Evolutionary Genomics of Structural Variation in Asian Rice (Oryza sativa) Domestication

Yixuan Kou,†,1,2 Yi Liao,†,1 Tuomas Toivainen,1,3 Yuanda Lv,1 Xinmin Tian,4 J.J. Emerson,1 Brandon S. Gaut *,1 and Yongfeng Zhou*,1,5

1Department of Ecology and Evolutionary Biology, UC Irvine, Irvine, CA
2Laboratory of Subtropical Biodiversity, Jiangxi Agricultural University, Nanchang, China
3Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland
4Department of Biological Sciences, College of Life Science and Technology, Xinjiang University, Urumqi, China
5Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

†These authors contributed equally to this work.

*Corresponding authors: E-mails: yongfez1@uci.edu; bgaut@uci.edu.

Abstract

Structural variants (SVs) are a largely unstudied feature of plant genome evolution, despite the fact that SVs contribute substantially to phenotypes. In this study, we discovered SVs across a population sample of 347 high-coverage, resequenced genomes of Asian rice (Oryza sativa) and its wild ancestor (O. rufipogon). In addition to this short-read data set, we also inferred SVs from whole-genome assemblies and long-read data. Comparisons among data sets revealed different features of genome variability. For example, genome alignment identified a large (~4.3 Mb) inversion in indica rice varieties relative to japonica varieties, and long-read analyses suggest that ~9% of genes from the outgroup (O. longistaminata) are hemizygous. We focused, however, on the resequencing sample to investigate the population genomics of SVs. Clustering analyses with SVs recapitulated the rice cultivar groups that were also inferred from SNPs. However, the site-frequency spectrum of each SV type—which included inversions, duplications, deletions, translocations, and mobile element insertions—was skewed toward lower frequency variants than synonymous SNPs, suggesting that SVs may be predominantly deleterious. Among transposable elements, SINE and mariner insertions were found at especially low frequency. We also used SVs to study domestication by contrasting between rice and O. rufipogon. Cultivated genomes contained ~25% more derived SVs and mobile element insertions than O. rufipogon, indicating that SVs contribute to the cost of domestication in rice. Peaks of SV divergence were enriched for known domestication genes, but we also detected hundreds of genes gained and lost during domestication, some of which were enriched for traits of agronomic interest.

Key words: structural variation, rice, domestication, transposable element insertions, gene gain and loss.

Introduction

Structural variants (SVs) are commonly defined as differences between individuals in genome order or DNA content that span >50 bases in length (Alkan et al. 2011; Tattini et al. 2015; Gaut et al. 2018). They remain a relatively mysterious feature of plant genomes, for at least three reasons. The first is their contribution to phenotypes. Numerous examples indicate that they can affect phenotypes (Zmienko et al. 2014; Gaut et al. 2018), such as the transposable element (TE) that inserted near a grapevine myb gene and caused a shift in berry color from red to green (Kobayashi et al. 2004). More generally, however, it is not clear how often and how many SVs contribute to phenotypic traits. The second is their prevalence. The fact that plant genomes vary substantially in size among individuals (Diez et al. 2013; Gordon et al. 2017; Roessler et al. 2019) suggests that SVs are common, and we now know that they contribute to remarkable fluidity in genic content. For example, ~15% of genes within grapevine cultivars are hemizygous due to SVs that cause allelic loss (Zhou et al. 2019) and 10% of genes are nonsyntenic between maize inbred lines (Sun et al. 2018). Yet, for most plant species, we have little idea of the full extent of SVs, how they vary across different SV types—such as insertions, deletions, duplications, inversions, and translocations—or the rate at which they
Asian rice (Oryza sativa) and its wild relative O. rufipogon, with our primary goal to compare SVs in wild accessions (Zhao et al. 2018). Although SVs have been studied in Asian rice, these previous studies either did not investigate SVs in wild germplasm (Carpentier et al. 2019; Fuentes et al. 2019) or were based on small (n = 13) samples of wild accessions (Zhao et al. 2018). As a result, comparisons of SV diversity between wild and cultivated rice have been limited, and SVs have not yet been used to gain insights into rice domestication. Thus far, SV population frequencies have been compared only between wild and cultivated grapevines (Zhou et al. 2019). In this case, SVs provided unique insights into regions of the genome that have been under artificial selection, and they also reflected an increased genetic burden associated with clonal propagation. It remains an open question, however, whether other domesticated plants have increased SV burdens relative to their wild relatives, whether SVs will commonly provide useful insights into genomic regions subjected to positive selection and whether domestication commonly includes shifts in genic content. The identification of genes gained and lost during domestication is an area of growing emphasis, because such genes may contribute to agronomic phenotypes (Gaut et al. 2018).

Here, we take advantage of the remarkable array of publicly available rice data, including high-coverage resequencing data, long-read data, and genome assemblies to study SVs in cultivated and wild rice accessions. After inferring SVs from these different data sets, we assess the reliability of SV calls across data sets, including a comparison to a previous SV study (Fuentes et al. 2019). We then focus on a highly curated set of biallelic SVs: (1) to compare SV population frequencies between wild and cultivated taxa, (2) to investigate MEI
frequencies across different TE families, and (3) to evaluate SVs with respect to common features associated with domestication, such as its genetic cost and signatures of positive selection.

**Results**

**Comparing SVs among Short-Read, Long-Read, and Whole-Genome Data**

We gathered three data sets to identify, contrast, and confirm SV calls: (1) 14 whole-genome assemblies, (2) Pacific Biosciences (PacBio) SMRT long reads from a subset of ten accessions included in whole-genome assembly, and (3) paired-end resequencing data from 347 wild and cultivated accessions that met our filtering criteria (see Materials and Methods), especially high enough mapped coverage (>15×) to aid accurate SV detection.

Focusing first on the resequencing data, we identified 38,717,560 SNPs and small variants (i.e., indels <50 bp). As expected (Garris et al. 2005; Huang et al. 2012; Wang et al. 2018; Zhao et al. 2018), phylogenetic analyses based on SNPs revealed clear group-specific clades, with the aus-indica clade separated from the japonica-aromatic clade (fig. 1A). The exception to clear separation was *O. nivara*, which nested within rufipogon clades, but this result was consistent with previous analyses based on wide sampling of *O. nivara* and rufipogon (Cai et al. 2019). In the phylogeny, rufipogon was represented by two major clades: one that branched next to the outgroups (*O. meridionalis* and *O. longistaminata*) and derived primarily from India and Southeast Asia, and a second that rooted between indica and japonica and included accessions primarily from China. We include this information to emphasize population structure among wild rice accessions, which prompted us to include only one clade for domestication analyses (see below).
majority of SV inferences within cultivated rice agreed with previous work, suggesting that our combination of high-coverage data with two SV-callers performed well.

Third, we compared SVs based on short-read data with SVs inferred from two other types of sequence data: whole-genome assemblies and SMRT reads (supplementary table S2, Supplementary Material online). SVs were inferred from genome assemblies based on pairwise genome alignments using MUMmer (Marçais et al. 2018) and LASTZ (Harris 2007) (see Materials and Methods). MUMmer detects DEL, DUP, INV, TRA, and INS events, and we further delineated INS events into MEI and non-MEI events. Across the pairwise comparison of 14 genomes, including an O. longistaminata (Reusch et al. 2018) outgroup, we detected a total of 390,823 SVs (table 1 and supplementary table S1, Supplementary Material online). Notably, the SVs included a ~4.3-Mb (coordinates: 11,296,942–15,576,712 bp in the O. longistaminata reference) homozygous inversion that spanned the centromere of chromosome 6 and differed between the four indica assemblies and assemblies from other Oryza species (fig. 1C and supplementary figs. S1 and S2, Supplementary Material online).

