Structural variants in 3000 rice genomes

Roven Rommel Fuentes,[1,2,13] Dmytro Chebotarov,[1,13] Jorge Duitama,[3,4] Sean Smith,[5] Juan Fernando De la Hoz,[4] Marghoob Mohiyuddin,[6] Rod A. Wing,[1,7,8] Kenneth L. McNally,[1] Tatiana Tatarinova,[9,10,11,12] Andrey Grigoriev,[5] Ramil Mauleon,[1] and Nickolai Alexandrov[1]

[1] International Rice Research Institute, Laguna 4031, Philippines; [2] Bioinformatics Group, Wageningen University and Research, 6708 PB Wageningen, the Netherlands; [3] Systems and Computing Engineering Department, Universidad de Los Andes, Bogotá 111711, Colombia; [4] Agrobiodiversity Research Area, International Center for Tropical Agriculture (CIAT), Cali 6713, Colombia; [5] Biology Department, Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey 08102, USA; [6] Roche Sequencing Solutions, Belmont, California 94002, USA; [7] Arizona Genomics Institute, University of Arizona, Tucson, Arizona 85721, USA; [8] King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia; [9] Department of Biology, University of La Verne, La Verne, California 91750, USA; [10] Vavilov Institute of General Genetics, Moscow 119333, Russia; [11] A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127051, Russia; [12] Laboratory of Forest Genomics, Siberian Federal University, Krasnoyarsk 660041, Russia

Investigation of large structural variants (SVs) is a challenging yet important task in understanding trait differences in highly repetitive genomes. Combining different bioinformatic approaches for SV detection, we analyzed whole-genome sequencing data from 3000 rice genomes and identified 63 million individual SV calls that grouped into 1.5 million allelic variants. We found enrichment of long SVs in promoters and an excess of shorter variants in 5′ UTRs. Across the rice genomes, we identified regions of high SV frequency enriched in stress response genes. We demonstrated how SVs may help in finding causative variants in genome-wide association analysis. These new insights into rice genome biology are valuable for understanding the effects SVs have on gene function, with the prospect of identifying novel agronomically important alleles that can be utilized to improve cultivated rice.

[Supplemental material is available for this article.]

Genomics accelerates biotechnological discoveries and advances in crops and livestock, particularly by identifying genetic markers and characterizing molecular mechanisms behind desirable traits that will aid in generating new varieties through marker-assisted breeding and genome editing. This is of particular importance for rice, which needs an estimated 26% increase in yield to meet the global demand by the year 2030 under constraints such as less arable land, less water, and severe environmental stresses due to climate change (Seck et al. 2012).

To help address this yield gap, we intend to catalog all natural variation that exists in cultivated and wild rice and utilize that information to identify genes and genomic regions that can be used to drive the next generation of super crops. As an initial foray, we resequenced 3010 rice genomes (3K RG) and discovered ~20 million SNPs upon alignment to the Nipponbare reference sequence (Alexandrov et al. 2014; The 3000 rice genomes project 2014). Further efforts expanded this database by integrating short insertions and deletions (indels) into the data set (Mansueto et al. 2017). Recent studies, however, reveal that single-nucleotide polymorphisms (SNPs) do not capture the entire spectrum of variations contributing to phenotypic differences, and structural variants also play an important role (Saxena et al. 2014; Francia et al. 2015).

Detection and characterization of structural variants (SVs) has revolutionized the understanding of the landscape of genomic variation in different species. A structural variant is commonly defined as a change in the genome (relative to a reference genome) that has a different copy number (i.e., gain, loss, deletion), orientation, or chromosomal location (Medvedev et al. 2009; Escaramís et al. 2015). In human genomes, structural variants account for more varying base pairs than SNPs (Alkan et al. 2011; Baker 2012; Sudmant et al. 2015); yet, in plants, studies of SVs are still limited (Saxena et al. 2014). Although less common than SNPs, structural variants have a greater potential to impact function due to their larger size and the possibility of altering gene structure, dosage, or location (Layer et al. 2014).

After the discovery that structural genomic variation in human genomes is common, more SV studies were initiated in other species, from the agriculturally important (Swanson-Wagner et al. 2010) to extinct ones (Smith et al. 2017b). However, identification of SVs generally has lagged behind finding single-nucleotide variants due to the lack of high-quality reference genomes (Escaramís et al. 2015) and robust methods, both of which are needed to discover and genotype SVs. In plants, structural variants are not recognized as polymorphisms affecting individual plants but as differentiating elements between cultivars/accessions of one species (Francia et al. 2015). Maize became the first plant species to be extensively interrogated to discover hundreds of SVs. Although the number of SVs detected was later found to be an

[13] **These authors contributed equally to this work.**

Corresponding authors: nickolai.alexandrov@gmail.com, ttatarinova@laverne.edu, ja.duitama@uniandes.edu.co, andrey.grigoriev@rutgers.edu, r.mauleon@irri.org

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.241240.118. Freely available online through the *Genome Research* Open Access option.

underestimate, the high level of SVs in maize was unprecedented among higher eukaryotes (Żmieńko et al. 2014). Another large plant genome sequencing initiative started in 2008 developed a catalog of genetic variation in 1135 *Arabidopsis* accessions (The 1001 Genomes Consortium 2016).

Several studies in plants have already shown the association between structural variants and plant phenotypes (Żmieńko et al. 2014). For example, the increased copy number of *Vrn-A1* and *Ppd-B1* genes in wheat causes late flowering and early flowering, respectively (Würschum et al. 2015). Furthermore, a specific tandem duplication in wheat that covers the *Rht-D1b* gene results in a >70% reduction in plant height (Li et al. 2012). SVs have also been linked to stress tolerance phenotypes in crop plants such as boron tolerance in barley (Sutton et al. 2007) and nematode resistance in soybean (Cook et al. 2012).

Previous studies on rice (*O. sativa*) have identified structural variants by comparison of rice genome to its closest relatives in genus *Oryza* (Hurwitz et al. 2010) and between representatives of its major subgroups (Schatz et al. 2014) and elucidated association between structural variants and rice phenotypes using multiple rice accessions (Xu et al. 2012; Duitama et al. 2015). Examples of SVs affecting rice traits include the 17.1-kb tandem duplication at the *GL7* locus (Wang et al. 2015) that increases grain length, the 1.2-kb deletion in *qSW5* that alters grain width (Shomura et al. 2008), the 833-bp deletion that causes dwarf phenotypes and smaller grains (Ashikari et al. 1999), and the 10-bp deletion that results in slender grains (Wang et al. 2012). Recently, an extensive study on genomic variants including 90,000 SVs larger than 100 bp in the 3K RG was published, relying on a single SV caller (Wang et al. 2018) applied to a subset of samples with high coverage.

The mutational mechanism for structural variants formation includes nonallelic homologous recombination, nonhomologous end-joining (NHEJ), shrinking or expansion of variable number tandem repeats, and transposable element insertion (TEI) (Lam et al. 2010; Yi and Ju 2018). In the human genome, NHEJ and TEI are the major mechanisms for SV formation (Lam et al. 2010; Yang et al. 2013).

Structural variants can be classified in the following types: deletions, insertions, duplications (tandem and interspersed), inversions, and translocations. There are five general strategies to detect SVs based on analysis of data from high-throughput sequencing data using short reads: paired-end mapping (RP) (Chen et al. 2009; Sindi et al. 2009), split-read mapping (SR) (Schröder et al. 2014), read depth (RD) (Abyzov et al. 2011; Duitama et al. 2014; Smith et al. 2015), de novo assembly (AS) (Narzisi et al. 2014; Rizk et al. 2014; Yang et al. 2015), and a combination of the preceding approaches (CB) (Ye et al. 2009; Rausch et al. 2012; Layer et al. 2014; Mohiyuddin et al. 2015; Smith et al. 2017a). Each of these strategies has different strengths and weaknesses in detection, depending on variant type, sequence length, and reference genome quality and complexity; hence, applying complementary methods and combining results can overcome some of the limitations inherent to these different approaches (Alkan et al. 2011).

