPs to sSNPs is similar across taxa

in the BH dataset; that is, the proportion of dSNPs to sSNPs is not substantially elevated in *O. sativa* (Table 1). To gain further insight into this observations, we performed forward simulations of populations that differed in the presence or absence of three important features: a domestication bottleneck, positive selection and a shift to inbreeding at the time of the domestication bottleneck. To perform these forward simulations, we made assumptions about the bottleneck size during domestication (Caicedo et al., 2007), the dominance coefficient (h=0.5) and the distribution of fitness effects of new mutations (see Methods).

Figure 6 presents the simulation results, which can be summarized as follows. First, the effect of a bottleneck is to increase proportion of dSNPs to sSNPs relative to a non-bottlenecked population. Second, the net effect of inbreeding, relative to an outbred population, is to lower the ratio of dSNPs to sSNPs, consistent with the action of purging of deleterious variants (Arunkumar et al., 2015). Third, under our simulation conditions, positive selection has only slight effects on the ratio of the number of dSNPs to sSNPs (Figure 5), probably because positive selection affects both SNP categories. Overall, the combination of a domestication bottleneck, a shift to inbreeding and positive selection produces a ratio of dSNPs to sSNPs that is similar to that of an outbred population under our simulation conditions.

DISCUSSION

Recent focus on the population genetics of dSNPs in humans (Henn et al., 2015; Henn et

al., 2016), plants (Lu et al., 2006; Gunther and Schmid, 2010; Mezmouk and Ross-Ibarra, 2014; Nabholz et al., 2014; Renaut and Rieseberg, 2015; Kono et al., 2016) and animals (Schubert et al., 2014; Marsden et al., 2016; Robinson et al., 2016) reflect an emerging recognition that dSNPs provide unique clues into population history, the dynamics of selection and the genetic bases of phenotypes. This is especially true for the case of domesticated species, where the increased frequency of deleterious variants reflect a potential "cost of domestication" (Schubert et al., 2014).

Our analyses unequivocally demonstrate a cost of domestication in Asian rice. We have detected this cost as a skew in the frequency spectrum in domesticated relative to wild populations (Figure 1), an increased proportion of derived dSNPs relative to derived sSNPs across frequency classes (Figure 2A&B), and an overall increase in the ratio of the frequency and number of deleterious variants to synonymous SNPs between wild and cultivated germplasm (Figure 2C). For all of these measures, the results were consistent between different types of presumably deleterious variants (i.e., dSNPs vs. LoF variants), different methods to predict deleterious SNPs (PROVEAN vs SIFT; Figs. S1 and S3), and rice datasets (BH data vs. 3K data) that were based on different accessions, levels of coverage and SNP detection methods. We note, however, that our analyses of the 3K and BH data relied on the same *O. rufipogon* accessions, which may artificially emphasize similarities between the two datasets.

Our analyses are subject to caveats and assumptions. While we have tried to overcome potential pitfalls by using multiple approaches (different datasets, SNP calling

methods, dSNP predictors, and SS inference metrics), important limitations remain. One is the potential for a reference bias, because the use of the *japonica* reference is expected to decrease the probability that a *japonica* variant (as opposed to an *indica* variant) returns a low PROVEAN or SIFT score (Lohmueller et al., 2008). We have adjusted for this bias by submitting the ancestral allele -- rather than the reference allele -- to annotation programs (Kono et al., 2016). Without this adjustment, a reference bias was patently obvious, because the SFS of *japonica* dSNPs lacked a high frequency peak, and the U-shape of tSNPs became commensurately more extreme. We cannot know that we have corrected completely for reference bias. We do, however, advocate caution when interpreting results from dSNP studies that make no attempt to correct for reference bias, because the effect can be substantial.

Our treatment of reference bias requires accurate ancestral inferences. To date, many studies of Asian rice have relied on outgroup sequences from *O. meridionalis* e.g., (Caicedo et al., 2007; Gunther and Schmid, 2010), a species that diverged from *O. sativa* ~2 million years ago (Zhu and Ge, 2005). When we used *O. meridionalis* as the sole outgroup, we inferred a U-shaped SFS in wild *O. rufipogon*, which is suggestive of consistent parsimony misinference of the ancestral state (Keightley et al., 2016). We instead inferred ancestral states relative to a dataset of 93 accessions of African wild rice (*O. barthii*) (Wang et al., 2014). *O barthii* is closer phylogenetically to *O. sativa* than *O. meridonalis*, but *O. barthii* sequences form clades distinct from *O. sativa* (Zhu and Ge, 2005). Even so, we have found that ~10% of SNPs sites with minor allele frequencies >

5% are shared between African wild rice and Asian rice.

These shared polymorphic sites could mislead inference of the ancestral state, but we do not believe that the use of *O. barthii* this has distorted our primary inferences, for two reasons. First, systematic misinference of the ancestral state should lead to a U-shaped SFS, which is lacking from *O. rufipogon*. Instead, the U-shaped SFS is unique to *O. sativa* and differentiates wild from domesticated species. Second, we have confirmed our inferences by using *O. meridonalis* and *O. barthii* together as outgroups (Keightley et al., 2016), considering only the sites where the two agree on the ancestral state. The use of two outgroups decreases the number of SNPs with ancestral states by ~10% and ~15% for the BH and 3K datasets, but all analyses based on these reduced SNP sets were qualitatively identical to those with only an *O. barthii* outgroup (e.g., Figure S18).

The components of cost

Our principle finding is that the frequency of derived dSNPs has shifted from wild *O. rufipogon* to domesticated Asian rice and that shift is more pronounced for dSNPs than sSNPs. There are at least four major evolutionary factors that could drive these patterns: *i*) population size, particularly bottlenecks associated with domestication (Caicedo et al., 2007; Zhu et al., 2007), *ii*) linkage effects, especially to selective sweeps (Hartfield and Otto, 2011; Marsden et al., 2016), *iii*) relaxed selection on wild traits that are no longer important under cultivation (Renaut and Rieseberg, 2015) and, finally, *iv*) inbreeding, because the domestication of rice included a shift from an outcrossing to a selfing

breeding system.

We cannot apportion the proportion of cost attributable to each of four forces, but we can provide some insights as to whether each has an effect. First, consider population size and recall that *japonica* rice has ~2-3 fold smaller N_e than *indica* (Huang et al., 2012b). If shifts in population sizes have played a role in the accumulation of high frequency dSNPs in *japonica* and *indica*, we expect *japonica* to have a more pronounced shift in high frequency derived dSNPs, owing to lower efficacy of selection in smaller populations. This is indeed what we find (Figure 2C), consistent with an effect of N_e. This observation agrees with simulation results, because they show a pronounced effect between bottlenecked vs. non-bottlenecked populations (Figure 6).

Our work also shows that the second force - linkage - influences the accumulation of deleterious variants within the rice genome, because low recombination regions of the rice genome have higher ratios of dSNPs to sSNPs than the remainder of the genome (Figure 3). This enrichment of dSNPs in low recombination regions appears to be a general phenomenon, based on studies in *Drosophila* (Campos et al., 2014), sunflower (Renaut and Rieseberg, 2015) and soybean (Kono et al., 2016). It remains unclear whether differences between high and low recombination regions of the genome are driven by lower N_e in regions of low recombination (Hill and Robertson, 1966; Felsenstein, 1974a; Charlesworth et al., 1993) or by linkage effects to positively selected variants (Begun and Aquadro, 1992).

A more interesting aspect of linkage is the expected enrichment of dSNP frequencies

near genes that have experienced selective sweeps (SS) (Hartfield and Otto, 2011). This enrichment is readily detectable for *indica* rice (Figure 5) and largely consistent – but less pronounced – within *japonica* rice (Figure S17). We suspect that the enrichment of dSNPs in SS regions is not as pronounced in *japonica* rice because of its lower genetic diversity. Lower diversity has two confounding effects: it makes it more difficult to identify discrete SS regions, and the demographic history that leads to lower diversity (e.g., lower N_e in *japonica*) may drive more accumulation of high frequency derived dSNPs throughout the genome, independent of sweeps. Nonetheless, one must be careful about our conclusions, because there is circularity in our identification of SS regions and comparisons of SFS between SS and non-SS regions. This circularity comes from the fact that most methods that detect SS regions, including π_d/π_w, rely to some extent on a skew of the SFS. The relationships among population size, SS regions and their contributions to the 'cost of domestication' merit continued study.