In addition to whole-genome assemblies, we collected raw SMRT reads from the subset of six indica, three japonica, and one O. longistaminata accession that were used in genome alignment (table 1 and supplementary table S2, Supplementary Material online) and then mapped SMRT reads to the Nipponbare genome (Zhang et al. 2018) using Minimap2 (Li 2018). SVs were called across all ten samples using the Sniffles pipeline (Sedlacek et al. 2018), identifying 531,926 SVs (table 1 and supplementary table S1, Supplementary Material online), again detecting DEL, DUP, INV, TRA, and INS events and further separating INS events into MEI and non-MEI events. SMRT reads also provide the unique opportunity to investigate within-genome hemizygosity, based on the presence of alternative reads that do and do not span a genomic region (Vondras et al. 2019; Zhou et al. 2019). Applying this approach to a japonica individual (Nipponbare), an indica individual (93-11), and a wild outcrossing relative O. longistaminata, we detected similar numbers of presence–absence variants (PAVs) in Nipponbare (257 DELs and 31 DUPs) and 93-11 (306 DELs and 18 DUPs). For these two accessions, only 0.73% (385) and 0.35% (171) of genes were hemizygous, whereas 56-fold more PAVs were found in the outcrossing O. longistaminata accession, resulting in estimated hemizygosity for 8.89% (3,895) genes.

Finally, we compared the three data sets for three SV categories that were most comparable across methods: DEL, DUP, and INV events. We first contrasted the genome and SMRT-read data sets. For the three categories, more SVs were detected with SMRT-reads (72,492) than with genome assemblies (64,262), despite the fact that data for the former (n = 10) were a subset of the latter (n = 14; supplementary table S2, Supplementary Material online). Notably, the pipeline for genome alignments detected 10-fold fewer INV events, whereas SMRT reads yielded 10-fold fewer DUP events (supplementary fig. S3, Supplementary Material online), suggesting systematic biases in detection methods. Altogether,

### Table 1. The Number of Variants Detected by Illumina Short Reads, PacBio Long Reads, and Genome Alignment.

<table>
<thead>
<tr>
<th>Variant Type</th>
<th>Illumina</th>
<th>PacBio</th>
<th>Genome Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples</td>
<td>531,926</td>
<td>390,823</td>
<td></td>
</tr>
<tr>
<td>TRA</td>
<td>76,835</td>
<td>85,279</td>
<td></td>
</tr>
<tr>
<td>DUP</td>
<td>48,132</td>
<td>213</td>
<td></td>
</tr>
<tr>
<td>DEL</td>
<td>72,930</td>
<td>60,747</td>
<td></td>
</tr>
<tr>
<td>INV</td>
<td>341,752</td>
<td>11,532</td>
<td></td>
</tr>
<tr>
<td>INS</td>
<td>103,447</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>MEI</td>
<td>284,741</td>
<td>270,708</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>824,390</td>
<td>531,926</td>
<td>390,823</td>
</tr>
</tbody>
</table>

We called SVs from the set of 347 individuals using two approaches, following Zhou et al. (2019). First, to identify deletion (DEL), duplication (DUP), inversion (INV), and translocation (TRA) events, we combined population calls from Delly (Rausch et al. 2012) and Lumpy (Layer et al. 2014). The combined SVs were then filtered based on minimal quality scores and also on the requirement that a single SV had exact breakpoints shared among accessions (see Materials and Methods). Although this pipeline also calls insertion (INS) events, previous studies (Mahmoud et al. 2019; Zhou et al. 2019) suggest that INS are not accurately detected from short reads. INS may also overlap with MEIs, which we called using separate methods (see below). For those reasons, we excluded INS inferences from further analyses. Altogether, the approach led to a highly curated set of 824,390 biallelic SVs across the entire data set of 347 accesses (table 1 and supplementary table S1, Supplementary Material online).

We used multiple methods to assess the accuracy of the SV calls. First, as in previous papers (Gordon et al. 2017; Fuentes et al. 2019; Zhou et al. 2019), we assessed the congruence of population structure between SNP and SV calls from the same data—that is, the 347 resequenced individuals to see if they provided similar insights into population structure and history. To do so, we applied NGSadmix to the SNP data, allowing the number of groups (K) to range from 2 to 10. At K = 10, the SNP data recapitulated the expected population groupings—that is, indica and japonica were separate as were two distinct clades of rufipogon (fig. 1B). We then applied ADMIXTURE to the full data set of SVs, which identified the same major groups (fig. 1B). Per-individual assignments were strongly correlated between SV and SNP results (Pearson r = 0.853, P < 2.2 × 10^{-16}), suggesting the suitability of our SV calls for population genetic analyses.

Second, we compared our SV inferences with those from Fuentes et al. (2019), whose SVs were based on ten detection methods and 3,000 cultivars (Wang et al. 2018). For this comparison, we included only DEL, DUP, and INV events, because the size and location of TRA events were not always clear, thus leading to the potential for improper comparison. Our DEL, DUP, and INV SV calls overlapped with 90.14% of those identified by Fuentes et al. (2019). We did have an additional 9.64% of events, however, due to the added diversity within the rufipogon and outgroup samples. Nonetheless, the vast majority of SV inferences within cultivated rice agreed with previous work, suggesting that our combination of high-coverage data with two SV-callers performed well.
however, the two data sets had 57.53% of DEL, DUP, and INV events in common (fig. 1D and supplementary fig. S3, Supplementary Material online).

We then compared short-read SVs to the two other data sets, but this was an inherently biased process because the sample size was much larger for short-read data, yielding more SVs despite extensive filtering. Given this bias, we examined the overlap among data sets in a directional manner, asking: How often do Illumina SV calls identify SVs found within the other two data sets? The Illumina data set included 70.07% of DEL events identified within the whole-genome and SMRT-read data sets. 75.24% of DUP events, and a lower proportion (43.22%) of INV events (supplementary fig. S3, Supplementary Material online). Summing across the three SV types, the short-read data identified 65.35% of SVs found in the two other data sets (fig. 1D). Notably, this level of correspondence (65.35%) exceeded that between the genome and SMRT-read data sets (57.53%), again suggesting the short-read population calls were reasonable.

Population Properties of SVs

Given concordant information with SNPs and SVs and notable overlaps among long-read and short-read SVs, we focused on short-read SVs to investigate population dynamics. For simplicity, we narrowed our focus to the three groups with the largest population samples: rufipogon, indica (n = 96), and japonica (n = 106). For rufipogon, we examined only the clade of n = 40 accessions that appeared to be truly wild, based on their phylogenetic position (fig. 1A), recognizing that combining the two distinct clades would produce skewed population statistics.

We first characterized chromosomal positions of SVs, plotting SV diversity using sliding window analyses for each taxon (e.g., supplementary figs. S4–S6, Supplementary Material online). Visually, there were no compelling patterns that suggested particular regions were more prone to specific SV events. However, SV and SNP diversity were slightly but significantly correlated across chromosomal windows in all three population groups (Pearson r = 0.0332, P = 1.07 × 10⁻⁵ for rufipogon; Pearson r = 0.0637, P = 2.2 × 10⁻¹⁶ for indica; Pearson r = 0.0494, P = 1.07 × 10⁻¹⁰ for japonica; fig. 2A and supplementary fig. S7, Supplementary Material online).