Despite the development of many SV callers, SV discovery remains challenging due to the complexity of some structural variant events and their occurrence in repetitive regions (Sudmant et al. 2015). For example, 45% of the rice genome consists of repetitive sequences (Ouyang and Buell 2004), complicating read mapping and reducing accuracy of breakpoint predictions. Aside from the performance of the callers, the nature of the data set greatly influences the quality of prediction. Many studies suggest that sensitivity, specificity, and breakpoint accuracy are dependent on read length, insert size, and physical coverage (Alkan et al. 2011). Because the average sequence coverage of the 3K RG data set is 14× depth, the use of one single method for SV detection may result in a high error rate.

In this study, we combined multiple approaches and developed a robust SV prediction pipeline to identify more than 63 million structural variants grouped into 1.5 million SV events across 3000 rice genomes and performed further analyses to confirm their accuracy. This set of SVs represents an important public resource cataloging genome variation across the main rice varieties and provides new insights for the discovery of genes related to different traits and for studying the possible roles of structural variants in rice.

Results

SV clusters number 1.5 million within *O. sativa*

Based on the benchmark of 10 diverse SV-finding algorithms (Supplemental Methods; Supplemental Fig. S1; Supplemental Table S2), we built a custom SV calling pipeline and used it to detect deletions, insertions, tandem duplication, and inversions on the 3K RG data set and alignment files (https://aws.amazon.com/public-datasets/3000-rice-genome/). Instead of relying on a single caller, we combined multiple variant callers with the best sensitivity and precision across different sizes and types of SVs. We identified a total of 63,441,115 SV calls (Table 1) across the 3K RG data set and grouped them into 1.5 million SVs clusters or events. The clusters were defined by grouping together SV calls in different samples that are likely to correspond to single evolutionary events. This grouping was based on the similarity in sizes and positions of SVs, with an average distance between breakpoints on either side of each cluster of 2.2% of the respective SV length (Supplemental Fig. S2). The frequency distribution for each type of SV follows the power law (Fig. 1A), consistent with expectation from the neutral theory of evolution (Fu 1995). Compared to SVs discovered by Wang et al. (2018), our data set covered 80.5% of their detected SV sites (Supplemental Fig. S3) within the same subset of samples with high coverage; however, we also report insertions and variants smaller than 100 bp, applied more stringent clustering criteria, and used all 3K RG samples for SV detection.

To further validate SVs detected by our pipeline, we compared the reference genome of Nipponbare (IRGSP 1.0) (Kawahara et al. 2013) with the published genome of N 22 (Pacific Biosciences [PacBio] assembly) (Stein et al. 2018) by visually inspecting predicted

Table 1. Distribution of structural variants per SV type

SV type	Number of calls (from all samples)	Number of events (clusters)	Number of events (MAF > 0.01)	Transposable elements (events)[c]
Deletion	44,143,199	834,763	116,733	106,983
Insertion[a]	14,411,023	413,740	82,720	16,159
Duplication	3,631,860	78,879	11,729	3485
Inversion	1,255,033	210,301	3362	7946
CNV[b]	31,093,780	207,927	129,088	44,021

[a]Detected in 562 high-coverage samples.
[b]Detected in 938 samples with >15× read depth and normal distribution; 50% reciprocal overlap was used for TE classification.
[c]SV size > 50 bp.

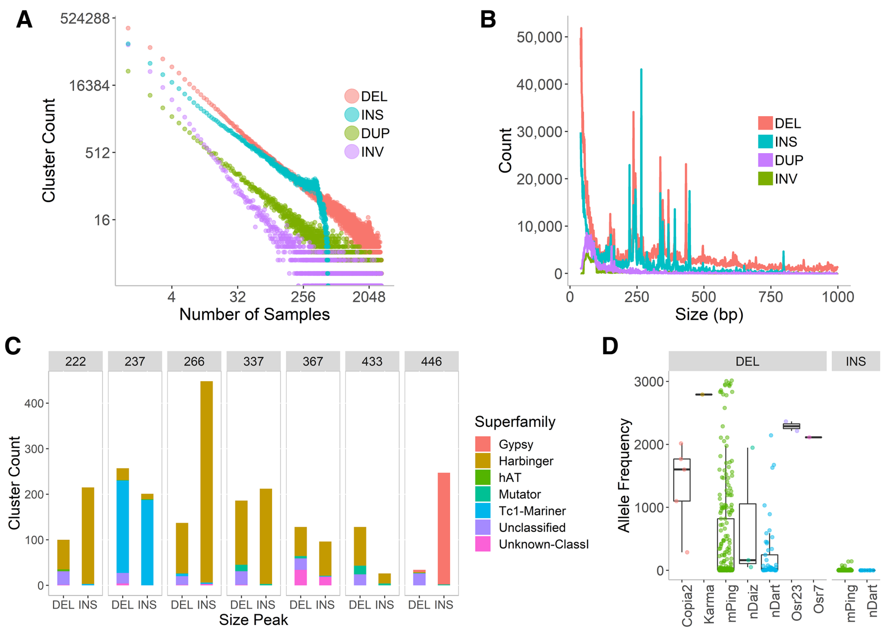

Figure 1. Distribution and classification of SVs. (*A*) Frequency of observations per SV cluster. Only 562 high-coverage samples were used for insertion detection. (*B*) Distribution of variant sizes by SV type. (*C*) Classification of variants in each peak (cluster frequency > 10 samples). (*D*) Frequencies of events with 98% sequence identity to known or potentially active TEs in rice.

variants using a dot plot display (Krumsiek et al. 2007) of the two genomes aligned against each other (Supplemental Fig. S17) and calculated false positive rates (FPR) and false negative rates (FNR) for random variants predicted in the sample CX368, an N 22 accession. The FPR for deletions is 14%, whereas duplications and inversions have much higher FPRs, 40% and 75%, respectively (Supplemental Table S4). Predicted false positive rates of the pipeline across different types of variants compare favorably with the performance of individual tools; for example, see extensive benchmarking of several leading SV callers on human genomes from Illumina sequencing (Smith et al. 2017b). The false negative rate for detecting deletions is ∼40%, consistent with the use of multiple callers whereby caller-specific SVs are often discarded and dependent on the quality of the sequences. The number of detected inversions and tandem duplications is very low compared to the other types. Although this observation can reflect a limitation of the detection pipeline, it is consistent with studies in other organisms (Quinlan et al. 2010). Due to the lower coverage of sample CX368 and its non-normal read depth distribution (Supplemental Figure S18a), we were not surprised by somewhat higher FPR and FNR, and we could not validate insertions and CNVs predicted based on read depth following this procedure (for precision of CNV, see Supplemental Figure S18b).

Transposable elements

Transposable elements play a major role in creating SVs. Among the 8.7% SV events (17.3% SVs calls) that have 80% reciprocal overlap (at least 80% of the TE is covered by SV, and at least 80% of the SV is covered by TE) with known transposable elements (TEs) in Nipponbare annotated in the RiTE database (Copetti et al. 2015), the Harbinger superfamily had the largest SV event contribution, followed by the Tc1-Mariner and Mutator TE families (Supplemental Table S1). Of the remaining clusters, 42.5% also overlap on at least 50% of their length with the TE and repetitive regions, comprising ∼45% of the rice genome (Ouyang and Buell 2004), but with lower intersection percentages.

Most peaks in the distribution of SV sizes (Fig. 1B) match TEs defined in the RiTE database with major peaks associated with events related to the Tc1-mariner (DTT), Harbinger (DTH), Mutator (DTM), Gypsy (RLG), and SINE (RSU) families of transposable elements (Fig. 1C). Peaks at 237, 433, and 466 bp consisted of the OsT38 family of Tc1-mariner (Lu et al. 2012) elements, mPing elements (Jiang et al. 2003; Naito et al. 2014), and the long terminal repeat (LTR) of the gypsy-type retrotransposon RIRE2 (Ohtsubo et al. 1999), respectively. Last, we found that 87.6% of the 155-bp duplication peak is composed of centromeric repeats (SRC).

Using the MEME Suite (Bailey et al. 2009), we analyzed the deletion sequences corresponding to the 237- to 238-bp peaks and found that 81.1% have rice-specific 95-bp terminal inverted repeats (TIR). Insertions of 237 and 238 bp were found to have the same 95 bp TIR and were classified as Tc1-mariner elements, one of the superfamilies that generates the majority of the miniature inverted-repeat transposable elements (MITEs) in rice (Han et al. 2013). Of the deletions with TIR, 90% (40,212) belong to 191 (28%) clusters with a frequency above 0.1. We identified 730 genes that have insertion/deletion of MITEs in the 3′ UTRs, which may result in the translational repression of the gene as shown by Shen et al. (2017). Their conserved lengths are consistent with observations of the OsT38 family of Tc1-mariner (Lu et al. 2012).