As an aside, it is worth briefly focusing on the locations of SS regions identified by three different methods (Figure 5). To our surprise, the regions defined by the three methods rarely overlapped (Table 3), such that the three methods identified almost completely independent regions of the genome. The lack of convergence among methods probably reflects the fact that different tests are designed to capture different signals of selection. However, the results are also sobering, because overlaps in SS regions have been used by a number of groups to argue for or against independent domestication of *indica* and *japonica* rice (He et al., 2011; Molina et al., 2011). Recently, both Huang et al.

(2012b) and Civian et al. (2015) have argued for independent domestication events for *japonica* and *indica* based on the observation that there is little overlap between the SS regions in *japonica* and *indica*. [The Civian *et al*. (2015) analyses may also have other flaws (Huang and Han, 2015)]. The fact that we find little overlap among SS regions identified by different methods mirrors the lack of overlap of SS regions identified across the human genome by different studies (Akey, 2009), between domesticated grasses (Gaut, 2015), and between independent domestication events of common bean (Gaut, 2015). Because the inferred locations of SS regions vary markedly by method, sampling and taxon, they should be interpreted with caution, particularly as markers of independent domestication events.

The third potential contributor to the preferential accumulation of high frequency derived dSNPs is relaxed selection on traits that are under strong purifying selection in the wild but less critical under cultivation (Renaut and Rieseberg, 2015). Unfortunately, we cannot yet ascertain the degree to which this shift contributes to the accumulation of dSNPs in rice or other domesticated crops.

The fourth and final potential evolutionary force is a shift in mating system, because rice became predominantly selfing during domestication. It is generally thought that a shift to selfing offers advantages for an incipient crop, such as reproductive assurance, reduced opportunities for gene flow between an incipient crop and its wild ancestor (Dempewolf et al., 2012), and the creation of lines that "breed true" for agronomically advantageous traits (Allard, 1999). This shift may also affect the accumulation of

deleterious mutations, but the effect can be difficult to predict, because of antagonistic effects (Arunkumar et al., 2015). On one hand, inbreeding increases homozygosity, exposing recessive deleterious mutations to natural selection (Lande and Schemske, 1985) and potentially leading to the purging of deleterious alleles (Charlesworth and Willis, 2009). On the other hand, inbreeding reduces both the population size and effective recombination rates (Nordborg, 2000), thereby reducing the efficiency of selection and contributing to the retention and possible fixation of deleterious variants (Takebayashi and Morrell, 2001).

We have used forward simulations to examine the interplay between inbreeding and demographic (bottleneck) effects under parameters similar to those of *O. sativa* domestication. These simulations are unlikely to precisely mimic rice genome history, but they do offer some insight into relative effects among evolutionary forces. Our forward simulations are consistent with the possibility that a change in mating system during domestication provides a benefit, in terms of purging deleterious variants, as measured by the ratio of dSNPs to sSNPs (Figure 6). These interactions require further study, because this potential benefit may vary with bottleneck size, dominance coefficients (h), patterns of positive selection, and the timing of recovery from demographic events (Brandvain and Wright, 2016).

Similar analyses that take into account population size, inbreeding and their interplay have been performed for domesticated dogs (Marsden et al., 2016). Dog domestication includes two stages: a population bottleneck associated with domestication ~15,000 years

ago (Vonholdt et al., 2010) and inbreeding within the last few hundred years to produce modern breeds. Marsden et al (2016) have argued that the domestication bottleneck, rather than inbreeding, has had a larger effect on the accumulation of deleterious genetic variation in dogs. In rice, however, selfing likely coincides with the domestication bottleneck, such that the bottleneck and selfing have roughly similar durations and commensurately similar effects on the ratio of dSNPs to sSNPs over the limited parameters of our simulation (Figure 6). In the long term, selfing species are likely to fall victim to Muller's ratchet and go extinct (Takebayashi and Morrell, 2001). However, in the short term, a shift to selfing during domestication may provide a benefit that ameliorates some features of the "cost of domestication".

MATERIALS AND METHODS

Sequence polymorphism data

All of the data used in this study are publicly available. Illumina paired-end reads for the BH and 3K dataset were downloaded from the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) (see Tables S1 and S2 for accession numbers). The 3K accessions were chosen randomly among the total set of accessions with >12X coverage for an equal representation (*n*=15) of *indica*, tropical japonica and temperate *japonica* rice accessions. We also downloaded resequencing reads from *O. barthii* to polarize SNPs as either derived or ancestral. Sequencing reads for 93 *O. barthii* accessions (Wang et al., 2014) were obtained from the Sequence Read Archive (SRA) database in National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/sra/) (see Table S3 for accession numbers). Sequencing reads for another outgroup taxon, *O. meridonalis* were obtained from NCBI (BioProject No: PRJNA264483) (Zhang et al., 2014).

Read alignment and SNP detection

Paired-end reads for *O. sativa* and *O. rufipogon* data were assessed for quality using FastQC V0.11.2, and then preprocessed to filter adapter contamination and low quality bases using Trimmomatic V0.32 (Bolger et al., 2014). The trimmed reads were mapped to the reference genome for *japonica* rice (MSU V7), which was downloaded from the

Rice Genome Annotation Project (http://rice.plantbiology.msu.edu). Mapping was performed with the ALN and SAMPE commands implemented in the software Burrows-Wheeler Aligner (BWA) V0.7.8 (Li and Durbin, 2010), using default parameters. All reads with a mapping quality score of < 30 were discarded.

The method of SNP calling varied with the dataset. For the BH data, alignment files from BWA mapping were processed further by removing PCR duplicates and by conducting indel realignments using Picard tools and GATK, and then used as input for ANGSD V0.901, which is designed to deal with sequences of low depth (Korneliussen et al., 2014). ANGSD was run with the command line:

```
angsd -b BAMLIST -anc OUTGROUP –out OUTFILE -remove_bads -uniqueOnly 1
-minMapQ 30 -minQ 20 -only_proper_pairs 1 -trim 0 -minInd NUMBER -P
CPUNUMBERS -setMinDepth 3 -setMaxDepth 15 -GL 1 -doSaf 1 -doMaf 2
-SNP_pval 1e-3 -doMajorMinor 1 -baq 1 –C 50 –ref REFSEQ
```

We considered only SNPs that had between 3X and 15X coverage, with the high-end implemented to avoid regions with copy number variation (Huang et al., 2012b). For SNP calling, we used only uniquely mapping reads, and bases with quality score of < 20 were removed. SNP sites with >50% missing data were discarded.

For the higher coverage '3K' dataset, we used SAMtools V1.2 (Li et al., 2009) and GATK V3.1 (McKenna et al., 2010) to call SNPs. After mapping reads of each accession onto the reference genome, alignments were merged and potential PCR duplications were removed using Picard tools V1.96

(http://sourceforge.net/projects/picard/files/picard-tools/1.96/). Unmapped and non-unique reads were filtered using SAMtools V1.2. We realigned reads near indels by using the IndelRealigner and BaseRecalibrator packages in GATK to minimize the number of mismatched bases. The resulting mapping alignments were used as input for UnifiedGenotyper package in GATK and for SAMtools. SNPs that were identified by both tools, with no missing data and a minimum phred-scaled confidence threshold of 50, were retained. Subsequently, SNP calls were further refined by using the VariantRecalibrator and ApplyRecalibration packages in GATK on the basis of two sets of "known" rice SNPs (9,713,967 and 2,593,842) that were downloaded from the dbSNP and SNP-Seek databases (Alexandrov et al., 2015). These same SNP detection methods were applied to the subset of 29 *O. rufipogon* accessions the highest coverage - i.e. >4x (Table S1).