We also calculated the unfolded site-frequency spectra (SFS) of the three taxa for a sample of ten individuals with high-coverage and little missing data (fig. 2B–D). Each SFS included four SV types (DUP, DEL, TRA, and INV), along with sSNPs and nSNPs and 284,741 MEIs, which we called separately using two separate pipelines (see Materials and Methods). The SFSs revealed three salient features of SV polymorphism. First, there were demonstrable differences among taxa, because there was a higher proportion of fixed variants (and fewer intermediate variants) in cultivated rice compared with rufipogon. The U-shape of the SFS from cultivated rice had been noted previously and is consistent with both enhanced genetic drift during a domestication bottleneck (Caicedo et al. 2007) and a shift in mating system. Second, in all three taxa, there was a lower proportion of fixed SVs than fixed sSNPs and nSNPs. The distributions for each SV type were significantly different from the sSNP distribution in all three taxa (P < 0.01, Kolmogorov–Smirnov test). Assuming sSNPs provide a reasonable “neutral” control, the leftward shift in the SFS suggests either that SV variants were deleterious, on an average, or that they have higher mutation rates than SNPs such that many new events have not had the opportunity to rise in frequency. Finally, the SFS varied among SV types. In all three taxa, INV events had the most extreme SFS; in each group, >90% of INV events were identified in three or fewer individuals, suggesting either strong selection or perhaps detection biases (see Discussion).

The SFSs suggest that MEIs and all SV types have lower population frequencies, on an average, than sSNPs in all three taxa. As a consequence, SVs may have generally lower linkage disequilibrium (LD) values than SNPs, with the potential for a faster decay of LD over physical distance. We calculated LD in each of the three taxa based on sSNPs, SVs, and both sSNPs+SVs, using the squared correlation coefficients (r²). The SNP data confirmed previous observations that LD decays more slowly in japonica than either indica or rufipogon (Mather et al. 2007) (fig. 2E). For example, r² for SNPs remained ~0.2 over a distance of ~100 kb for japonica, but was ~0.1 for indica and <0.05 for rupifogon over the same physical distance. Note, however, that SVs had lower r² values than SNPs for all taxa, with values that exceeded 0.1 only over very short (<15 kb) distances. The r² values were even lower when based on both sSNP+SV data (fig. 2E). These results may have important implications for detecting the effect of SVs on phenotypes.

SVs and Domestication

It is an open question in other domesticated taxa whether SV burden increases as a consequence of domestication, whether SVs provide useful insights into genomic regions potentially subjected to positive selection during domestication and whether domestication shifts genetic content. We investigated these features with our set of SVs, using the previously described population samples of indica (n = 96) and japonica (n = 106) and rufipogon (n = 40) for all analyses.

SV Burden

Because the SFS of SVs and MEIs imply that they may be deleterious, we predicted that they contribute to the deleterious load, reflecting the cost of domestication in rice (Lu et al. 2006; Günther and Schmid 2010; Liu et al. 2017). We evaluated cost by calculating the additive SV + MEI burden per individual, which is the number of derived heterozygous sites (the heterozygote burden) plus two times the number of homozygous variants (the recessive burden) (Henn et al. 2016). Comparing the additive SV + MEI burden across the three taxa, it was 35% and 25% higher on an average for japonica and indica relative to rufipogon (P < 0.005 for both contrasts, t-test; fig. 3A). Given the differences in mating system between cultivated and wild rice, we also expected the recessive burden to be the primary contributor to differences in the additive burden. This expectation held, because the recessive burden was >72% of the additive burden for both cultivars, but the proportion was lower (67%)
for rufipogon (fig. 3A). These patterns—that is, higher additive and especially recessive burdens—held across variant types, with the apparent exception of DEL events (supplementary fig. S8, Supplementary Material online).

Divergence between Domesticated Taxa and Rufipogon

In theory, the genes that contribute to domestication can be identified as regions of marked chromosomal divergence between wild and cultivated samples. We compared rufipogon to both indica and japonica by estimating SNP and SV divergence in fixed 20-kb windows across the genome. We calculated divergence with two measures ($F_{ST}$ and $D_{xy}$; supplementary figs. S9 and S10, Supplementary Material online) but focus on $F_{ST}$ results here for simplicity. Across the entire genome, mean $F_{ST}$ estimates were substantially higher for SNPs (indica–rufipogon 0.293 ± 0.134; japonica–rufipogon 0.375 ± 0.185).

**Fig. 2.** Population information about SVs. (A) The plot graphs SNP and SV average pairwise diversity ($\pi$) for the rufipogon sample across 20-kb windows of the genome, with the line indicating the correlation, which is weakly positive but significant (Pearson $r = 0.0332$, $P = 1.07 \times 10^{-5}$). Similar graphs for the japonica and indica samples are in supplementary figure S7, Supplementary Material online. Plots (B), (C), and (D) show the unfolded SFS of different types of SVs in (B) rufipogon, (C) indica, and (D) japonica. Each SFS contains synonymous SNPs (Syn), nonsynonymous SNPs (Nsyn), and SV events that fall into the duplication (DUP), deletion (DEL) translocation (TRA), mobile element insertion (MEI), and inversion (INV) categories. (E) The decay of linkage disequilibrium (LD) of SNPs and SVs measured by $r^2$ for the three population groups based on SNPs, SVs, and SNPs + SVs.
0.485 ± 0.181) than for SVs (indica–rufipogon 0.122 ± 0.079; japonica–rufipogon 0.259 ± 0.141), reflecting the fact that SVs were typically at lower population frequencies (fig. 2A).

We contrasted the two cultivars to rufipogon and ranked the top 1% \( F_{ST} \) windows (or 187 of 18,654 windows throughout the genome) for both SNPs and SVs (fig. 3B and supplementary fig. S10, Supplementary Material online). Only a small number of \( F_{ST} \) windows were with the top 1% for both SNPs and SVs (supplementary fig. S11, Supplementary Material online); for example, we detected 26 such windows for the indica–rufipogon comparison. Although a small number, 26 is far more than the ~2 windows expected at random (\( P < 10^{-3} \), permutation test), suggesting that the SNP and SV data do capture some common signatures (as is expected, given that they are in the same genomic window). Similarly, we detected 12 such windows in the japonica–rufipogon comparison, again representing an enrichment over a random draw (\( P < 10^{-3} \), permutation test). Of these, only one window overlapped between japonica and indica; it contained a gene (LOC_Os02g43800) that was annotated as a retrotransposon protein. Thus, arguably the strongest signal of positive selection based on \( F_{ST} \)—that is, in a window identified from both SNPs and SVs across both cultivated taxa—contained a gene without obvious agronomic implications.

We examined the shared SNP–SV windows for potential candidate genes (supplementary table S3, Supplementary Material online). For example, of the 82 genes contained in the 26 windows for the indica–rufipogon comparison, 31 were annotated as expressed or hypothetical proteins, and 14 were annotated as TE-related; neither category were obvious candidates to contribute to agronomic phenotypes. Most of the remaining 37 genes were assigned putative functions, including a ribosomal protein, a male sterility protein, small auxin up-regulated genes, receptor like-kinase genes, and genes with other functions. GO analyses indicated that the 82 genes were enriched for a variety of putative functions, including cellular components extrinsic to membranes and biological processes related to superoxides (supplementary table S4, Supplementary Material online). Similarly, the shared SNP–SV peaks between japonica and rufipogon contained 36 genes in 12 windows, with 15 genes assigned putative functions and GO enrichment in DNA replication and other
functions. For completeness, we also analyzed the set of genes identified in SNP-only (489) or SV-only (374) peaks for each taxon (supplementary table S3, Supplementary Material online). Not surprisingly, the genes were enriched for a variety of GO-based functions (supplementary table S4, Supplementary Material online).