Other known active transposable elements also matched several events (Fig. 1D). Supplemental Figure S19 shows more TE families that matched SVs and that mPing inserts preferentially in introns, whereas nDart tends to insert in 5′ UTRs.

Supplemental Table S1 presents statistics of transposable and repeat elements (size >50 bp) among the detected structural variants. Even after aggregation of insertions and duplications, detected events are significantly rarer than deletion events: Counting SV events, the ratio of the number of deletions to the combined number of insertions and duplications is 5.45. There are several reasons for this imbalance. First, when sequences of any two genomes are compared, deletions in one genome are detected as insertions in the other. Because we do not have the ancestral genome, unaffected by expansion of TE, we must compare three thousand genomes of cultivated rice to the Nipponbare reference. All insertions of transposable elements in the evolution of *Japonica* rice or even particular to Nipponbare will be predicted as deletions in other accessions in a reference-based analysis. Second, insertion and other events (especially longer ones) in accessions different from Nipponbare are much harder to detect using short NGS reads compared to the deletion events. Hence, a large proportion of nondeletion events, therefore, may go undetected. Comparison of length of different TEs supports the importance of this factor. A typical LINE element can be as long as 6000 base pairs, whereas

LTRs range from 100 to 5000 base pairs. Supplemental Table S1 shows that predicted LINE element deletions are almost 100 times more common compared to predicted insertions, whereas for various subclasses of LTR, the ratio can be as small as 1.5.

This tendency of excess deletions occurs for both retrotransposons (Class I; copy and paste) and DNA transposons (Class II; cut and paste). Overall, there are more deletion events for Class II transposons. This is a combined methodological and biological effect that can be illustrated by the following thought experiment. Assuming that there are copy-and-paste TE events in each of the 3K rice genomes. When projected onto the reference genome, they should be manifested as insertions. Because insertions are hard to detect using NGS, many of them go undetected. When a cut-and-paste event occurs in the same genomes, the "cut" site is seen as a deletion event, and the "paste" site can be either detected or missed. This leads to a greater number of deletions of Class II TEs (Supplemental Table S1). The copy-and-paste mechanism ensures a relatively higher number of insertions for Class I TEs. Our observations support both of these hypotheses in which the ratio of the number of detected deletions to insertions and duplications for Class I is 2.36, and for Class II (cut-and-paste mechanism), it is 6.77.

To normalize the described imbalance effect produced by the methodology, for each type of SV instead of raw numbers, we compared the distribution (percentage) of events intersecting each TE family with the general distribution of TEs annotated in the rice reference genome. If SVs are randomly distributed, both distributions should be similar. For CNV and tandem duplications, we identify mostly Gypsy and Copia Class I elements (Supplemental Data, sheet "TE SVs"). Copia elements are more enriched in duplication (17.25%) than in deletion CNVs (10.46%). For the Class II elements, CNVs show enrichment for CATCA elements, probably because they are longer on average (860 bp) than other Class II elements. The Helitron family appears to be enriched in deletion CNVs (7,18%). Deletions, insertions, and inversions found by RP approaches are comprised mostly of Harbinger, Mutator, and Mariner Class II elements. Despite the larger numbers, percentages of deletions in Class I elements are similar to the general percentages of these elements. In contrast, Gypsy elements are enriched in insertions (17.15%) and inversions (12.52%). Cluster sizes for deletions related to Copia and Gypsy elements are much larger on average than clusters for other elements. This suggests that many of these events may be true insertions in *Japonica*. Despite the shortcomings of each method to detect SVs, our data agree with the expected footprints of historical activity of Class I and Class II transposable elements in contributing to variety-specific (or type-specific) differences in genome structure.

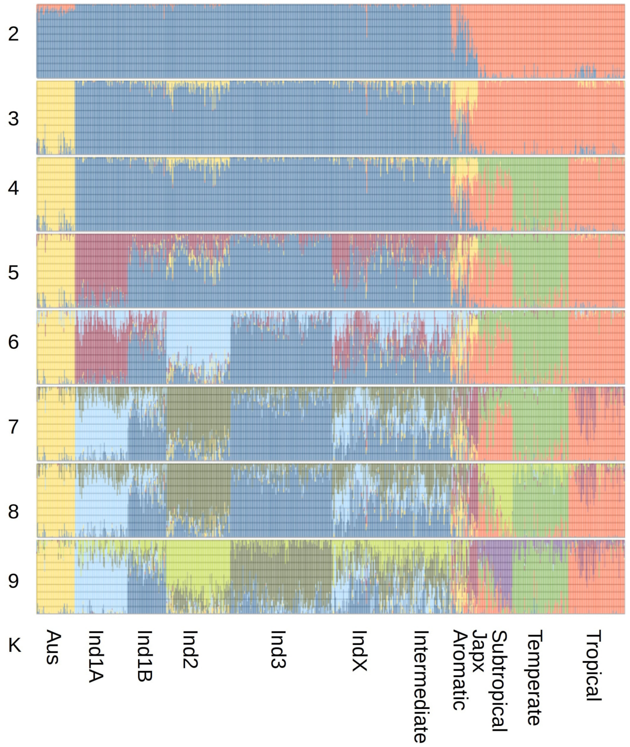

Figure 2. Structure analysis based on selected CNVs and assuming K= [2, …, 9] subpopulations.

Population structure derived from SV calls

To validate the catalog of SVs described in this study, we selected different types of variants and performed population structure analyses taking individual SV calls as alleles of genetic markers to verify whether this structure is consistent with that inferred from SNP markers. For the case of copy number variants (CNVs) genotype calls, we selected 7515 CNVs genotyped in at least 800 of 938 selected samples (for details, see Supplemental Methods; Supplemental Fig. S13) and having the major allele in a maximum of 99% of the samples. Then, using the predicted copy number of these CNVs in each accession as alleles of genetic markers, we performed a population stratification analysis with Structure (Pritchard et al. 2000) and verified that the population structure derived from CNVs is consistent with that obtained from genome-wide SNPs (Wang et al. 2018). Figure 2 shows that indeed, CNV genotyping data can distinguish the three major rice subpopulations of: *Indica*, *circum-Aus*, and *Japonica*. At $K=4$, *Japonica* is separated into temperate and tropical types. At $K=5$, the *Indica* group 1A emerges. From $K=6$ to $K=9$, *Indica* is further divided in the groups *Indica* 1A, 1B, 2, and 3, whereas *Japonica* is divided in temperate, tropical, subtropical, and admixed *Japonica* types. All of these groups are consistent with the clustering derived from SNP markers. We also tried to reconstruct population structure from 2839 CNVs genotyped in at least 2000 of the 3023 samples, located in nonrepetitive regions of the genome that have the major allele in at most 80% of the samples. In this case, the main populations could still be differentiated but the signal was less clear (Supplemental Fig. S14a). This was probably a result of the larger percentage of missing data in this data set (27.9%) compared to that of the data set shown in Figure 2 (9.02%) as well as the lower-quality predictions of CNVs for samples sequenced at low (<15×) average read depths.

We also conducted principal component analysis (PCA) on the deletion data set using all and high-coverage samples and found agreement with clustering defined by genome-wide SNP data (Wang et al. 2018). In particular, the first two principal components (PCs) separate major groups (Supplemental Fig. S4c), and PC 6 and 7 separate the *Indica* subgroups (Supplemental Fig. S4b,d).

Distribution of SVs relative to gene models

About 74.6% of SV clusters lie in intergenic space, but only 5.8% intersect with exonic regions. Nevertheless, in the 3K RG data, we found that 72.6% of the gene models supported by full-length mRNA sequence overlap with SV clusters having MAF > 0.005, and 47.6% of their coding regions are affected by SVs. Structural variants occur more often in intergenic and promoter regions and are depleted in genic regions, especially in coding DNA sequence (CDS) (Fig. 3A). Figure 3B shows the similarity between distributions of SVs and SNPs (Tatarinova et al. 2016; Triska et al. 2017), where a higher density of variation was observed in the intergenic space. The excess of SV events upstream of core promoters is likely due to transposon-related SVs.