Finally, sequence reads for the outgroup dataset were aligned to the reference genome using stampy V1.0.21 (Lunter and Goodson, 2011), and then a pseudo-ancestral genome sequence was created using ANGSD (Korneliussen et al., 2014) with the parameters "-doFasta 2 -doCounts 1". This pseudo-ancestral genome was used to determine the ancestral state of each SNP in *O. sativa* and *O. rufipogon*.

SNP Annotation and Deleterious Mutation Prediction

SNPs were annotated using the latest version of ANNOVAR (Wang et al., 2010) relative to the *japonica* reference genome (MSU v 7.0). SNPs were annotated as

synonymous, nonsynonymous, intergenic, splicing, stop-gain and stop-loss related. Throughout the study, we combined SNPs that contribute to splicing variation, stop-gain and stop-loss and called them loss-of-function (LoF) mutations.

To discriminate putatively deleterious nSNPs from tolerant nSNPs, nSNPs were predicted as deleterious or tolerated using PROVEAN V1.1.5 against a search of the NCBI nr protein database (Choi et al., 2012). To reduce the effects of reference bias, predictions of deleterious variants were inferred using the ancestral (rather than the reference) variant. Following previous convention (Renaut and Rieseberg, 2015), we considered an nSNP to be a deleterious dSNP if it had a PROVEAN score ≤ -2.5 and a tolerant tSNP when a PROVEAN score was > -2.5. To assess consistency, we also employed SIFT (Kumar et al., 2009) to predict nSNPs as dSNPs or tSNPs. For these analyses, a nSNP was defined as a dSNP if it had a normalized probability < 0.05, and an nSNP was predicted to be a tSNP with a SIFT score ≥ 0.05.

Calculating site frequency spectra

Following Huang et al. (2012b), we separated the BH dataset of 1,212 accessions into five populations: *indica*, *japonica* (mostly temperate) and three *O. rufipogon* subpopulations (W_I, W_{II}, and W_{III}) The five subpopulations were composed of 436, 330, 155, 121, and 170 individuals, respectively (Table S1).

To calculate the site frequency spectrum (SFS) for subpopulations, we initially projected the sample size of all five subpopulations to that smallest W_{II} population of

n=121. However, many of the 121 accessions had low sequencing depth and high levels of missing data. We therefore focused on the W_{II} population to find criteria suitable for inclusion. Ultimately, we sought to retain ≥ 90% of SNP sites within each SNP category, which resulted in a sample size of n = 70 for the W_{II} population. Accordingly, we randomly sampled n = 70 individuals from the remaining four subpopulations, so long as the sample retained ≥ 90% of SNP sites for each category, to mimic the W_{II} sample.

Given a sample of n = 70 for each of the five subpopulations, the SFS for each subpopulation was calculated using the formula proposed by (Nielsen et al., 2005), where the *O. barthii* sequence was used as an outgroup to determine the polarity of the mutations.

$$p_{i,70} = k^{-1} \sum_{j=1}^{k} \frac{\binom{f_j}{i}\binom{n_j - f_j}{70 - i}}{\binom{n_j}{70}} \quad (1)$$

In this formula (1), $p_{i,70}$ represents the probability of the derived allele frequency (DAF) of SNPs found in i individuals in a sample size of 70; k is the total number of SNPs in the dataset; n_j and f_j are the sample size and the number of derived alleles of the jth SNP, respectively. The SFS was calculated in an identical manner with data from the 3K dataset except the sample sizes for each population were n = 15 instead of 70. The SFS for sSNPs, tSNPs, dSNPs and LoF SNPs were compared with the Kolmogorov-Smirnov test.

$R_{A/B}$ - A relative measure of dSNPs frequency enhancement

We adopted a metric to assess the accumulation of deleterious variants in either cultivated or wild rice populations (Xue et al., 2015). In this analysis, the statistic $L_{A,B}(C)$ compares two populations (A and B) within a given particular category, C, of SNP sites (e.g., dSNPs). It was calculated by counting the derived alleles found at specific sites in population A rather than B and then normalized by the same metric calculated in synonymous sites (S). The calculation of $L_{A,B}(C)$ was:

$$L_{A,B}(C) = \frac{\sum_{i \in C} f_i^A (1 - f_i^B)}{\sum_{j \in S} f_j^A (1 - f_j^B)} \quad (2)$$

where f_i^A and f_i^B are the observed derived allele frequency at each site i in populations A and B, respectively, and S refers to sSNPs. The ratio $R_{A/B}(C) = L_{A,B}(C) / L_{B,A}(C)$ then measures the relative number of derived alleles that occur more often in population A than that in population B. To obtain the standard errors of $R_{A/B}(C)$ we used the weighted-block jackknife method (HR, 1989), where each of the tested SNP datasets was divided into 50 contiguous blocks and then the $R_{A/B}(C)$ values were recomputed. A P value was assigned by using a Z score assuming a normal distribution (Do et al., 2015).

Calculation of recombination rate

The high-density rice genetic map was downloaded from http://rgp.dna.affrc.go.jp/E/publicdata/geneticmap2000/index.html, on which a total of 3,267 EST markers were anchored. We extracted the sequences of these markers from the

dbEST database in NCBI, which were used as query to perform a BLAST search against the rice genome sequences (MSU V7) to annotate their physical positions. Finally, we normalized the recombination rate to centiMorgans (cM) per 100kb between different markers, and then calculated the average recombination rate in 3 or 2Mb window segments for the BH and 3K datasets.

Identification of selective sweep regions

Both SweeD (Pavlidis et al., 2013) and XP-CLR (Chen et al., 2010) were used for identifying selective sweep (SS) regions separately in *indica* and *japonica* populations. SweeD was used with a sliding window size of 10kb, and the *O. barthii* genome sequence (Zhang et al., 2014) was used as an outgroup to determine whether alleles were ancestral or derived. XP-CLR was applied to the 3K datasets along with a subset of 29 *O. rufipogon* individuals for which we could infer explicit genotypes. Both packages were applied with 5% cutoffs to define putative sweep regions.

We calculated the percentage of genes overlapping between two sets of SS regions, defined as:

```
Overlap%= number of genes in common/ ((number of genes in the first set of SS regions + number of genes in the second set of SS regions)-number of genes in common))*100
```

Forward simulations

To examine the relative effects of demography, inbreeding and positive selection on the deleterious variants, we conducted forward simulations using the software SLiM V1.8 (Messer, 2013). SLiM includes both selection and linkage in a Wright–Fisher model with nonoverlapping generations. Similar to previous demographic studies of Asian rice domestication (Caicedo et al., 2007), we simulated a population of $N = 10000$ individuals, which were run for $10N$ generations to reach equilibrium. We then introduced a domestication bottleneck of size $N_b/N = 0.01$ at generation $11N$ until generation $15N$, when the population size recovered to size N. For the selfing population, the population switched from outcrossing to total inbreeding (inbreeding coefficient $F = 1$) at the beginning of the domestication bottleneck. We adjusted population size after the outcrossing-selfing transition by calculating the reduction in silent genetic diversity ($\theta_w = 4N_e\mu$, where θ_w is genetic diversity, N_e is effective population size and μ is mutation rate). To simplify the simulation, we assumed a constant mutation rate ($\mu = 6.5 \times 10^{-9}$ substitutions per site per generation) (Gaut et al., 1996)) and recombination rate ($\rho = 4 \times 10^{-8}$ recombinants per generation) (Gaut et al., 2007) across the single chromosome of 100 Mb with alternating 400 bp of noncoding and 200 bp of coding DNA. All noncoding and 75% of coding sequences were selectively neutral ($s = 0$). The remaining 25% of coding sequences were under negative selection with s following a gamma distribution with shape parameter 0.3 and mean -0.05. For the simulations with positive selection, we introduced 20 predetermined mutations with s drawn from an exponential distribution of mean 0.05 at the beginning of domestication. Positive selection was applied throughout

the entirety of the population simulation, not only during domestication. For all mutations under positive or negative selection, we assumed a dominance coefficient $h = 0.5$ (i.e., an additive model). In total, we simulated under six models: 1) outcrossing population with constant population size (out); 2) selfing population with comparable population size (inb); 3) bottleneck without the outcrossing-inbreeding transition (bot+out); 4) bottleneck with the outcrossing-inbreeding transition (bot+inb); 5) bottleneck with positive selection (bot+out+pos) and 6) bottleneck with both positive selection and the outcrossing-inbreeding transition (bot+inb+pos). The results for each model were summarized over 20 separate runs of SLiM; the SLiM input is available as Supplementary Text.