Because we were interested in the utility of SVs to detect selection events, we also took a candidate gene approach to assess whether SVs enhanced their identification. To do this, we focused on a set of 15 known domestication and improvement genes (Huang et al. 2012; Wang et al. 2018) to ascertain whether they were identified in F\textsubscript{ST} scans more often than expected at random (table 2). Among the 15, six were within the top 1% of F\textsubscript{ST} windows between rufipogon and either indica or japonica (table 2); three of these were found with SNPs alone (including TAC1, a gene implicated in tillering, and shattering genes Sh1 and Sh4) and three more with SNPs alone ( Bh4, Dwarf4, and SAG13) (table 2). Overall, the set of 15 genes was highly enriched to be within F\textsubscript{ST} peaks at the 1% and 10% levels for both SNPs (P < 1.19 \times 10^{-15} for 1% peaks, P < 6.67 \times 10^{-8} for 10% peaks, Wilcoxon–Mann–Whitney test) and SVs (P < 2.20 \times 10^{-16} for 1% peaks, P < 2.54 \times 10^{-14} for 10% peaks, Wilcoxon–Mann–Whitney test), suggesting that SVs do have some utility for identifying domestication genes. We note, however, that candidate gene enrichment using SVs was not as evident based on D\textsubscript{xy} (supplementary table S5, Supplementary Material online).

Table 2. Previously Identified Putative Domestication or Improvement Genes with Their ID, Function, and % Ranking in F\textsubscript{ST} Windows.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Gene ID (MSU7/IRGSP1.0)</th>
<th>Function</th>
<th>F\textsubscript{ST} (i–r)</th>
<th>F\textsubscript{ST} (j–r)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SNPs</td>
<td>SVs</td>
<td>SNPs</td>
</tr>
<tr>
<td>sh4</td>
<td>LOC_Os06g10350/Os06g04700</td>
<td>Shattering</td>
<td>0.18</td>
<td>17.05</td>
</tr>
<tr>
<td>qWS5</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Grain width</td>
<td>79.07</td>
<td>1.44</td>
</tr>
<tr>
<td>qSH1</td>
<td>LOC_Os06g10350/Os06g0474000</td>
<td>Shattering</td>
<td>58.21</td>
<td>94.51</td>
</tr>
<tr>
<td>Sd1</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Grain quality</td>
<td>32.07</td>
<td>42.17</td>
</tr>
<tr>
<td>Wx1</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>47.21</td>
<td>52.79</td>
</tr>
<tr>
<td>Badh2.1</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>58.21</td>
<td>94.51</td>
</tr>
<tr>
<td>Rc</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>32.07</td>
<td>42.17</td>
</tr>
<tr>
<td>Bh4</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>47.21</td>
<td>52.79</td>
</tr>
<tr>
<td>PROG1</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>32.07</td>
<td>42.17</td>
</tr>
<tr>
<td>OsCl</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>47.21</td>
<td>52.79</td>
</tr>
<tr>
<td>TAC1</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>32.07</td>
<td>42.17</td>
</tr>
<tr>
<td>Dwarf4</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>47.21</td>
<td>52.79</td>
</tr>
<tr>
<td>SAG13</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>32.07</td>
<td>42.17</td>
</tr>
<tr>
<td>TB1</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>47.21</td>
<td>52.79</td>
</tr>
<tr>
<td>Sh1</td>
<td>LOC_Os06g04200/Os06g0133000</td>
<td>Flavor or fragrance</td>
<td>32.07</td>
<td>42.17</td>
</tr>
</tbody>
</table>

We again investigated the top 1% of 20-kb windows for both SVs and SNPs (fig. 3C and supplementary fig. S12, Supplementary Material online). Qualitatively the results exhibited fewer but clearer peaks compared with F\textsubscript{ST} troughs, several of the 15 genes were located in F\textsubscript{ST} peaks between indica and japonica, including Dwarf4, Sd1, and TB1 (supplementary table S6, Supplementary Material online). Altogether, these analyses reinforce the complex history of Asian rice domestication by suggesting that some, but not all, known domestication genes may have a history of introgression between japonica and indica.

Selective Sweeps within Domesticates

We also searched for taxon-specific signals of selective sweeps using the composite likelihood method (Pavlidis et al. 2013). We again investigated the top 1% of 20-kb windows for both SVs and SNPs (fig. 3C and supplementary fig. S12, Supplementary Material online). Qualitatively the results exhibited fewer but clearer peaks compared with F\textsubscript{ST} troughs, several of the 15 genes were located in F\textsubscript{ST} peaks between indica and japonica, including Dwarf4, Sd1, and TB1 (supplementary table S6, Supplementary Material online). Altogether, these analyses reinforce the complex history of Asian rice domestication by suggesting that some, but not all, known domestication genes may have a history of introgression between japonica and indica.
Gene Gain and Loss

We focused on the subset of SVs that included genes and determined whether they were private (i.e., variable within only one taxon) or fixed (i.e., in alternative states) between taxa. For example, between rufipogon and japonica, we detected 114,840 SVs that were private in rufipogon; 144,600 that were private in japonica; 180,798 that were shared SVs between taxa; and only one fixed SV, corresponding to one gene that was annotated as related to retrotransposition. Focusing on the subset of private SVs that include genetic DUP and DEL events, we found 148 genes gained and 138 lost in our japonica sample relative to the rufipogon sample. Similarly, we detected 3,410 genes gained and 181 lost in indica relative to rufipogon (supplementary table S10, Supplementary Material online). Some of the genes lost during domestication had validated functions related to physiological and morphological traits, such as sterility (LOC_Os01g11054/Os01g208700), culm leaf (LOC_Os04 g39780/Os04g0473900), flowering (LOC_Os03g05680/Os03g0151300), and stress tolerance (LOC_Os01g64970/Os01g0869900). In addition to these functions, genes inferred to be gained during domestication involved functions that contribute to eating quality, including starch storage and biosynthesis (e.g., LOC_Os01g65670/Os01g0878700, LOC_Os02g 32350/Os02g0523500, LOC_Os03g09250/Os03g0127200, LOC_Os03g49350/Os03g0700400, LOC_Os03g52340/Os03g 0733800, and LOC_Os09g26880/Os09g0404000) (Wang et al. 2018). A complete list of private genes and their GO enrichments are listed in supplementary tables S10 and S11, Supplementary Material online; the important point is that SV analyses between Asian rice and rufipogon yield reasonable candidate genes for traits involved in domestication or improvement.