Figure 3C shows that there is a significant difference in distributions between short (<40 bp long) and long deletions. Short deletions peak in the 5′ UTR region, and long deletions are most frequent in the promoter region. Additionally, we examined short indels identified by the GATK (McKenna et al. 2010) pipeline and found that the peak in the 5′ UTR consists mostly of variants with sizes in multiples of 3 nt (Supplemental Fig. S5; Supplemental Fig. S15). Using the repeat finder SSRIT (Temnykh et al. 2001), we found that 42.4% of the small deletions within the region [transcription start site (TSS),TSS + 125] are simple sequence repeats (SSRs) with trimer or hexamer motifs (Supplemental Data, sheet "SSRs in UTRs"). Abundance of short indels in 5′ UTRs can be explained by high density of SSRs (mostly triplets) (Lawson and Zhang 2006), resulting in low sequence complexity (Supplemental Fig. S6).

To evaluate the effects of SVs on regulatory elements, for each position around the TSS, we performed a test for independence between the presence of transcription factor binding sites (TFBS) and deletions along the promoter sequence and found a significant negative correlation between the presence of TFBS and deletions near TSSs (Fig. 3D). In positions where *P*-values are close to 1 [log (P) close to zero], there is no mutual avoidance between TFBS and deletions. In the region [TSS-200, TSS + 50], there is significant mutual avoidance between TBFS and deletions, because the deletions in this area may be detrimental to the plant. Therefore, we hypothesize that presence of a functional TFBS in the core promoter decreases the chance of a structural variant to be retained.

Association of SV-rich regions and stress response genes

Using 100-kb sliding windows with 50-kb overlaps, we computed the density of SNPs (Mansueto et al. 2017) and SVs across the rice genome and found that SNPs and SVs correlate ($r = 0.52$; $P = 2.2 \times 10^{-16}$) (Supplemental Fig. S7). We retrieved windows with SV counts of twice or more than the mean and performed an enrichment analysis using GO annotations. Supplemental Figure S8 shows highly similar distributions of SNPs and SVs across the genome and the colocalization of SV spikes and with enriched GO categories like "stress response" (Fisher's exact test; $P = 2.8 \times 10^{-5}$). At least 83.4% of the genes from the enriched categories overlapped a deletion (MAF > 0.005) (Supplemental Data, sheet "SV-Rich Genes").

We also investigated possible functional roles of the genes affected by CNVs by selecting genes for which at least 80% of its genomic location was covered by a CNV used to perform structure analysis. Executing ontology term enrichment analysis by agriGO (Tian et al. 2017), we found that genes affected by CNVs were enriched for the biological processes of cell death and response to stress (Supplemental Fig. S14b). Enriched molecular functions include kinase activity and nucleotide binding (Supplemental Fig. S14c). This result is consistent with the previous analysis of Bai et al. (2016) on a more limited data set of deletions occurring in 50 accessions and suggests that copy number variation could play a role in the plant defense system.

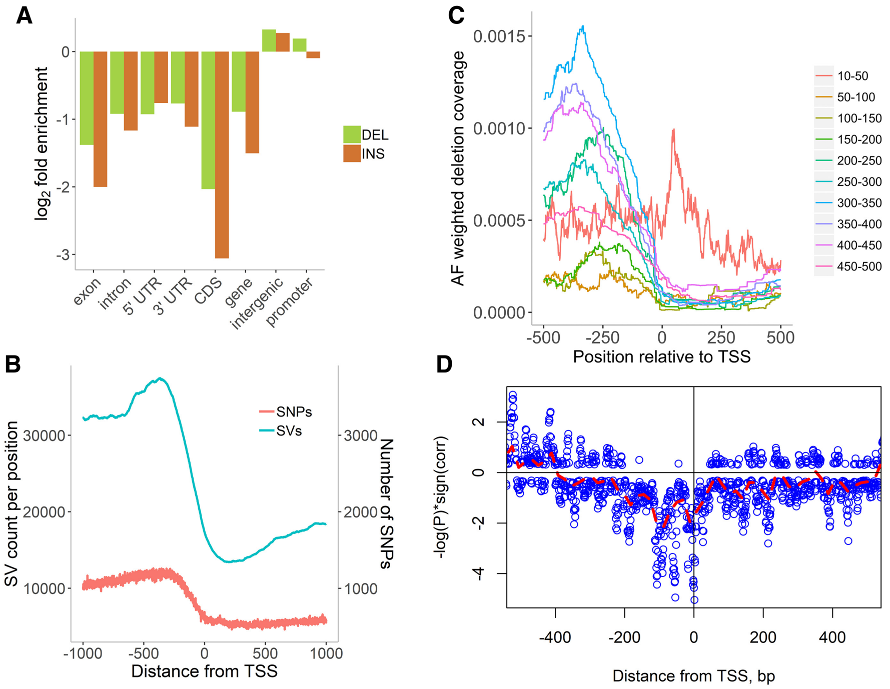

Figure 3. SVs in genome features. (*A*) Enrichment/depletion of deletions (green) and insertions (orange) in various genomic regions. As expected, genic regions have fewer SVs than intergenic ones, with CDSs and exons being the most conserved regions. (*B*) Distribution of deletion and insertion clusters near the transcription start site (TSS). Although the total number of SNPs is much larger than SV clusters, SVs affect more positions. The bump at about −366 bp just before the core promoter is explained by longer SVs associated with transposons. (*C*) Distribution of the number of deletions in the vicinities of start and end of transcription and translation (Supplemental Fig. S16). (*D*) *P*-values of the independence tests between predicted TFBS and deletions. Strong anti-correlation is observed at the TSS and ∼100 bp upstream. Distribution of *P*-values shows that in the core promoter area ([TSS-200, TSS]), deletions and TFBS are not independent.

Known SVs at important loci

To test whether our pipeline detected known structural variants in rice, we examined a set of selected genes with known structural variants. An important gene in rice known as *GW5* was found to be associated with rice grain width and weight (Shomura et al. 2008). A study revealed that a deletion in *qSW5*, a QTL for seed width which contains *GW5*, has played an important role in increased yield during rice domestication. Only 390 bp of *GW5* can be mapped to Chromosome 5 of Nipponbare, with a larger part of the gene overlapping a 1212-bp deletion in the Nipponbare genome. Analysis of our insertion data set revealed 17 samples that contain the corresponding insertion of 1212 bp. Using historical phenotyping data we were able to confirm its association with grain weight, although the *P*-value appeared to be barely

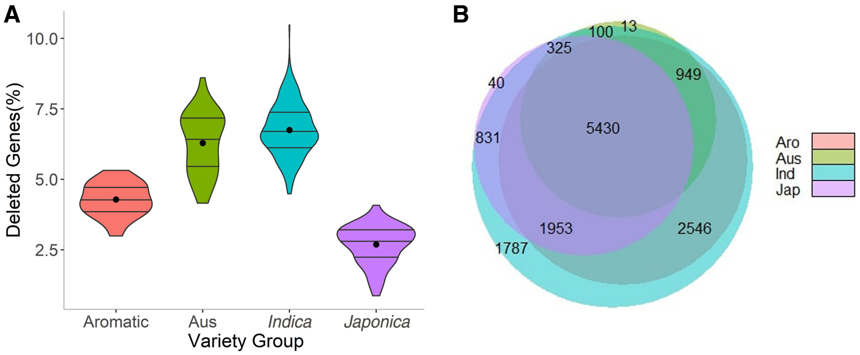

Figure 4. Deleted genes in variety groups. (*A*) Percentage of deleted genes in each variety group. (*B*) Number of deleted genes (frequency ≥ 5) that are unique or shared between variety groups. Note that the number of the deleted genes in *Japonica* is lower can be explained by the bias introduced by using Nipponbare genome as a reference.

significant due to sample size (Supplemental Figure S9a).

Many studies have associated copy number variants with changes in gene expression levels and various adaptive traits. A study of the *GL7* locus (Wang et al. 2015) revealed that a 17.1-kb tandem duplication is responsible for a long-grain phenotype in selected rice varieties. We identified 111 varieties in the 3K RG data set with this causative duplication. Historical data confirmed the association with grain length (Supplemental Fig. S9b).