ACKNOWLEDGEMENTS

We thank J. Aguirre, and T. Kono for reading earlier versions of the ms, and J. Aguirre for assistance. J. Ross-Ibarra also provided comments and a revised version of XP-CLR. Q. Liu is supported by the National Natural Science Foundation of China (grant no. 31471431) and the Training Program for Outstanding Young Talents of Zhejiang A&F University. YZ is supported by the International Postdoctoral Exchange Fellowship Program 2015 awarded by the Office of China Postdoctoral Council. PLM is supported by NSF Plant Genome Program (DBI-1339393). BSG is supported by NSF IOS-1542703.

REFERENCES

Akey, J.M. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* **19**: 711-722.

Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R.R., Ulat, V.J., Chebotarov, D., Zhang, G., Li, Z. *et al.* 2015. SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res* **43**: D1023-7.

Allard, R.W. 1999. History of plant population genetics. *Annu Rev Genet* **33**: 1-27.

Arunkumar, R., Ness, R.W., Wright, S.I. and Barrett, S.C. 2015. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics* **199**: 817-829.

Begun, D.J. and Aquadro, C.F. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. *Nature* **356**: 519-520.

Bolger, A.M., Lohse, M. and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Brandvain, Y. and Wright, S.I. 2016. The Limits of Natural Selection in a Nonequilibrium World. *Trends Genet* **32**: 201-210.

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

Campos, J.L., Halligan, D.L., Haddrill, P.R. and Charlesworth, B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in Drosophila melanogaster. *Mol Biol Evol* **31**: 1010-1028.

Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., Grenier, J.C., Gbeha, E., Hamdan, F.F., Girard, S. *et al.* 2013. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* **9**: e1003815.

Charlesworth, B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63**: 213-227.

Charlesworth, B., Morgan, M.T. and Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289-303.

Charlesworth, D. and Willis, J.H. 2009. The genetics of inbreeding depression. *Nat Rev Genet* **10**: 783-796.

Chen, H., Patterson, N. and Reich, D. 2010. Population differentiation as a test for selective sweeps. *Genome Res* **20**: 393-402.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. and Chan, A.P. 2012. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**: e46688.

Civian, P., Craig, H., Cox, C.J. and Brown, T.A. 2015. Three geographically separate domestications of Asian rice. *Nature Plants* **1**: 15164.

Dempewolf, H., Hodgins, K.A., Rummell, S.E., Ellstrand, N.C. and Rieseberg, L.H. 2012. Reproductive isolation during domestication. *Plant Cell* **24**: 2710-2717.

Do, R., Balick, D., Li, H., Adzhubei, I., Sunyaev, S. and Reich, D. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet*

Eyre-Walker, A. 2010. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc Natl Acad Sci U S A* **107 Suppl 1**: 1752-1756.

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

Felsenstein, J. 1974a. The evolutionary advantage of recombination. *Genetics* **78**: 737-756.

Felsenstein, J. and Yokoyama, S. 1976. The evolutionary advantage of recombination. II. Individual selection for recombination. *Genetics* **83**: 845-859.

Felsenstein, J. 1974b. The evolutionary advantage of recombination. *Genetics* **78**: 737-756.

Gao, L.Z. and Innan, H. 2008. Nonindependent domestication of the two rice subspecies, Oryza sativa ssp. indica and ssp. japonica, demonstrated by multilocus microsatellites. *Genetics* **179**: 965-976.

Gaut, B.S. 2015. Evolution Is an Experiment: Assessing Parallelism in Crop Domestication and Experimental Evolution: (Nei Lecture, SMBE 2014, Puerto Rico). *Mol Biol Evol* **32**: 1661-1671.

Gaut, B.S., Morton, B.R., McCaig, B.C. and Clegg, M.T. 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, B.S., Wright, S.I., Rizzon, C., Dvorak, J. and Anderson, L.K. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat Rev Genet* **8**: 77-84.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H. *et al.* 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**: 92-100.

Gunther, T. and Schmid, K.J. 2010. Deleterious amino acid polymorphisms in Arabidopsis thaliana and rice. *Theor Appl Genet* **121**: 157-168.

Hartfield, M. and Otto, S.P. 2011. Recombination and hitchhiking of deleterious alleles. *Evolution* **65**: 2421-2434.

He, Z., Zhai, W., Wen, H., Tang, T., Wang, Y., Lu, X., Greenberg, A.J., Hudson, R.R. and Wu, C.I. 2011. Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genet* **7**: e1002100.

Henn, B.M., Botigue, L.R., Bustamante, C.D., Clark, A.G. and Gravel, S. 2015. Estimating the mutation load in human genomes. *Nat Rev Genet* **16**: 333-343.

Henn, B.M., Botigue, L.R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B.K., Martin, A.R., Musharoff, S., Cann, H., Snyder, M.P. *et al.* 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci U S A*

113: E440-9.

Hill, W.G. and Robertson, A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8**: 269-94.

Huang, P., Molina, J., Flowers, J.M., Rubinstein, S., Jackson, S.A., Purugganan, M.D. and Schaal, B.A. 2012a. Phylogeography of Asian wild rice, Oryza rufipogon: a genome-wide view. *Mol Ecol* **21**: 4593-4604.

Huang, X. and Han, B. 2015. Rice domestication occurred through single origin and multiple introgressions. *Nature Plants* **DOI: 10.1038/NPLANTS.2015.207**:

Huang, X., Kurata, N., Wei, X., Wang, Z.X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W. *et al.* 2012b. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**: 497-501.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C. *et al.* 2012c. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet* **44**: 32-39.

Keightley, P.D., Campos, J.L., Booker, T.R. and Charlesworth, B. 2016. Inferring the Frequency Spectrum of Derived Variants to Quantify Adaptive Molecular Evolution in Protein-Coding Genes of Drosophila melanogaster. *Genetics* PMID 27098912.

Kono, T.J., Fu, F., Mohammadi, M., Hoffman, P.J., Liu, C., Stupar, R.M., Smith, K.P., Tiffin, P., Fay, J.C. and Morrell, P.L. 2016. The role of deleterious substitions in crop genomes. *bioRxiv* **http://dx.doi.org/10.1101/033175**:

Korneliussen, T.S., Albrechtsen, A. and Nielsen, R. 2014. ANGSD: Analysis of Next

Generation Sequencing Data. *BMC Bioinformatics* **15**: 356.

Kryukov, G.V., Pennacchio, L.A. and Sunyaev, S.R. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* **80**: 727-739.

Kumar, P., Henikoff, S. and Ng, P.C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**: 1073-1081.

HR, K. 1989. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* 1217-1241.

Lande, R. and Schemske, D.W. 1985. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* **39**: 24-40.

Li, H. and Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li, J.Y., Wang, J. and Zeigler, R.S. 2014. The 3,000 rice genomes project: new opportunities and challenges for future rice research. *Gigascience* **3**: 8.

Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R. *et al.* 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**:

994-997.

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

Lu, J., Tang, T., Tang, H., Huang, J., Shi, S. and Wu, C.I. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet* **22**: 126-131.