MEIs for Specific TE Families

We called MEIs separately and assigned them to specific families, providing an opportunity to compare population dynamics among TE families and types. We separated MEIs into ten distinct TE families—Gypsy, Copia, LINE, SINE, CACTA, hAT, Mutator, Harbinger, Mariner, and Helitron elements—and calculated their SFSs based on calls from two methods, PoPoolationTE2 and TE-locate. Focusing on the PoPoolationTE2 results from rufipogon for simplicity (fig 4A), but with similar results from two methods used to call MEIs (supplementary figs. S14 and S15, Supplementary Material online), all TE families had fixation frequencies lower than sSNPs, with each distribution significantly different from sSNPs (P < 0.01, Kolmogorov–Smirnov test). However, there were also marked differences among TE families (fig 4A and supplementary figs. S14 and S15, Supplementary Material online). The most obvious deviation was for SINE and Mariner elements, for which only a small proportion of MEIs were fixed; hAT and Harbinger elements also demonstrated a substantial leftward relative to Gypsy, Copia, and other element types. Consistent with this observation, estimation of the distribution of fitness effects (DFEs) suggests that selection was more severe against these four TE families (fig 4B), with the lowest proportion of putatively adaptive variants (α) for SINEs (α = 0.12%), followed by Mariner (α = 1.52%), Harbinger (α = 6.32%), and hAT (α = 5.91%) (fig 4C).

What might cause apparent differences in population dynamics among TE families? One explanation concerns the timing of TE insertion; if SINEs have been active more recently, it is possible that the SFS reflects a lack of sufficient time for insertions to reach fixation. To test this idea, we estimated the insertion time of individual elements within the japonica reference (see Methods), producing a distribution of insertion times for all ten families (fig 4D). The distribution of insertion times was similar among TE families (P > 0.05, t-test) with Gypsy and Copia elements (but not SINEs) biased toward slightly more recent insertions. Hence, more recent activity does not seem to be an adequate explanation for the SFS of SINE and Mariner elements. Another explanation is stronger purifying selection against some element families. To assess this idea, we examined the distribution of MEIs relative to genes in the japonica reference. We found that a lower proportion of SINE, Mariner, Harbinger, and hAT MEIs were inserted within exons relative to other TE families (supplementary fig S14, Supplementary Material online). This observation could be fueled by insertion biases, but they may also point to stronger selection against these four families. Consistent with the latter interpretation, the ratio of homozygous to heterozygous variants was lower for these four families than for the other families (fig 4E), suggesting stronger selection when these elements are uncovered from a heterozygous state and experience recessive selection.

Discussion

SVs remain unexplored for most crops, and this is particularly true with respect to comparing population frequencies between crops and their wild relatives. Here, we have performed a genome-wide analysis of SVs in Asian rice and its wild progenitor O. rufipogon, with the goal of understanding more about the evolutionary processes that act on them and their fate during domestication. Most of our inferences have been based on SVs called from a large (347 accession) data set consisting of high-coverage (average 50×, median 28×), short-read data. Given these calls, it is important to note two important caveats. First, we have focused on biallelic SVs that were useful for population genetic inference, meaning that the SVs were filtered both for quality and to avoid complex events, such as overlapping SVs. Given this curation, it is important to convey that we did not expect, nor intend, our set of SVs to represent a comprehensive catalog, as was the intent for a previous study that lacked rufipogon samples (Fuentes et al. 2019). Second, like previous studies (Carpentier et al. 2019; Fuentes et al. 2019), all of our inferences rely on mapping to the Nipponbare genome, which may introduce a reference bias that makes it more difficult to identify novel SVs from non-japonica samples. Nonetheless, three features of our SV calls suggest they are reasonable: (1) they provide population structure information that is highly concordant to (and strongly correlated with) information from sSNPs (fig 1B), (2) they overlap substantially (>90%) with SVs...
reported previously, based on different data sets and analytical methods (Fuentes et al. 2019), and (3) there is some agreement across data types, where overlap is less impressive but still substantial (fig. 1D).

**SV Comparisons among Data Sets**

Our results are consistent with previous work claiming that different data types provide different information. For example, genome alignments can be poor for detecting heterozygous SVs, because primary assemblies ignore alternative haplotypes (Mahmoud et al. 2019). Instead, they tend to be best at identifying large SV events, such as large insertions (Nattestad and Schatz 2016). In contrast, SMRT-read mapping should be efficient for most SV detection, outperforming short-read data (Sedlazeck et al. 2018; Chaisson et al. 2019). However, the development of SMRT-read methods is still nascent (Mahmoud et al. 2019) and by no means perfected, as perhaps evidenced by the very low number of DUP events that were detected with SMRT reads (table 1 and supplementary table S1, Supplementary Material online). Finally, paired-end short reads are probably accurate for SV calls, given sufficient coverage (Layer et al. 2014), but they often miss large and complex SV events (Sedlazeck et al. 2018).
Given the strength and limitations of different approaches, our comparative data sets provide different insights into rice SVs. For example, whole-genome alignments reveal the presence of a large centromeric inversion on chromosome 6 that differentiates indica from japonica and the outgroups in our sample (fig. 1C). This inversion has been reported previously to contain 404 genes (Du et al. 2017), and it was also identified in O. brachyantha compared with indica assemblies (Liao et al. 2018). Another SV feature is genic hemizygosity, which can be estimated reasonably from SMRT reads (Zhou et al. 2019). Based on remapping SMRT reads to the Nipponbare reference, we estimate that <1% of genes are hemizygous in a japonica and an indica accession, which is consistent with the expectation of high homozygosity for predominantly selfed lineages. These estimates are low enough that they may reflect the false-positive rate of the method. In contrast, ~9% of genes are hemizygous for the O. longistaminata individual. Superficially 9% seems high, but it is similar to the ~10% PAV differences between inbred lines of maize (Sun et al. 2018) and lower than the ~10–14% genic hemizygosity of grapes (Vondras et al. 2019; Zhou et al. 2019). This observation adds to a growing appreciation that outcrossed plants harbor a substantial fraction of SVs that lead to genic hemizygosity.

Our filtering of short-read SVs was purposefully biased against the detection of complex, overlapping SV regions, because we sought to identify discrete biallelic loci for population genetic analysis. It is worth emphasizing, however, that complex SV regions do exist. For example, we used genome alignments to investigate SVs in the region surrounding an NBS-LRR gene (LOC_Os01g05600/Os01g0149350) that was detected as gained during domestication, based on short-read SVs. In this region the cultivated individuals in our sample have an additional NBS-LRR gene relative to the wild individuals; among cultivars, the region is marked by the movement of both DNA and RNA transposons that alter distances among genes (fig. 5A). Similarly, we examined the Submergence1 (SUB1) region (fig. 5B), which contains CNVs that affect flooding tolerance (Xu et al. 2006; Mickelbart et al. 2015). This locus contains a cluster of three ethylene response factor (ERF) genes, Sub1A, Sub1B, and Sub1C. Among them, only the Sub1A-1 (an allele of the Sub1A gene) confers flooding tolerance and it was present in only a few indica cultivars (fig. 5B). Altogether, these analyses accentuate the prevalence of SVs in rice (Wang et al. 2018; Fuentes et al. 2019) and the fluidity of Oryza genomes (Stein et al. 2018; Zhao et al. 2018).