A study on anaerobic germination previously discovered a 20.9-kb deletion of the *AG1* locus in several varieties included in the 3K RG (Kretzschmar et al. 2015) data set. In our predicted SV data set, we found 156 samples with this deletion, 39 of which had an additional deletion of ∼100 bp within the *AG1* locus and 149 belong to the *Indica* group. In addition, we found 176 samples with a novel smaller deletion (185–467 bp) at the locus instead of the longer variant previously reported. Thus, our study confirmed the presence of known SVs in important genomic regions and expanded the set of genotypes and haplotypes with novel SVs to examine in further association studies (Supplemental Fig. S9; Supplemental Data sheet “Known SVs in Genes”).

Utility of SV data sets for GWAS

SVs have been recognized as the causative mutations for many traits. Thus, the ability to conduct association studies with SVs should aid in gene discovery. Although SV detection is known to have a higher error rate than SNP calling, we theorize that with enough coverage, our predicted SVs can be used in genome-wide association studies (GWAS). As an example, we conducted a GWAS for a seed coat color trait using a previously published SNP data set, merged with genotype data for an insertion site (Chr 07: 6068071) at the red pericarp (*Rc*; LOC_Os07g11020) gene locus, known to be the causative mutation for this trait (Sweeney et al. 2006). Supplemental Figure S10a shows that the insertion is the most significant point in the GWAS plot at the *Rc* locus.

We also tested for SV effects associated with grain length. The major peak (Supplemental Fig. S10b) coincides with the *LONG KERNEL 3* gene, which is known to regulate grain size (Takano-Kai et al. 2009). The 350-nt-long deletion within the peak truncated the longest CDS of the gene and is likely to be the causative variation. The other peak on Chromosome 11 has an indel as the most significant variation. This region does not have annotated protein-coding genes and is more difficult to interpret.

Knowing that the presence of SVs is likely to have a significant impact on a gene function, we also identified 710 functionally characterized genes in the Q-TARO database (Yonemaru et al. 2010) that intersect with deletions (MAF > 0.005) (Supplemental Data, sheet “Q-TARO Genes”), among these, 308 are completely deleted in some rice accessions.

Deleted genes

We defined genes as deleted when their coding sequence is deleted over its entire length in at least one sample out of 562 high-coverage samples. Figure 4A shows the fraction of deleted genes in each sample, and Figure 4B illustrates the distributions of deleted genes among the variety groups. We used GO annotations and MAPMAN curated pathways (Thimm et al. 2004) to compare enriched biological themes of deleted temperate *Japonica* (Nipponbare) genes with the core gene set of genes that are present in all varietal groups. Results showed an enrichment of a core set genes having basal housekeeping functions and processes (e.g., developmental and catabolic processes, DNA-binding transcription factor activity, etc.). For *circum-Aus* and *Indica*-deleted genes, biological themes were indicative of adaptive functions/processes (e.g., response to oxidative stress, abiotic stress pathways, defense response), hinting at domestication/selection events that these variety groups underwent through their history of cultivation in diverse environments (Supplemental Data, sheet “Deleted Genes”). Deleted genes in the *circum-Basmati* (aromatic) group did not show any overrepresented adaptive themes, but were enriched for housekeeping functions (cell growth, carbohydrate metabolism), somewhat supporting the known cultivation history of *circum-Basmati* varieties in a smaller geographic region as compared to the *Indica* and *circum-Aus* varietal groups.

Discussion

We present the results of one of the largest studies on structural genome variations within a crop species. We carefully assessed the performance of different methods in discovering SVs with simulated short reads. We then combined them into one pipeline and assembled a comprehensive data set of SVs produced from the 3000 rice genomes project data set.

We found ∼1.5 million SV events (clusters) longer than 9 nt that are distributed across the Nipponbare reference genome. In the manual validation, we found complex events that “confuse” the SV callers, such as short deletions in large interspersed duplications, palindromes, and terminal repeats. It is also important to note that some events may be found in large translocated regions, which could not be differentiated by the pipeline due to the limitation of detecting translocations and events contained within them. Furthermore, our deletion detection has a lower false positive rate compared to the other SV types. Although transposable elements complicate SV detection, true TE events were accurately predicted by the pipeline (see validation results).

At least 17% of the SV calls longer than 50 bp are associated with transposable elements, which contribute significantly to genomic variation in plants (Wendel et al. 2016). Based on their distribution near genes, we hypothesize that transposons play an extensive role in gene regulation, which is consistent with other studies (Naito et al. 2009; Han et al. 2013; Shen et al. 2017). Population structure, revealed by the deletion (Supplemental Fig. S4c) and CNV data sets, identified the same subgroups as genome-wide SNP analysis (Wang et al. 2018), providing additional validation of the identified SVs.

We observed a significant difference between distributions of short and long indels near TSSs. Longer deletions most frequently occurred in promoter regions, whereas short deletions were preferentially found in or near 5′ UTRs. The peak of long deletions in promoter regions at ~360 bp upstream of the TSS is consistent with previous observations made for transposons (Han et al. 2013) and may be explained by easier accessibility of these regions for transposon insertion (Naito et al. 2009). The short deletions peak in 5′ UTRs can be explained by lower complexity of 5′ UTRs having numerous short sequence repeats. The greatest contribution to this peak was from deletions with lengths divisible by three, which is consistent with the SSR length distribution. We also detected a significant anti-correlation between the presence of SVs and TFBSs at ~100 bp upstream of TSS, implying negative selection against deletion of important regulatory elements in these regions.

The abundance of SVs that have high sequence similarity to known transposable elements suggests that many SVs are products of TE activity. The higher number of TEs in the upstream regions of genes (Naito et al. 2009; Han et al. 2013), where promoters and regulatory motifs reside, indicates that SVs may be important agents for gene expression pleiotropy that is often observed in stress responsive genes. Studies in both human and plant genomes have found structural variants and transposable elements that are associated with aberrant expression of nearby genes (Lu et al. 2012; Wei and Cao 2016; Chiang et al. 2017). Previous studies in maize (Lu et al. 2015), cucumber (Zhang et al. 2015), soybean (McHale et al. 2012), *A. thaliana* (Debolt 2010), and 50 rice accessions (Xu et al. 2012) also associated high level of SVs in proximal locations to stress response or disease defense genes. Makarevitch et al. (2015) reported that small numbers of maize TE families may contribute to abiotic stress responses by providing stress responsive enhancer-like functions to nearby genes. They also reported that specific insertions of TEs near genes are often polymorphic within a species, in agreement with our observations across the 3K RG.

Although third-generation sequencing technologies can assemble high-quality rice genomes and assess structural variation through comparative genomics, it is unlikely that for the foreseeable future they will be applied to a large set of varieties within a species. Hence, bioinformatic analysis of short reads is currently the most practical way to assess the diversity of structural variation within a species as performed in this study. Future studies may use different reference genomes from other variety groups, and inclusion of new samples resequenced at higher depths would allow better profiling of longer insertions. Our SV data set will enable rice geneticists to explore variability that is normally missing in SNP-based genome-wide association studies. Moreover, the variability described in this analysis can be used as a hypothesis generator to identify genetic causes of different important traits through future functional studies.

Methods

Evaluation of SV callers

We benchmarked a set of SV callers to identify a subset to integrate into a discovery pipeline. The benchmarking pipeline was designed so that the performance of SV callers could be evaluated with respect to variant types—deletion (DEL), insertion (INS), inversion (INV), tandem duplication (DUP), and translocation—and variant sizes, binned according to lengths: A (50–150 bp), B (151–500 bp), C (500–5000 bp), D (5–50 kb), E (50–250 kb), and F (0.25–1 Mb). Test data were designed to replicate the 14× average coverage of the 3K RG data set with 1000 introduced variations per SV type for the first four bins, and 200 and 100 variations for bins E and F, respectively. Genomic sequences with DEL, DUP, and INV variants were created by SVSIM (https://github.com/mfranberg/svsim) using the Nipponbare RefSeq (IRGSP 1.0), and sequencing reads were simulated by WGSIM (https://github.com/lh3/wgsim) with 83-bp read lengths, 500-bp insert sizes (SD = 50), and 0.02 error rates. It is worth clarifying that all DUP events simulated by SVSIM were tandem duplications (TDs). For insertion types, simulated paired-end reads of the Nipponbare RefSeq were aligned to another reference genome with randomly deleted regions. For translocation types, random regions in the Nipponbare RefSeq were deleted and inserted into regions either in the same or another chromosome. The Burrows–Wheeler Aligner's (BWA) (Li and Durbin 2010) paired-end module was used in mapping reads to a reference.