Lunter, G. and Goodson, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* **21**: 936-939.

Marsden, C.D., Ortega-Del Vecchyo, D., O'Brien, D.P., Taylor, J.F., Ramirez, O., Vila, C., Marques-Bonet, T., Schnabel, R.D., Wayne, R.K. and Lohmueller, K.E. 2016. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc Natl Acad Sci U S A* **113**: 152-157.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Messer, P.W. 2013. SLiM: simulating evolution with selection and linkage. *Genetics* **194**: 1037-1039.

Mezmouk, S. and Ross-Ibarra, J. 2014. The pattern and distribution of deleterious

mutations in maize. *G3 (Bethesda)* **4**: 163-171.

Molina, J., Sikora, M., Garud, N., Flowers, J.M., Rubinstein, S., Reynolds, A., Huang, P., Jackson, S., Schaal, B.A., Bustamante, C.D. *et al.* 2011. Molecular evidence for a single evolutionary origin of domesticated rice. *Proc Natl Acad Sci U S A* **108**: 8351-8356.

Morrell, P.L., Buckler, E.S. and Ross-Ibarra, J. 2011. Crop genomics: advances and applications. *Nat Rev Genet* **13**: 85-96.

Nabholz, B., Sarah, G., Sabot, F., Ruiz, M., Adam, H., Nidelet, S., Ghesquiere, A., Santoni, S., David, J. and Glemin, S. 2014. Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima). *Mol Ecol* **23**: 2210-2227.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J. *et al.* 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**: 923-9.

Oka, H.I. 1988. Origin of cultivated rice. 254.

Oka, H.I. and Miroshima, H. 1967. Variations in he breeding systems of a wild rice, Oryza perennis. *Evolution* **21**: 249-258.

Pavlidis, P., Zivkovic, D., Stamatakis, A. and Alachiotis, N. 2013. SweeD:

likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol* **30**: 2224-2234.

Peischl, S., Dupanloup, I., Kirkpatrick, M. and Excoffier, L. 2013. On the accumulation of deleterious mutations during range expansions. *Mol Ecol* **22**: 5972-5982.

Project, I.R.G.S. 2005. The map-based sequence of the rice genome. *Nature* **436**: 793-800.

Renaut, S. and Rieseberg, L.H. 2015. The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops. *Mol Biol Evol* **32**:2273-83

Robinson, J.A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B.Y., vonHoldt, B.M., Marsden, C.D., Lohmueller, K.E. and Wayne, R.K. 2016. Genomic Flatlining in the Endangered Island Fox. *Curr Biol* **26**: 1183-1189.

Sang, T. and Ge, S. 2007. The puzzle of rice domestication. *J Int. Plant Bio* **49**: 760-768.

Schubert, M., Jonsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., Albrechtsen, A., Dupanloup, I., Foucal, A., Petersen, B. *et al.* 2014. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc Natl Acad Sci U S A* **111**: E5661-9.

Takebayashi, N. and Morrell, P.L. 2001. Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. *Am J Bot* **88**: 1143-1150.

Vonholdt, B.M., Pollinger, J.P., Lohmueller, K.E., Han, E., Parker, H.G., Quignon, P.,

Degenhardt, J.D., Boyko, A.R., Earl, D.A., Auton, A. *et al.* 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**: 898-902.

Wang, K., Li, M. and Hakonarson, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164.

Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., Zuccolo, A., Song, X., Kudrna, D., Ammiraju, J.S. *et al.* 2014. The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. *Nat Genet* **46**: 982-988.

Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D. and Gaut, B.S. 2005. The effects of artificial selection on the maize genome. *Science* **308**: 1310-1314.

Xue, Y., Prado-Martinez, J., Sudmant, P.H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D.N. *et al.* 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**: 242-245.

Zhang, L.B., Zhu, Q., Wu, Z.Q., Ross-Ibarra, J., Gaut, B.S., Ge, S. and Sang, T. 2009. Selection on grain shattering genes and rates of rice domestication. *New Phytol* **184**: 708-720.

Zhang, Q.J., Zhu, T., Xia, E.H., Shi, C., Liu, Y.L., Zhang, Y., Liu, Y., Jiang, W.K., Zhao, Y.J., Mao, S.Y. *et al.* 2014. Rapid diversification of five Oryza AA genomes

associated with rice adaptation. *Proc Natl Acad Sci U S A* **111**: E4954-62.

Zhu, Q., Zheng, X., Luo, J., Gaut, B.S. and Ge, S. 2007. Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. *Mol Biol Evol* **24**: 875-888.

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

Table 1: The number and category of SNPs identified from different datasets

		SNP Type[1]					
Sample	*n*	ncSNP	sSNP	dSNP	tSNP	LoF	Total[2]
BH data							
indica	436	323,182	28,765	7,579	22,577	1,279	388,615
japonica	330	206,915	15,767	4,640	13,230	797	243,964
W_I	155	1,608,310	91,985	23,195	66,641	4,976	1,816,094
W_{II}	121	3,548,048	150,522	37,483	108,480	8,415	3,888,315
W_{III}	170	3,054,763	147,824	35,795	105,455	7,641	3,381,359
3K data							
indica	15	3,262,864	118,582	22,293	87,225	5,788	3,534,837
japonica (temperate)	15	2,634,868	97,922	18,894	72,823	4,751	2,861,999
japonica (tropical)	15	2,808,403	103,238	19,977	77,414	5,012	3,048,559

[1] ncSNP=non-coding; sSNP=synonymous; dSNP=deleterious; tSNP=tolerated; LoF = Loss of Function. dSNPs were predicted with PROVEAN.

[2] The total number of SNPs identified in the dataset is not equal to the sum of all SNP categories, because some nonsynonymous SNPs were filtered for quality after PROVEAN analysis and therefore were not categorized.

Table 2: Correlation coefficients comparing rice diversity statistics to recombination rate (cM/100kb)

	sSNPs[1]		dSNPs		dSNPs/sSNPs	
Data Set	r^2	*p*-value[2]	*r*	*p*-value	*r*	*p*-value
BH *indica*	0.297	$6.68e^{-04}$	0.319	$2.46e^{-04}$	-0.166	0.061
BH *japonica*	0.209	$1.81e^{-02}$	0.263	$2.68e^{-03}$	-0.094	0.291
3K *indica*	0.358	$3.77e^{-07}$	0.338	$1.72e^{-06}$	-0.306	$1.70e^{-05}$
3K *japonica* (temperate)	0.339	$1.62e^{-06}$	0.306	$1.72e^{-05}$	-0.279	$9.47e^{-05}$
3K *japonica* (tropical)	0.375	$9.11e^{-08}$	0.324	$4.87e^{-06}$	-0.268	$1.79e^{-04}$

[1] sSNPs, dSNPs and their ratio were calculated in non-overlapping 3Mb windows for the BH dataset and 2Mb windows for the 3K dataset.

[2] *r* is the Pearson correlation coefficient, with corresponding *p*-value.

Table 3: The number and percentage of SS regions identified by different methods, based on the 3K data.

	indica		***japonica* (temperate)**		***japonica* (tropical)**	
	No.	**Extent[2]**	**No.**	**Extent[2]**	**No.**	**Extent[2]**
Huang et al (2012b)	84[1]	9.98%	103[3]	15.32%	103[3]	15.32%
SweeD	485	4.61%	461	4.76%	389	4.81%
XP-CLR	161	5.02%	160	8.41%	171	5.62%

[1] Based on 60 SS regions identified as specific to *indica,* which overlapped with 31 of 55 regions identified in the combined samples of *indica* and *japonica* rice, for a total of [60+(55-31)]=84.

[2] Extent = the percentage of the reference genome covered by SS regions.

[3] Based on 62 SS regions identified as specific to *japonica,* which overlapped with 14 of 55 regions identified in the combined samples of *indica* and *japonica* rice, for a total of [62+(55-14)]=103.