**SVs Are Typically at Lower Population Frequencies than SNPs**

In each of the three taxa, the SFS of each SV type differs significantly from the SFS of “neutral” (4-fold synonymous) sSNPs. Previous work has shown that plant SVs tend to be deleterious in planet populations (Flagel et al. 2014; Gaut et al. 2018; Zhou et al. 2019). Our results are consistent with this view, but such differences could also be caused by different mutational mechanisms and rates. Nonetheless, the SFSs also suggest heterogeneity among SVs, because they follow an apparent hierarchy in which INV events have the most left-leaning SFS, followed by MEI, TRA, DEL, and DUP events. One pertinent question is whether detection biases somehow fuel this heterogeneity because, for example, INV events have a
lower percentage of overlap among data sets than DEL and DUP events (fig. 1D and supplementary fig. S3, Supplementary Material online). There are systematic biases for all SV types—for example, ascertainment biases—that may tend to skew the SFS leftward (Emerson et al. 2008). The question here is whether INVs are especially prone to these biases. We do not believe that this is the case for three reasons: (1) the SFS are based on the individuals with highest coverage, which limits false-negative results (Cridland and Thornton 2010); (2) some studies suggest similar false-discovery rate across SV types (Layer et al. 2014); and (3) SMRT-read analyses also indicate that INV events are found at low frequency, because 86% of INVs are found in only one indica individual, which is significantly lower than other SV types (P < 0.05, t-test). Overall, our results suggest that polymorphic INV events are at especially low frequencies in population samples, suggesting the possibility that they are particularly deleterious.

SVs are typically found at lower population frequencies than SNPs, which affects LD. We have shown that LD between SNPs and SVs declines far more rapidly than that of SNPs alone and slightly more rapidly than that of SVs alone (fig. 2F). This relationship may provide a challenge for association analyses, because it implies that causative SVs in rice will not be easily tagged by anchoring SNPs. There is remarkably little information about LD based on SVs, but thus far it seems as if SVs are usually not high LD with SNPs across plants. For example, 20% of maize copy number variants (Chia et al. 2012), 27% of maize SVs (Sudmant et al. 2015), and ~70% of arabidopsis MEIs cannot be anchored by nearby SNPs (Stuart et al. 2016). SVs appear to have substantial explanatory power when they are included in GWA analyses (Chia et al. 2012; Yao et al. 2015; Fuentes et al. 2019), but they may require methods that incorporate the possibility of detecting associations with SVs (Voichek and Weigel 2020).

We also examined the SFS for potential differences in the population dynamics of MEIs from different TE families, given that previous work focused only on the frequency of retrotransposons and only in cultivated accessions (Carpentier et al. 2019). We found that several element families, but particularly SINE elements, have dramatically different population frequencies than other element families. Relatively few SINEs were fixed, and most were found at low frequency. This finding does not appear to be due to more recent activity (fig. 4D) but rather to stronger selection, as implied by the low frequency of SINEs in and near genes (supplementary figs. S14 and S15, Supplementary Material online) and the relative dearth of homozygous variants (fig. 4E). This last observation supports the growing consensus that deleterious variants can accrue as heterozygotes because they are typically under recessive selection (Zhou et al. 2017; Huber et al. 2018; Zhou et al. 2019), but the mating system of rice ensures that new, heterozygous TE insertions do not remain heterozygous for long. We note that in our study retrotransposon insertions (Gypsy and Copia elements) were not found at particularly low frequencies based on two detection methods. That is, unlike Carpentier et al. (2019), we did not find that ~50% of retrotransposon insertions are at low (<5%) population frequencies (fig. 4A and supplementary figs. S14 and S15, Supplementary Material online). We suspect that the principal difference between their study and ours is that they focused on full-length—and therefore presumably functional and recently active—elements.

SVs and Domestication

To date, the fate of SVs during domestication has been investigated in only one domesticated taxon, grapevines (Zhou et al. 2019), where SVs provided insights into regions of the genome that may have been under artificial selection and into the SV burden associated with cultivation. Grapevines are, however, distinct from rice and other annuals in that they are clonally propagated and lack evidence of a domestication bottleneck (Myles et al. 2011), a history that leads to the accumulation of recessive deleterious mutations that do not affect load (Zhou et al. 2017) and leads to increased SV numbers in the domesticate (Zhou et al. 2019). Here, we have characterized SVs in both cultivated and wild rice to begin to extend the many previous studies of rice domestication to include this novel genomic component.

Domestication bottlenecks accelerate genetic drift, which can contribute to a cost of domestication. An appropriate measure of cost is the average number of deleterious variants per genome (d_g) (Moyers et al. 2018). Interestingly, d_g is not expected to vary substantially before and after a demographic shift under some conditions, such as a strict genetic bottleneck with outcrossing and additive (h = 0.5) variants (Simons et al. 2014). However, it can vary substantially with deviations from these conditions (Henn et al. 2016). For example, clonal variants tend to accrue recessive deleterious variants over time, because these variants are under selective advantage (Zhao et al. 2017). Similarly, forward simulations have shown that moderately and weakly deleterious variants accumulate under various demographic regimes (Robinson et al. 2018). Here, we have shown that the SV burden is elevated by 25–35% in our japonica and indica samples relative to rufipogon (fig. 3A). Although the estimated increase of d_g undoubtedly depends on the composition of the samples under comparison, this observation is consistent with previous studies suggesting a cost of domestication in Asian rice (Lu et al. 2006; Günther and Schmid 2010; Liu et al. 2017). Nonetheless, the potential for an SV-associated cost in rice has not been demonstrated previously.

Positive selection can also be pervasive during domestication (Wright et al. 2005; Doebly et al. 2006; Hufford et al. 2012; Zhou et al. 2017). We examined the SV and SNP data for signals of positive selection throughout the genome, relying either on divergence between rufipogon or signals of selective sweeps within cultivars. For divergence, as measured by F_{ST}, only a small number of windows fell within the top 1% of peaks for both SVs and SNPs, with 12 windows for japonica–rufipogon and 26 for indica–rufipogon. Although small, the numbers are enriched relative to random expectation. None of these windows contains obvious candidate genes, at least under our examination, but the lists of genes within F_{ST} peaks may prove useful for researchers studying domestication and
improvement traits in indica and japonica (supplementary tables S3 and S4, Supplementary Material online). The shared windows between SNPs and SVs also do not include any of the set of 15 domestication genes (table 2), but these genes are enriched significantly in high-ranking $F_{ST}$ peaks when SNPs and SVs are considered separately, lending some credibility to the basic approach. Importantly, $F_{ST}$ analyses suggest that SV calls aid the detection of divergent genomic regions, because some of the 15 genes were detected with SVs only (table 2). We thus believe that some of the regions identified by SV divergence between rufipogon and cultivated rice represent bona fide selection events.

We also searched for signals of positive selection using population-specific signals of selective sweeps. One expects a priori that SVs have lower power to detect selection, given that there are fewer of them and that they tend to be at lower standing population frequencies than SNPs (so that the relative effect of sweeps is less readily detectable relative to background frequencies). Consistent with this premise, few windows of putative selection overlapped between SNPs and SVs, although there was a slight enrichment for overlapping windows in japonica. These regions again yielded no obvious candidate genes for contributing to domestication traits, but again the list of genes may prove useful for functional analyses in rice (supplementary tables S8 and S9, Supplementary Material online).

Finally, our analysis of private SVs in cultivars versus rufipogon yielded a set of genes enriched for agronomic traits. We found hundreds of genes that differed between rufipogon and either indica or japonica (supplementary tables S10 and S11, Supplementary Material online). Many of them are implicated to contribute to abiotic stresses, such as salt tolerance, and eating quality, including starch storage and biosynthesis. Clearly, our work represents only a preliminary resolution of this question in rice; further work is necessary both to confirm the private status of genes across additional samples and to substantiate their function and potential phenotypic impacts. With few exceptions (Zhao et al. 2018; Hübner et al. 2019), the topic of gene gain and loss during domestication has been systematically understudied (Gaut et al. 2018). Our results nonetheless suggest that the genomic effects of domestication include substantial genic variation.