A prediction of an SV caller was considered correct if it passed 90% minimum reciprocal overlap (RO) and its breakpoint error (*e*), defined as the sum of distances between the breakpoint starts and ends of the predicted SV(*p*) and the simulated event (*S*), was less than 10 bp (allowing for microhomologies around breakpoint sites) (Schröder et al. 2014) or <10% of *length(p)+length(S)* (requiring less error for events with size <50 bp). However, these conditions were too strict to detect duplications that have poor breakpoint resolutions. In Supplemental Figure S11, the sensitivity of the callers on different ROs suggested a 70% threshold and no constraint for *e* to evaluate fairly duplication breakpoints.

Discovery pipeline

Based on benchmarking results, Pindel (Ye et al. 2009) was selected as the main variant caller for the pipeline since it consistently called more precise predictions across almost all bins of deletions, tandem duplications, and inversions even though it had relatively lower sensitivity for the largest events. DELLY (Rausch et al. 2012), GROM (Smith et al. 2017a), and LUMPY (Layer et al. 2014) were added to improve sensitivity and support of predictions especially for larger variants. For insertions, both MetaSV (Mohiyuddin et al. 2015) and MindTheGap (Rizk et al. 2014) were chosen to complement each other for better sensitivity for short and long insertions in the 562 highest-coverage samples since assembly-based algorithms require high coverages for accurate prediction. Interspersed duplications can only be detected using read depth signals; therefore, we analyzed copy number variation (CNV) predictions from NGSEP (Duitama et al. 2014) as a separate data set.

Given the calls predicted in each sample, merging results from all callers required a minimum reciprocal overlap (RO) to classify whether or not calls were similar. A common merging strategy is to require each variant prediction to be supported by at least two callers; however, this may increase false negatives when some selected callers perform poorly for some size ranges. To address this, our pipeline retained

1. all Pindel results with QUAL = PASS, regardless of other callers' support; for inversion, all Lumpy calls; and
2. results from other callers, not supported by Pindel, when supported by at least two callers, using the following criteria: If merging of callers is between Lumpy and another caller, 70% RO was required; otherwise, 90% RO.

These criteria were applied to all SV types except insertions that require a different clustering approach.

Supplemental Figure S12 shows the breakpoint accuracy of the callers based on sensitivity improvement using either 70% or 90% RO. The 90% threshold for reciprocal overlap was reduced to 70% to consider inaccuracy of duplication breakpoints from Lumpy. To merge insertion calls from MetaSV and MindTheGap, the pipeline clustered all INS sites that were within 10 bp of one another and included all unique insertions assembled by the callers.

To create a map of variant sites across all samples, we clustered variant calls across different samples and pooled them as follows. All events that overlapped by at least 1 bp were initially grouped together. For each group, a graph was built with SVs as nodes with edges connecting SVs that have at least 90% reciprocal overlap, or at least 70% RO with a *breakpoint error* of at most 10 bp. The latter condition allows for grouping of small events from different samples that have <90% RO but very small boundary differences. Each group was then split into connected components of the graph. For each connected component, we computed a distance matrix using the absolute value of the distances between breakpoints of two variants divided by their total lengths. Then, hierarchical clustering by complete linkage was performed using the distance matrix with a cutoff of 0.1 for height, yielding the final clusters. We consider each cluster to represent a single ancestral event inherited by a subset of our sample. This stage is done for each variant type except insertions, which were clustered by grouping events that are at most 10 bp apart.

The pipeline described above uses SV callers selected for their performance on detecting insertions, deletions, inversions, and tandem duplications. However, the use of read depth signals allows for the discovery of larger copy number variants, including interspersed duplications and those variants located in complex genomic regions that RP and SR methods find difficult to detect (Medvedev et al. 2009; Zhao et al. 2013). NGSEP, one of the callers using RD signals, was used to compile a CNV data set (Supplemental Table S3).

Validation

With the published "N 22::IRGC 19379-1" hereafter referred to as N 22 (NCBI Assembly ASM195236v1) (Stein et al. 2018) and Nipponbare reference IRGSP 1.0 (Kawahara et al. 2013), we validated random SVs predicted in CX368, an N 22 accession in the 3K RG data set. Random SVs were selected and manually inspected in a dot plot alignment between N 22 and Nipponbare generated using Gepard (Krumsiek et al. 2007). Events found in long deleted regions were further analyzed using NCBI BLAST to determine if they occur in translocated regions in the Nipponbare reference. False positive rate was computed per SV type depending on the number of predicted calls that were inconsistent with the dot plots. Some false positive calls may also be private to CX368 (a different N 22 sample) and not to the N 22 reference genome. Because we mainly expect to discover tandem duplications using the pipeline, interspersed duplications were given lower weights of being true positives. For computing false negative rates, we focused on deletions and identified 20 events between N 22 and Nipponbare using the dot plots in randomly selected locations and validated if they were predicted in CX368.

The scripts used for the structural variant discovery pipeline are available at https://github.com/rrfuentes/SV_Discovery as well as in Supplemental Code.

Identification and analysis of transcription factor binding sites

We extracted regulatory regions [TSS-500, TSS + 500] for all "high confidence" rice genes defined by Tatarinova et al. (2016). The distribution of transcription factor binding sites (TFBSs) in these regulatory regions were analyzed with the MATCH algorithm (Kel et al. 2003) using the TRANSFAC database (Wingender et al. 1996) comprised of 764 plant position weight matrices. For each genomic position, we calculated the fraction of genes that have a TBFS in this position and computed a probability that each position x is covered by a putative regulatory element, $P_x(TFBS)$. Then for each position in this region, we calculated the fraction of all genes that have a structural variant and TBFS covering the same position, resulting in the probability $P_x(TFBS \cap SV)$. The conditional probability was calculated as follows:

$$P_x(SV|TFBS) = \frac{P_x(TBFS \cap SV)}{P_x(TFBS)}.$$

Sequence complexity

To rule out the influence of sequence complexity on the efficiency of read mapping and variant calling, we calculated sequence complexity profiles around transcription start sites. For every sequence around a transcription start site [TSS − 1000, TSS + 1000], we calculated the Linguistic Complexity (CL)

$$CL = \left(\sum\nolimits_{i=1}^{N} V_i\right) / \left(\sum\nolimits_{i=1}^{N} V_{maxi}\right),$$

where N is the window size (10 bp in our case), V_i is the number of words of size i in the window, and V_{maxi} is the maximum possible number of words of length i. For a window of size N, and alphabet size K, this number is calculated according to the following formula: $V_{maxi} = \min(K^i, N - i + 1)$ (Orlov and Potapov 2004). CL is the ratio of the observed number of different words of size 1, …, N in a given window divided by the sum of maximum possible number of different words for a fixed window length. After calculating the complexity profile for every sequence, we averaged across all rice promoters.

Deleted genes

For each sample, all non-TE genes that were completely deleted were identified and their sequences retrieved. These sequences were compared against all insertion sequences to remove all of those that may have been translocated or have similar copies in other parts of the genome. Using the variety group assignment from Wang et al. (2018), we computed the average number of genes deleted per variety group. This analysis focused on genes present in the Nipponbare genome.

Gene enrichment analysis

To identify overrepresented (enriched) biological themes of genes covered by large deletions, we compared the list of these subset genes against the "population" of genes in the genome that is annotated within a given system of classifying genes (e.g., "Biological Process" in Gene Ontology). "Hits" refers to genes in the population falling within the gene category in question. As an example, "Population hits" for the GO annotation Biological Process "cellular stress response to acidic pH" refers to the number of genes falling within the category "cellular stress response to acidic pH" out of all genes in the population annotated with a Biological Process. Given the number of genes in the subset SV gene list that fall within a specific category (the "List hits"), the count of genes in the list (the "List total") and the corresponding "Population Hits" and "Population Total," the probability of seeing the number of "List Hits" in the "List Total" given the frequency of "Population Hits" in the "Population Total" is calculated with the Fisher's exact test and reported for overrepresentation analysis.