FIGURE LEGENDS:

Figure 1: The site frequency spectrum (SFS) for cultivated rice and *O. rufipogon*, based on BH data. Each row represents a different category of SNP, featuring from top to bottom: sSNPs, tSNPS, dSNPs, and LoF variants. The two columns represent *indica* rice on the left and *japonica* rice on the right. As per Huang et al (2012b), *indica* rice is contrasted to the accessions from wild population I (W_I) and *japonica* rice is contrasted to wild sample population III (W_{III}). The Density on the y-axis is the proportion of alleles in a given bin. Each graph reports the *p*-value of the contrast in SFS between cultivated and wild samples; a similar graph for the 3K data is presented in Figure S1.

Figure 2: Comparisons of the number of derived dSNP to sSNP between wild and cultivated samples based on their frequencies BH data. A) The ratio of the number of dNSPs to sSNPs (y-axis) at each derived allele frequency (x-axis) between *indica* rice and the W_I sample. B) The ratio of the number of dNSPs to sSNPs (y-axis) at each derived allele frequency (x-axis) between between *japonica* rice and the W_{III} sample. C) A measure of the relative accumulation of SNPs in *indica* or *japonica* rice compared to *O. rufipogon*. A value > 1.0 indicates an increased population density of that SNP type relative to wild rice. Bars indicate standard errors. A similar figure for the 3K data is presented in Figure S4.

Figure 3: Patterns of genomic variation relative to recombination, based on the BH dataset. The x-axis for each graph is the recombination rate (x-axis) as measured by centiMorgans (cM) per 100 kb. The y-axis varies by row. The top row is the ratio of deleterious to synonymous variants (y-axis) in 3Mb windows; the middle row is the density of dSNPs in 3Mb windows; and the bottom row is the density of sSNPs in 3Mb windows. The *p*-values for correlations are provided in the plot, but also within Table 2. A similar figure for the 3K data is provided in Figure S5.

Figure 4: A graph of the location of inferred SS regions along Chromosome 1 for the 3K *indica* dataset. The x-axis is the location along the chromosome, measured in base pairs. The top graph (red) indicates the ratio of π for the *indica* accessions against a set of wild accessions. The (grey) background represents values for windows of 10kb with a step size of 1kb. Values > 2.0 were omitted for ease of presentation, and the line was smoothed. The middle graph shows values of π for the *indica* accessions. The bars at the bottom represent inferred SS regions using SweeD and XP-CLR, along with predefined SS regions (BH) defined by Huang et al. (2012b). The red and blue colors are included to help differentiate SS regions; the orange bars represent additional SS regions defined by Huang et al. (2012b) on the basis of their combined *indica*+*japonica* dataset. The width of each bar is proportional to the length of the corresponding SS region along chromosome. Similar graphs for chromosomes 2 through 12 are available as supplemental figures (Figures S6 to S16).

Figure 5: A comparison between selective sweep (SS) and non-SS regions based on the *indica* 3K dataset. The rows correspond to different methods employed to detect sweeps, including pre-detected SS regions from Huang et al (2012b) (top row), SweeD (middle row), and XP-CLR (bottom row). The set of histograms on the right compare the density of either sSNPs or dSNPs in inferred SS regions against the remainder of the genome (non-SS regions).

Figure 6: The results of forward simulations. The y-axis denotes the ratio of the number of dSNPs and sSNPs. The x-axis defines the six models; 'out' represents an outbred (random mating) population with a constant population size; 'inb' represents an inbred (selfing) population with a comparable population size; 'bot+out' and 'bot+inb' represents outbred and inbred populations with a domestication bottleneck; 'bot+out+pos' and 'bot+inb+pos' include positively selected alleles.

FIGURE 1:

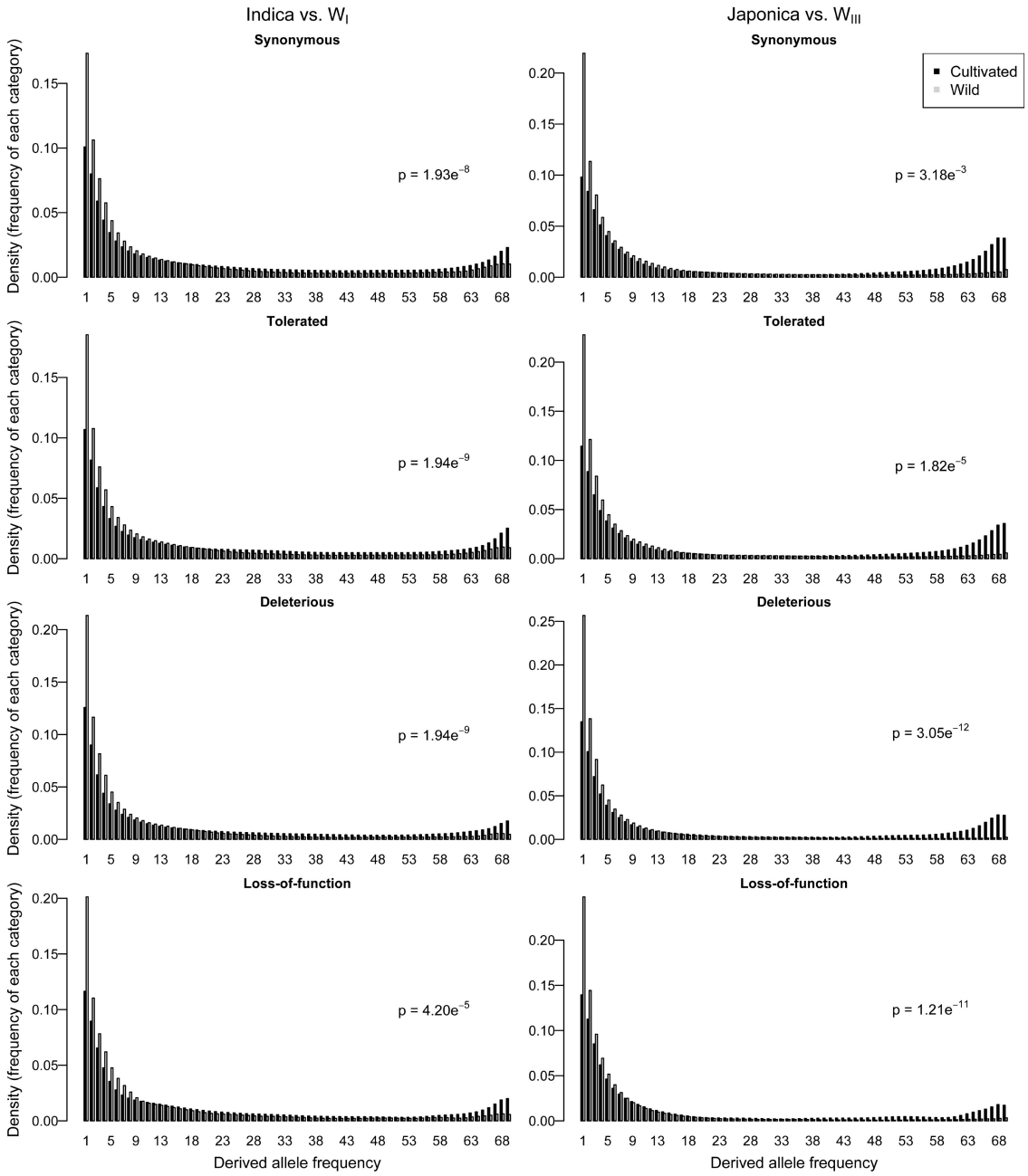

Indica vs. W_I
Japonica vs. W_III
Synonymous
Tolerated
Deleterious
Loss-of-function
Cultivated
Wild
p = 1.93e^-8
p = 3.18e^-3
p = 1.94e^-9
p = 1.82e^-5
p = 1.94e^-9
p = 3.05e^-12
p = 4.20e^-5
p = 1.21e^-11
Density (frequency of each category)
Derived allele frequency

FIGURE 2:

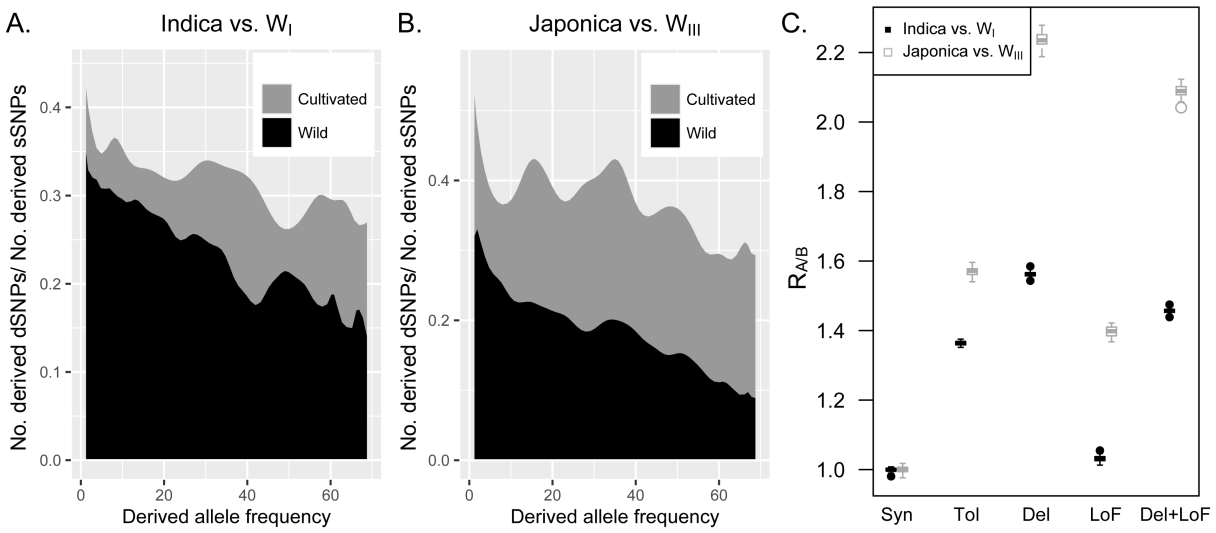

A.
Indica vs. W_I
Cultivated
Wild
No. derived dSNPs/ No. derived sSNPs
Derived allele frequency
B.
Japonica vs. W_{III}
Cultivated
Wild
No. derived dSNPs/ No. derived sSNPs
Derived allele frequency
C.
Indica vs. W_I
Japonica vs. W_{III}
$R_{A/B}$
Syn
ToI
Del
LoF
Del+LoF

FIGURE 3:

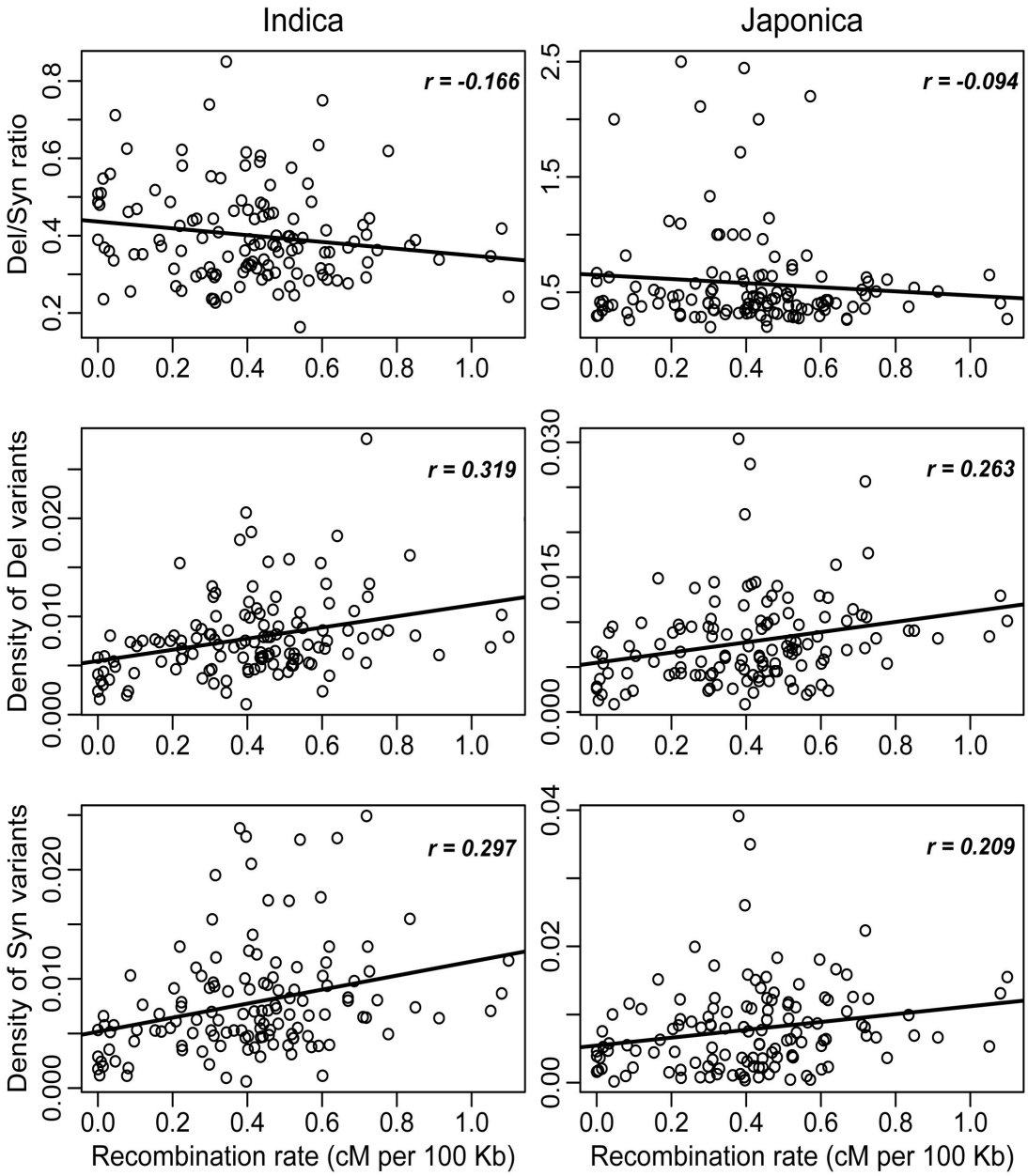

Indica
Japonica
Del/Syn ratio
Density of Del variants
Density of Syn variants
r = -0.166
r = -0.094
r = 0.319
r = 0.263
r = 0.297
r = 0.209
Recombination rate (cM per 100 Kb)
Recombination rate (cM per 100 Kb)

FIGURE 4:

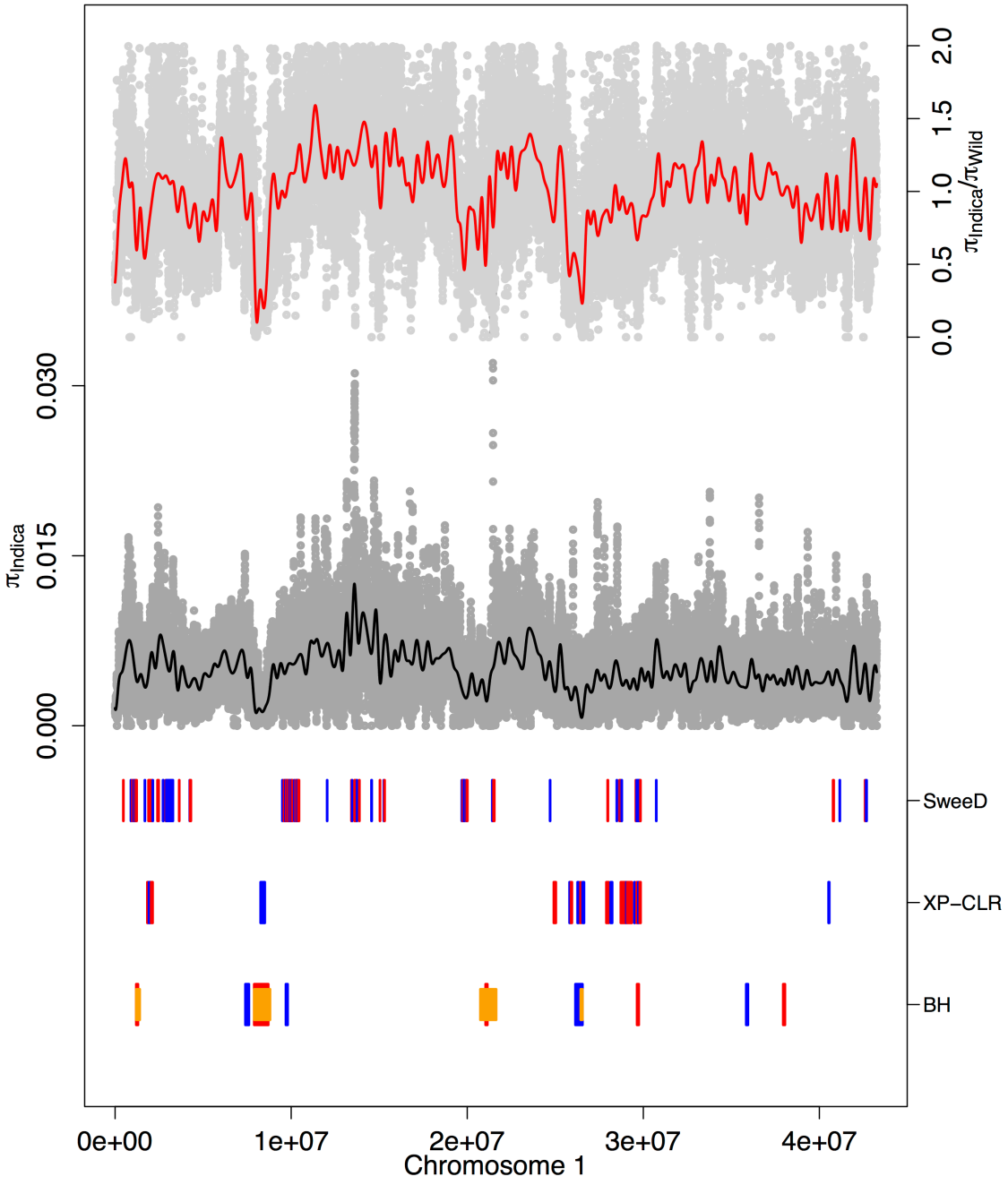

0.0
0.5
1.0
1.5
2.0
π_{Indica}/π_{Wild}
0.000
0.015
0.030
π_{Indica}
SweeD
XP−CLR
BH
0e+00
1e+07
2e+07
3e+07
4e+07
Chromosome 1

FIGURE 5:

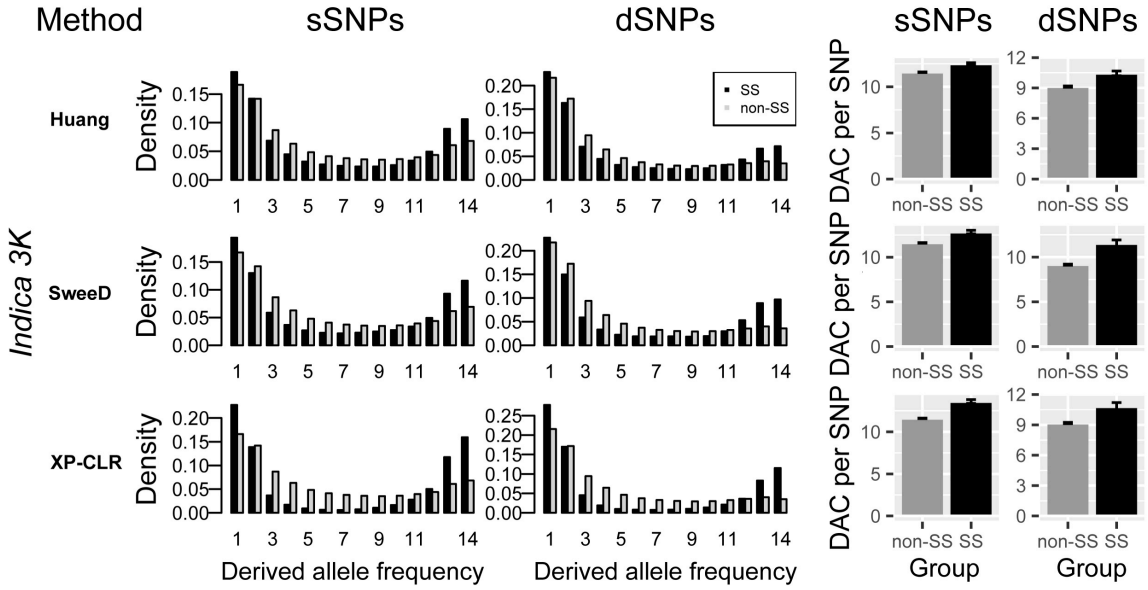
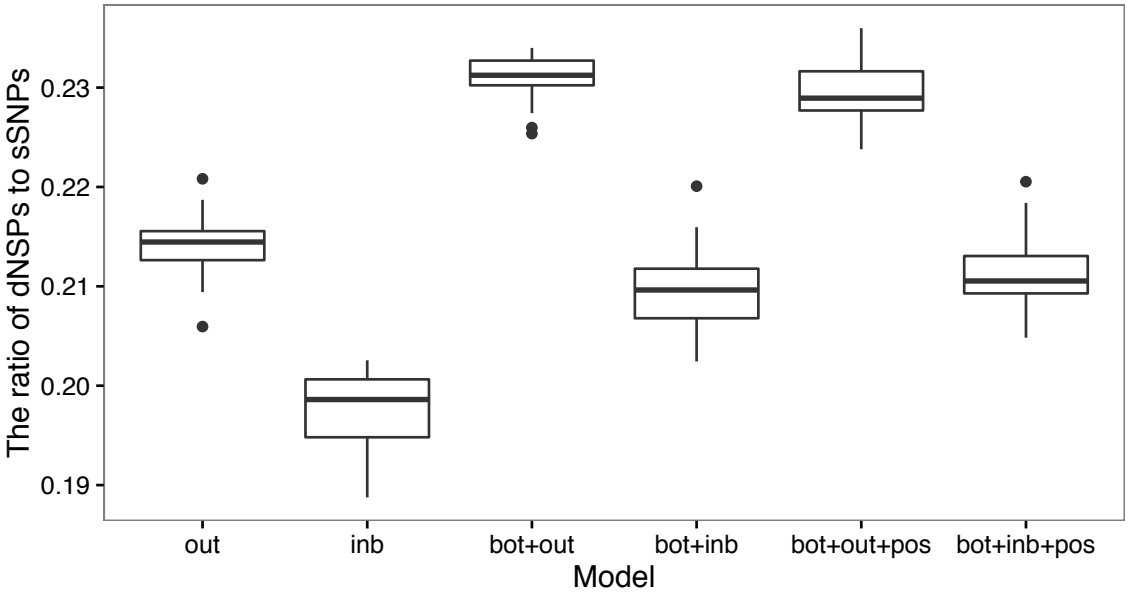

Method
sSNPs
dSNPs
sSNPs
dSNPs
Indica 3K
Huang
SweeD
XP-CLR
Density
SS
non-SS
Derived allele frequency
DAC per SNP
non-SS SS
Group

FIGURE 6:


The ratio of dNSPs to sSNPs
0.23
0.22
0.21
0.20
0.19
out
inb
bot+out
bot+inb
bot+out+pos
bot+inb+pos
Model