Materials and Methods

Data Samples and Preprocessing

We collected three kinds of data to detect SVs: whole-genome assemblies, PacBio SMRT reads, and paired-end short-read (Illumina) data. For whole-genomes, we downloaded 14 assemblies from previous publications (supplementary table S2, Supplementary Material online), including an O. longistaminata outgroup (Reuscher et al. 2018) that was used to infer the ancestral state of SVs. For the PacBio data, we gathered SMRT reads for ten accessions that were a subset of the whole-genome data set (supplementary table S2, Supplementary Material online) and included data from six indica, three japonica, and one O. longistaminata accession. For the third data set, we compiled paired-end, short-read resequencing data, requiring a minimum of 15× coverage per genome. We downloaded an initial set of genome sequences representing 393 individuals. The data for this article are all publicly available, with their sources listed in supplementary tables S12 and S13, Supplementary Material online.

Both SMRT and Illumina reads were preprocessed. SMRT reads were extracted and filtered from h5 files using Dextractor v1.0 (https://github.com/thegenemyers/DEXTRACTOR) with a minimum length 1,000 and a minimum quality score 750. Paired-end Illumina reads were trimmed to remove adapters and low-quality bases (<20) and filtered for reads <40 bp using Trimmomatic 0.36 (Bolger et al. 2014). The quality of raw and filtered reads was computed using FastQC 1.0.0 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Filtered short reads were then mapped to the reference genome japonica Nipponbare (Zhang et al. 2018) using BWA-MEM (Li and Durbin 2010). Reads with mapping qualities <10 were filtered to remove nonuniquely mapped reads using SAMtools 1.9 (Li 2011). To account for the occurrence of PCR duplicates introduced during library construction, we used MarkDuplicates in the picard-tools v1.119 (https://github.com/broadinstitute/picard) to remove reads with identical external coordinates and insert lengths. The bam files were then sorted and indexed using SAMtools for downstream analyses.

Joint Variant Calls across Population Samples

SNP Calling with Short-Read Data

We used an initial data set of 393 individuals with >15× mapped coverage that represented the five major rice subpopulations (Garris et al. 2005) (temperate japonica, tropical japonica, indica, aus, and aromatic), two wild relatives (O. rufipogon and O. nivara, which is often considered an annual form of rufipogon), and two outgroup species (O. meridionalis and O. longistaminata). We mapped these resequencing data to an updated version of the Nipponbare genome (Zhang et al. 2018), called SNPs, and then subjected the sample to clustering analyses. SNPs and short (<50 bp) indels were called for the entire data set of 393 individuals using the HaplotypeCaller in GATK v4.1.2.0. SNPs were filtered using the VariationFiltering in GATK v4.1.2.0, according to the following criteria: variant quality (QD) >2.0, quality score (QUAL) >40.0, mapping quality (MQ) >30.0, and <80% missing genotypes across all samples. More than 90% of the reads uniquely mapped to the reference genome after filtering (90.22% ± 3.79% for outgroup samples and 93.79% ± 5.02% for other samples). SNP variants were then annotated to be synonymous or nonsynonymous according to the gene annotation from MSU7 Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/) (Kawahara et al. 2013) using the SnpEff v4.0 (Cingolani et al. 2012) with the structural annotation of the reference based on Maker version 2.31.8 (Holt and Yandell 2011). These SNP variants were used for population assignment and inference but not for estimates of nucleotide diversity (see below).

Based on population structure and phylogenetic analyses of SNPs from the 393 accessions, we detected 46 individuals (11.39%) that did not cluster with their reported group of
origin, suggesting they were misidentified in public databases (supplementary table S12, Supplementary Material online). After their removal, the curated data set consisted of 347 accessions: 244 cultivated rice representing indica (n = 96), japonica (n = 106), aus (n = 24), and aromatic (n = 18) varieties; 97 wild rice, including O. rufipogon (n = 90) and O. nivara (n = 7) accessions; and outgroup accessions from O. meridionalis (n = 3) and O. longistaminata (n = 3) (supplementary table S13, Supplementary Material online). The mean coverage among the 347 accessions was 50×, with a range of 15× to 333× (supplementary table S13, Supplementary Material online), providing sufficient coverage for high-sensitivity SV calls (Layer et al. 2014).

SV Discovery
We used separate methods to infer SVs in the three different data sets. For genome alignment, we performed comparisons between pairs of assemblies with MUMmer v4 (Marçais et al. 2018). The minimum length of a single exact match (l ≤ 1000) and a cluster of matches (c < 1000), and the maximum diagonal difference between two adjacent anchors in a cluster (d ≤ 5) were set using the nucmer program (nucmer -maxmatch -noextend). Dot plots were generated to visualize chromosomal collinearity and large SVs, using mummerplot. SVs between pairs of assemblies were discovered using the show-diff program in MUMmer v4; collinear regions and SV breakpoints were shown using the show-coords program. The chromosome 6 inversion and other major SVs were verified in the IGV browser (Robinson et al. 2017) from bam files that mapped the SMRT and/or Illumina reads to both the reference genome. For DELLY and LUMPY, the joint algorithm and default parameters. Finally, the MEIs were genotyped for each individual based on the number of supporting reads; an MEI or non-MEI allele was genotyped as 1/1 (homozygous MEI); and the remaining cases were genotyped as 0/0 (homozygous non-MEI); an MEI supported by <4 reads at the breakpoints that support either allele. An MEI supported by <4 reads was genotyped as 0/0 (homozygous non-MEI); an MEI supported by most of the reads (<4 reads support non-MEI) was genotyped as 1/1 (homozygous MEI); and the remaining cases were genotyped as 0/1 (heterozygous MEI).

To assess the results of MEIs detected by PoPoolationTE2, we also identified MEIs across all 347 accessions using TElocate (Platzer et al. 2012), which is reported to have high sensitivity and precision in the detection of reference TE insertions (Vendrell-Mir et al. 2019). We again based our inferences on the TE reference in gff3 format. We included all major TE families with length >50 bp, which were extracted from the TE annotation of reference genome japonica Nipponbare (Zhang et al. 2018). Using the TE reference as the input file, the insertion events of TEs were then identified using the program TE locate.pl. Insertion events with <4 supporting reads were marked as missing, and >5 reads were considered evidence of presence of the TE in an individual. For this analysis, insertion events were further filtered if supported by less than five individuals (1% minimum frequency) across population samples.
Population Genetic Analyses

Our variant calls resulted in filtered bam files, a vcf file for SNPs, a vcf for SVs, and MEI genotypes based on population samples that were used in downstream evolutionary genomic analyses. Only biallelic variants were used. The annotation files used in this article, along with the unfiltered vcfs are available at https://zenodo.org/deposit/3758509.

Population Structure

We used the SNP vcfs to examine population structure. These analyses were performed in ANGSD v0.929 (Korneliussen et al. 2014) using genotype likelihoods in the beagle file as an input to NGsadmix. Population structure inference was based on SNP variants with a minimal quality score of 20 and a minimal mapping quality of 30. The number of genetic clusters (K) ranged from 2 to 10, and the maximum iteration of the EM algorithm was set to 2,000. The filtered SNPs across all samples were used to construct a phylogenetic tree in the FastTree v2.1.11 program with GTR+CAT model (Price et al. 2009) with O. longistaminata and O. meridionalis used as outgroups. The assignment tables from SNP and SVs were compared by flattening the matrix and calculating the Pearson correlation coefficient.