Genome-wide association studies

We conducted GWAS on the seed coat color and grain length phenotypes, recorded by the T.T. Chang Genetic Resources Center for the International Rice Genebank Collection information system at IRRI and retrieved from SNP-Seek, on a 365 high-sequence coverage subset of the 3K RG where the phenotype was measured: ("white") 291 samples; ("red") 74 samples. To construct the genotype file, we first merged two data sets (GATK SNP and small indels and the INS data set). Then, we merged this single-variant data set with a filtered and LD-pruned SNP data set ($MAF > 0.015$, max $miss = 0.2$, $r^2 \leq 0.8$ within 2 kb, total 889,903 SNPs).

We used linear mixed model association analysis implemented in GEMMA (Zhou and Stephens 2012), using a kinship matrix and the first five principal components for relatedness and population structure correction. The kinship matrix was computed by the GEMMA -gk command with default parameters. The PCA was computed using PLINK 1.9 (Purcell et al. 2007). We plotted Manhattan and QQ plots using the R package "qqman" (Turner 2014) with in-house modifications for graphics.

Data access

All SVs identified in this study are available at the SNP-Seek portal (http://snp-seek.irri.org) in the download section.

Acknowledgments

Work in the Grigoriev laboratory is supported by awards from the National Science Foundation (DBI-1458202, ACI-1548562), National Institute of Health (R15CA220059), and New Jersey Health Foundation (PC77-17). We thank XSEDE (MCB150014), PRAGMA, Universidad de los Andes, and the Department of Science and Technology, Advanced Science and Technology Institute of the Philippines for computing resources. We thank the Bill and Melinda Gates Foundation for support to the 3K RG via the GSR Phase 2 award OPPGD1393.

Author contributions: N.A. and A.G. conceived the study. R.R.F., D.C., J.D., T.T., A.G., and N.A. performed data analyses. J.D., J.F.D., S.S., A.G., and M.M. contributed improved software for SV detection. R.R.F., D.C., J.D., T.T., A.G., R.A.W., K.L.M., R.M., and N.A. interpreted results and wrote the manuscript.

References

The 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166:** 481–491. doi:10.1016/j.cell.2016.05.063

The 3000 rice genomes project. 2014. The 3,000 rice genomes project. *GigaScience* **3:** 7. doi:10.1186/2047-217X-3-7

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21:** 974–984. doi:10.1101/gr.114876.110

Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, Ulat VJ, Chebotarov D, Zhang G, Li Z, et al. 2014. SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res* **63:** 2–6. doi:10.1093/nar/gku1039

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12:** 363–376. doi:10.1038/nrg2958

Ashikari M, Wu J, Yano M, Sasaki T, Yoshimura A. 1999. Rice gibberellin-insensitive dwarf mutant gene *Dwarf 1* encodes the α-subunit of GTP-binding protein. *Proc Natl Acad Sci* **96:** 10284–10289. doi:10.1073/pnas.96.18.10284

Bai Z, Chen J, Liao Y, Wang M, Liu R, Ge S, Wing RA, Chen M. 2016. The impact and origin of copy number variations in the *Oryza* species. *BMC Genomics* **17:** 261. doi:10.1186/s12864-016-2589-2

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37:** 202–208. doi:10.1093/nar/gkp335

Baker M. 2012. Structural variation: the genome's hidden architecture. *Nat Methods* **9:** 133–137. doi:10.1038/nmeth.1858

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high resolution mapping of genomic structural variation. *Nat Methods* **6:** 677–681. doi:10.1038/nmeth.1363

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L; GTEx Consortium, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet* **49:** 692–699. doi:10.1038/ng.3834

Cook DL, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, et al. 2012. Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* **338:** 1206–1209. doi:10.1126/science.1228746

Copetti D, Zhang J, Baidouri ME, Gao D, Wang J, Barghini E, Cossu RM, Angelova A, Maldonado L CE, Roffler S, et al. 2015. RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics* **16:** 538. doi:10.1186/s12864-015-1762-3

Debolt S. 2010. Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* **2:** 441–453. doi:10.1093/gbe/evq033

Duitama J, Quintero JC, Cruz DF, Quintero C, Hubmann G, Foulquié-Moreno MR, Verstrepen KJ, Thevelein JM, Tohme J. 2014. An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. *Nucleic Acids Res* **42:** e44. doi:10.1093/nar/gkt1381

Duitama J, Silva A, Sanabria Y, Cruz DF, Quintero C, Ballen C, Lorieux M, Scheffler B, Farmer A, Torres E, et al. 2015. Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. *PLoS One* **10:** e0124617. doi:10.1371/journal.pone.0124617

Escaramís G, Docampo E, Rabionet R. 2015. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics* **14:** 305–314. doi:10.1093/bfgp/elv014

Francia E, Pecchioni N, Policriti A, Scalabrin S. 2015. CNV and structural variation in plants: prospects of NGS approaches. In *Advances in the understanding of biological sciences using next generation sequencing (NGS) approaches* (ed. Sablok G, et al.), pp. 211–232. Springer International Publishing, Cham, Switzerland.

Fu YX. 1995. Statistical properties of segregating sites. *Theor Popul Biol* **48:** 172–197. doi:10.1006/tpbi.1995.1025

Han Y, Qin S, Wessler SR. 2013. Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC Genomics* **14:** 71. doi:10.1186/1471-2164-14-71

Hurwitz BL, Kudrna D, Yu Y, Sebastian A, Zuccolo A, Jackson SA, Ware D, Wing RA, Stein L. 2010. Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus *Oryza*. *Plant J* **63:** 990–1003. doi:10.1111/j.1365-313X.2010.04293.x

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR. 2003. An active DNA transposon family in rice. *Nature* **421:** 163–167. doi:10.1038/nature01214

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu JZ, Zhou SG, et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6:** 4. doi:10.1186/1939-8433-6-4

Kel AE, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003. MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* **31:** 3576–3579. doi:10.1093/nar/gkg585

Kretzschmar T, Pelayo MA, Trijatmiko KR, Gabunada LF, Alam R, Jimenez R, Mendioro MS, Slamet-Loedin IH, Sreenivasulu N, Bailey-Serres J, et al. 2015. A trehalose-6-phosphate phosphatase enhances anaerobic germination tolerance in rice. *Nat Plants* **1:** 15124. doi:10.1038/nplants.2015.124

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23:** 1026–1028. doi:10.1093/bioinformatics/btm039

Lam HY, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28:** 47–55. doi:10.1038/nbt.1600

Lawson MJ, Zhang L. 2006. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol* **7:** R14. doi:10.1186/gb-2006-7-2-r14

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15:** R84. doi:10.1186/gb-2014-15-6-r84

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26:** 589–595. doi:10.1093/bioinformatics/btp698

Li Y, Xiao J, Wu J, Duan J, Liu Y, Ye X, Zhang X, Guo X, Gu Y, Zhang L, et al. 2012. A tandem segmental duplication (TSD) in green revolution gene *Rht-D1b* region underlies plant height variation. *New Phytol* **196:** 282–291. doi:10.1111/j.1469-8137.2012.04243.x

Lu C, Chen J, Zhang Y, Hu Q, Su W, Kuang H. 2012. Miniature inverted–repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. *Mol Biol Evol* **29:** 1005–1017. doi:10.1093/molbev/msr282

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, et al. 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun* **6:** 6914. doi:10.1038/ncomms7914

Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, Springer NM. 2015. Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet* **11:** e1004915. doi:10.1371/journal.pgen.1004915

Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, Sanciangco M, Palis K, Copetti D, Poliakov A, et al. 2017. Rice SNP-seek database update: new SNPs, indels, and queries. *Nucleic Acids Res* **45:** D1075–D1081. doi:10.1093/nar/gkw1135

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM. 2012. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol* **159:** 1295–1308. doi:10.1104/pp.112.194605

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. doi:10.1101/gr.107524.110

Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* **6:** 13–20. doi:10.1038/nmeth.1374

Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, Lam HYK. 2015. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31:** 2741–2744. doi:10.1093/bioinformatics/btv204

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461:** 1130–1134. doi:10.1038/nature08479

Naito K, Monden Y, Yasuda K, Saito H, Okumoto Y. 2014. *mPing*: the bursting transposon. *Breed Sci* **64:** 109–114. doi:10.1270/jsbbs.64.109

Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC. 2014. Accurate *de novo* and transmitted indel detection in exome-capture data using microassembly. *Nat Methods* **11:** 1033–1036. doi:10.1038/nmeth.3069

Ohtsubo H, Kumekawa N, Ohtsubo E. 1999. *RIRE2*, a novel *gypsy*-type retrotransposon from rice. *Genes Genet Syst* **74:** 83–91. doi:10.1266/ggs.74.83

Orlov YL, Potapov VN. 2004. Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res* **32:** W628–W633. doi:10.1093/nar/gkh466

Ouyang S, Buell CR. 2004. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res* **32:** D360–D363. doi:10.1093/nar/gkh099

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

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575. doi:10.1086/519795

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20:** 623–635. doi:10.1101/gr.102970.109

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28:** i333–i339. doi:10.1093/bioinformatics/bts378

Rizk G, Gouin A, Chikhi R, Lemaitre C. 2014. MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics* **30:** 3451–3457. doi:10.1093/bioinformatics/btu545

Saxena RK, Edwards D, Varshney RK. 2014. Structural variations in plant genomes. *Brief Funct Genomics* **13:** 296–307. doi:10.1093/bfgp/elu016

Schatz MC, Maron LG, Stein JC, Wences AH, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, et al. 2014. Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Genome Biol* **15:** 506. doi:10.1186/s13059-014-0506-z

Schröder J, Hsu A, Boyle SE, Macintyre G, Cmero M, Tothill RW, Johnstone RW, Shackleton M, Papenfuss AT. 2014. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics* **30:** 1064–1072. doi:10.1093/bioinformatics/btt767

Seck PA, Diagne A, Mohanty S, Wopereis MCS. 2012. Crops that feed the world 7: rice. *Food Security* **4:** 7–24. doi:10.1007/s12571-012-0168-1

Shen J, Liu J, Xie K, Xing F, Xiong F, Xiao J, Li X, Xiong L. 2017. Translational repression by a miniature inverted-repeat transposable element in the 3′ untranslated region. *Nat Commun* **8:** 14651–14651. doi:10.1038/ncomms14651

Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S, Yano M. 2008. Deletion in a gene associated with grain size increased yields during rice domestication. *Nat Genet* **40:** 1023–1028. doi:10.1038/ng.169

Sindi S, Helman E, Bashir A, Raphael BJ. 2009. A geometric approach for classification and comparison of structural variants. *Bioinformatics* **25:** 222–230. doi:10.1093/bioinformatics/btp208

Smith SD, Kawash JK, Grigoriev A. 2015. GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. *PeerJ* **3:** e836. doi:10.7717/peerj.836

Smith SD, Kawash JK, Grigoriev A. 2017a. Lightning-fast genome variant detection with GROM. *GigaScience* **6:** 1–7. doi:10.1093/gigascience/gix091

Smith SD, Kawash JK, Karaiskos S, Biluck I, Grigoriev A. 2017b. Evolutionary adaptation revealed by comparative genome analysis of woolly mammoths and elephants. *DNA Res* **24:** 359–369. doi:10.1093/dnares/dsx007

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, et al. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat Genet* **50:** 285–296. doi:10.1038/s41588-018-0040-0

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz HM, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394

Sutton T, Baumann U, Hayes J, Collins NC, Shi BJ, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, et al. 2007. Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* **318:** 1446–1449. doi:10.1126/science.1146853

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res* **20:** 1689–1699. doi:10.1101/gr.109165.110

Sweeney MT, Thomson MJ, Pfeil BE, McCouch S. 2006. Caught red-handed: *Rc* encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell* **18:** 283–294. doi:10.1105/tpc.105.038430

Takano-Kai N, Jiang H, Kubo T, Sweeney M, Matsumoto T, Kanamori H, Padhukasahasram B, Bustamante C, Yoshimura A, Doi K, et al. 2009. Evolutionary history of *GS3*, a gene conferring grain length in rice. *Genetics* **182:** 1323–1334. doi:10.1534/genetics.109.103002

Tatarinova TV, Chekalin E, Nikolsky Y, Bruskin S, Chebotarov D, McNally KL, Alexandrov N. 2016. Nucleotide diversity analysis highlights functionally important genomic regions. *Sci Rep* **6:** 35730–35730. doi:10.1038/srep35730

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* **11:** 1441–1452. doi:10.1101/gr.184001

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M. 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* **37:** 914–939. doi:10.1111/j.1365-313X.2004.02016.x

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z. 2017. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res* **45:** W122–W129. doi:10.1093/nar/gkx382

Triska M, Solovyev V, Baranova A, Kel A, Tatarinova TV. 2017. Nucleotide patterns aiding in prediction of eukaryotic promoters. *PLoS One* **12:** e0187243. doi:10.1371/journal.pone.0187243

Turner SD. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J Open Source Software* **3:** 731. doi:10.21105/joss.00731

Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X, Zeng R, Zhu H, Dong G, Qian Q, et al. 2012. Control of grain size, shape and quality by *OsSPL16* in rice. *Nat Genet* **44:** 950–954. doi:10.1038/ng.2327

Wang Y, Xiong G, Hu J, Jiang L, Yu H, Xu J, Fang Y, Zeng L, Xu E, Xu J, et al. 2015. Copy number variation at the *GL7* locus contributes to grain size diversity in rice. *Nat Genet* **47:** 944–948. doi:10.1038/ng.3346

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557:** 43–49. doi:10.1038/s41586-018-0063-9

Wei L, Cao X. 2016. The effect of transposable elements on phenotypic variation: insights from plants to humans. *Sci China Life Sci* **59:** 24–37. doi:10.1007/s11427-015-4993-2

Wendel JF, Jackson SA, Meyers BC, Wing RA. 2016. Evolution of plant genome architecture. *Genome Biol* **17:** 37. doi:10.1186/s13059-016-0908-1

Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24:** 238–241. doi:10.1093/nar/24.1.238

Würschum T, Boeven PH, Langer SM, Longin CF, Leiser WL. 2015. Multiply to conquer: copy number variations at *Ppd-B1* and *Vrn-A1* facilitate global adaptation in wheat. *BMC Genet* **16:** 96. doi:10.1186/s12863-015-0258-0

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, et al. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* **30:** 105–111. doi:10.1038/nbt.2050

Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L, et al. 2013. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153:** 919–929. doi:10.1016/j.cell.2013.04.010

Yang R, Nelson AC, Henzler C, Thyagarajan B, KaT S. 2015. ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and *de novo* assembly. *Genome Med* **7:** 127–127. doi:10.1186/s13073-015-0251-2

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25:** 2865–2871. doi:10.1093/bioinformatics/btp394

Yi K, Ju YS. 2018. Patterns and mechanisms of structural variations in human cancer. *Exp Mol Med* **50:** 98. doi:10.1038/s12276-018-0112-3

Yonemaru JI, Yamamoto T, Fukuoka S, Uga Y, Hori K, Yano M. 2010. Q-TARO: QTL annotation rice online database. *Rice* **3:** 194–203. doi:10.1007/s12284-010-9041-z

Zhang Z, Mao L, Chen H, Bu F, Li G, Sun J, Li S, Sun H, Jiao C, Blakely R, et al. 2015. Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *Plant Cell* **27:** 1595–1604. doi:10.1105/tpc.114.135848

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14** (Suppl. 11)**:** S1. doi:10.1186/1471-2105-14-S11-S1

Zhou X, Stephens M. 2012. Genome-wide efficient mixed model analysis for association studies. *Nat Genet* **44:** 821–824. doi:10.1038/ng.2310

Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. 2014. Copy number polymorphism in plant genomes. *Theor Appl Genet* **127:** 1–18. doi:10.1007/s00122-013-2177-7

Received June 29, 2018; accepted in revised form March 11, 2019.

CSH PRESS GENOME RESEARCH

Structural variants in 3000 rice genomes

Roven Rommel Fuentes, Dmytro Chebotarov, Jorge Duitama, et al.

Genome Res. 2019 29: 870-880 originally published online April 16, 2019
Access the most recent version at doi:10.1101/gr.241240.118