The population structure inference for SV variations were conducted using ADMIXTURE 1.3.0 with a block relaxation algorithm (Alexander et al. 2009). The termination criterion for the algorithm was when the log-likelihood increased by <0.0001 between iterations. The binary fileset (.bed) as ADMIXTURE’s input was created from the SV vcf by PLINK 1.90 (Purcell et al. 2007). For downstream population genetic analyses, we only chose samples with clear classifications supported by population structure analyses.

Population Genetic Statistics

Given SNP calls on the filtered set of 347 accessions, we used ANGSD v0.929 (Korneliussen et al. 2014) to estimate genome-wide nucleotide diversity, primarily to assess whether our samples reflected the well-substantiated hierarchy of diversity within Oryza and were therefore reasonable representatives of genetic diversity. The data did indeed recapitulate the known hierarchy—that is, rufipogon had higher diversity ($\theta_w = 0.0213 \pm 0.0022$, $\pi = 0.0129 \pm 0.0015$, $n = 90$) than indica ($\theta_w = 0.0100 \pm 0.0017$, $\pi = 0.0094 \pm 0.0018$, $n = 96$), which was more diverse than japonica ($\theta_w = 0.0057 \pm 0.0012$, $\pi = 0.0039 \pm 0.0012$, $n = 166$). Because these diversity estimates were based on genomic likelihoods, they were higher and likely more accurate (Kim et al. 2011) than previous genome-wide reports based on filtered SNPs (Huang et al. 2012).

Linkage disequilibrium decay along physical distance was measured by the squared correlation coefficients ($r^2$) between all pairs of SNPs, SVs, and all variants (SNPs + SVs) within a physical distance of 300 kb using PopLDdecay (Zhang et al. 2019). Genome-wide genetic diversity was assessed from genotype likelihood in the ANGSD v0.929 (Korneliussen et al. 2014), based on SNP variants. The —doSaf option was used to calculate the site allele frequency likelihood at all sites, and then the —realSFS was used to obtain a maximum likelihood estimate of the unfolded SFS using the EM algorithm (Kim et al. 2011). Population genetic statistics, including Watterson’s $\theta_w$ and pairwise differences $\pi$, were calculated for each population group using the thetaStat program (Korneliussen et al. 2014). Genetic diversity ($\pi$) for SNPs and SVs in each group was compared using VCFtools v0.1.15 (Danecek et al. 2011).

The unfolded SFS was calculated from the allele counts for each position using three O. longistaminata and three O. meridionalis accessions as outgroups. For japonica, indica, and rufipogon, we downsampled to ten samples with the highest coverage and the least missing data to calculate the SFS for each variant type, including sSNPs that were outside outlier windows based on SweeD analyses (see below), nSNPs (Nsyn), and SVs (DEL, DUP, TRA, INV, MEI). For MEIs, we also classified them into ten families, including four retrotransposon families (Gypsy, Copia, LINE, SINE) and six DNA transposon families (CACTA, Mutator, Helitron, hAT, Harbinger, Mariner), and calculated the SFS for each family. The number of derived alleles were calculated for each type of variant using O. longistaminata and O. meridionalis as outgroups. We excluded sites with missing data at all six outgroup samples in the SFS estimation. The genetic burden was calculated under additive ($2 \times$ homozygous variants + number of heterozygous variants) (Henn et al. 2016; Zhou et al. 2017).

Selection on SVs and SNPs

SweeD v3.3.2 (Pavlidis et al. 2013) was used to detect genomic signatures of selection in the indica, japonica, and rufipogon samples, based on a sliding window size of 20 kb. The genes underlying the outlier windows were then annotated based on the MSU7 annotation (Kawahara et al. 2013). Gene Ontology (GO) analyses for these genes were run in agriGO v2.0 (Tian et al. 2017).

We used the SweeD results to define neutral sSNPs, because we assumed that sSNPs outside putative selective sweeps were neutral. The neutral sSNPs were used for calculating the SFS; the sSNP SFS was compared with the SFS of other variant types using the Kolmogorov–Smirnov test in R v3.5.1. For the SFS of individual TEs, we used the unfolded SFS to estimate DFE and $\alpha$, using polyDVE v2.0 (Tataru and Bataillon 2019). The results were presented with 95% confidence intervals obtained from the inferred discretized DFEs from 20 bootstrap data sets.

Pairwise genetic differentiation ($F_{ST}$) and genetic divergence ($D_{xy}$) for SNPs and SVs along chromosome between each pair of three groups, indica, japonica, and rufipogon, were estimated using VCFtools v0.1.15 (Danecek et al. 2011) and PBScan v1.0 (Hämälä and Savolainen 2019) with 20-kb fixed windows. Genes underlying the $F_{ST}$ and $D_{xy}$ outlier windows were annotated based on the MSU7 annotation (Kawahara et al. 2013). GO analyses were conducted in agriGO v2.0 (Tian et al. 2017).

To examine gene gain and loss during domestication, we identified private sites in each species, and also identified
shared and fixed sites between each pair of the three groups (japonica, indica, and rufipogon). The corresponding genes were inferred based on the MSU7 annotation (Kawahara et al. 2013). GO analyses were conducted in agriGO v2.0 (Tian et al. 2017) for each category.

**TE Analyses on the Nipponbare Reference**

To calculate parameters such as TE insertion time and distance from gene, we focused on the Nipponbare reference and relied on its TE annotation (Zhang et al. 2018) and the gene annotation from MSU7 annotation (Kawahara et al. 2013). Given the TE annotations, a multiple alignment file was generated for each TE family using MAFFT v7.305b with FFT-NS-2 method (Katoh et al. 2002; Katoh and Standley 2013). A consensus sequence of each TE family was extracted from the multiple alignment, and the sequence divergence between each TE copy and the consensus sequence was calculated using EMBOSS 6.5.7.0 (Rice et al. 2000). The TE insert time for each TE copy was estimated based on the sequence divergence ($d_{st}$) and a substitution rate $6.5 \times 10^{-9}$ substitutions per site per year (Gaut et al. 1996).

**Data Availability**

The gene and TE annotations of the reference genome and raw SV calls are available at https://zenodo.org/record/3758509. The LASTZ Pipeline for SV detection is available at https://github.com/yiliao1022/lastz_sv_pipeline.

**Supplementary Material**

**Supplementary data** are available at Molecular Biology and Evolution online.

**Note.** — $F_{ST}$ (i–r) refers to $F_{ST}$ between indica and rufipogon, with additional columns for japonica–rufipogon (j–r). Cells are colored as to whether the gene is in a 1% peak (dark gray), a 10% peak (medium gray), or an extreme valley (≥90%) (light gray) based on either SNPs or SVs.

**Acknowledgments**

B.S.G. is supported by National Science Foundation Grant Nos 1741627 and 1655808. J.E.E. is supported by National Institutes of Health Grant No. R01GM123303-1. Y.K. was supported by the China Scholarship Council (201808360265) and the National Natural Science Foundation of China (31901222). This work was made possible, in part, through access to the Genomics High-Throughput Facility Shared Resource of the Cancer Center Support Grant CA62203 at the University of California, Irvine, and NIH shared-instrumentation grants 1S10RR025496-01, 1S10OD010794-01, and 1S10OD021718-01.

**References**


