Plant Communications
Research Article

CellPress
Partner Journal

Exploring transposable element-based markers to identify allelic variations underlying agronomic traits in rice

Haidong Yan[1], David C. Haak[1,2], Song Li[1,2], Linkai Huang[3] and Aureliano Bombarely[4,5,*]

[1]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061, USA

[2]Graduate Program in Genetics, Bioinformatics and Computational Biology (GBCB), Virginia Tech, Blacksburg, VA 24061, USA

[3]Department of Grassland Science, Animal Science and Technology College, Sichuan Agricultural University, Chengdu 611130, China

[4]Department of Bioscience, Universita degli Studi di Milano (UNIMI), 20133 Milano, Italy

[5]Instituto de Biologıa Molecular y Celular de Plantas (IBMCP), UPV-CSIC, 46022 Valencia, Spain

***Correspondence: Aureliano Bombarely (aureliano.bombarely@unimi.it)**

https://doi.org/10.1016/j.xplc.2021.100270

ABSTRACT

Transposable elements (TEs) are a major force in the production of new alleles during domestication; nevertheless, their use in association studies has been limited because of their complexity. We have developed a TE genotyping pipeline (TEmarker) and applied it to whole-genome genome-wide association study (GWAS) data from 176 *Oryza sativa* subsp. *japonica* accessions to identify genetic elements associated with specific agronomic traits. TE markers recovered a large proportion (69%) of single-nucleotide polymorphism (SNP)-based GWAS peaks, and these TE peaks retained ca. 25% of the SNPs. The use of TEs in GWASs may reduce false positives associated with linkage disequilibrium (LD) among SNP markers. A genome scan revealed positive selection on TEs associated with agronomic traits. We found several cases of insertion and deletion variants that potentially resulted from the direct action of TEs, including an allele of *LOC_Os11g08410* associated with plant height and panicle length traits. Together, these findings reveal the utility of TE markers for connecting genotype to phenotype and suggest a potential role for TEs in influencing phenotypic variations in rice that impact agronomic traits.

Key words: transposable element, marker, agronomic trait, rice, GWAS

Yan H., Haak D.C., Li S., Huang L., and Bombarely A. (2022). Exploring transposable element-based markers to identify allelic variations underlying agronomic traits in rice. Plant Comm. **3**, 100270.

INTRODUCTION

Taking advantage of phenotypic variations is an essential approach for improving the productivity of agricultural crops (Tilman et al., 2001). Most current breeding efforts worldwide focus on the identification of selectable genetic markers that are associated with key phenotypic variations (Bevan et al., 2017). Although these efforts have made great progress in identifying the genetic basis of crop traits and superior alleles for agronomically significant genes (Moose and Mumm, 2008; Xue et al., 2008; Gross and Olsen, 2010), our understanding of the genotype-phenotype connections in crops is still limited (Liu and Yan, 2019).

The development of high-throughput genotyping platforms for single-nucleotide polymorphisms (SNPs) has enabled the generation of novel methods for the detection of causal loci and the dissection of complex traits (De Wit et al., 2015). Genome-wide association studies (GWAS) were developed to take advantage of the revolution in marker density afforded by SNPs and high-throughput sequencing. Nonetheless, the very high density of SNP markers also leads to some challenges, such as the number of false-positive associations due to linkage between SNPs (Cantor et al., 2010). SNP-based GWAS may also fail to identify structural variants (SVs), which can explain as much or more phenotypic variation than SNPs (Zhou et al., 2019). Thus, recent efforts have focused on developing approaches that reduce SNP artifacts and on generating marker platforms that can complement the strengths of SNP-based pipelines (Braz et al., 2019).

Transposable elements (transposons; TEs) can induce mutations in gene sequences or alter gene regulation through *de novo*

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

insertion or by generating duplicate exons or genes through retrotransposition of RNA (Kiyosawa, 1972; Doebley et al., 1997; Hayashi and Yoshida, 2009; Studer et al., 2011). The majority of SVs are driven by TEs that are also an essential source of large-effect alleles (Lisch, 2013; Alonge et al., 2020; Dominguez et al., 2020; Della Coletta et al., 2021). Given the ubiquitous distribution of TEs that harbor rich genetic variation, Roy et al. (2015) proposed that TEs mined from genomes could be used as a source of genetic markers for molecular plant breeding. A recent study analyzed variations in transposon content and insertions in 83 maize inbred lines (Lai et al., 2017). In rice, the TRACKPOSON tool package was developed to identify retrotransposon TE insertion polymorphisms (TIPs) based on the 3,000 rice genome database (Carpentier et al., 2019; Akakpo et al., 2020). Another work leveraged whole-genome re-sequencing data to determine the contribution of TEs to tomato diversity (Dominguez et al., 2020). Although these efforts have associated numerous agronomic traits to particular TE insertions, TE-induced phenotypic variations in crops have not been fully clarified. Furthermore, the development of high-throughput approaches for identifying TE markers at a genomic scale has been challenging.

Here, we re-analyzed a set of whole-genome sequencing data (DRA004358) from 176 rice accessions (Yano et al., 2016). TEs were useful for detecting positive selection and phenotype associations from GWAS in this rice population. We also found that TEs may produce allelic diversity by cooperating with insertions/deletions that occurred in candidate gene body regions. Finally, we developed a comprehensive bioinformatic tool (TEmarker) for use with whole-genome shotgun sequencing data. This tool can combine results from different TIP calling tools, identify and remove potential false-positive TE insertions, and perform high-throughput genotyping. Collectively, these findings provide a framework for using TE data to advance molecular breeding efforts and provide novel evolutionary insights.

RESULTS

TE landscape

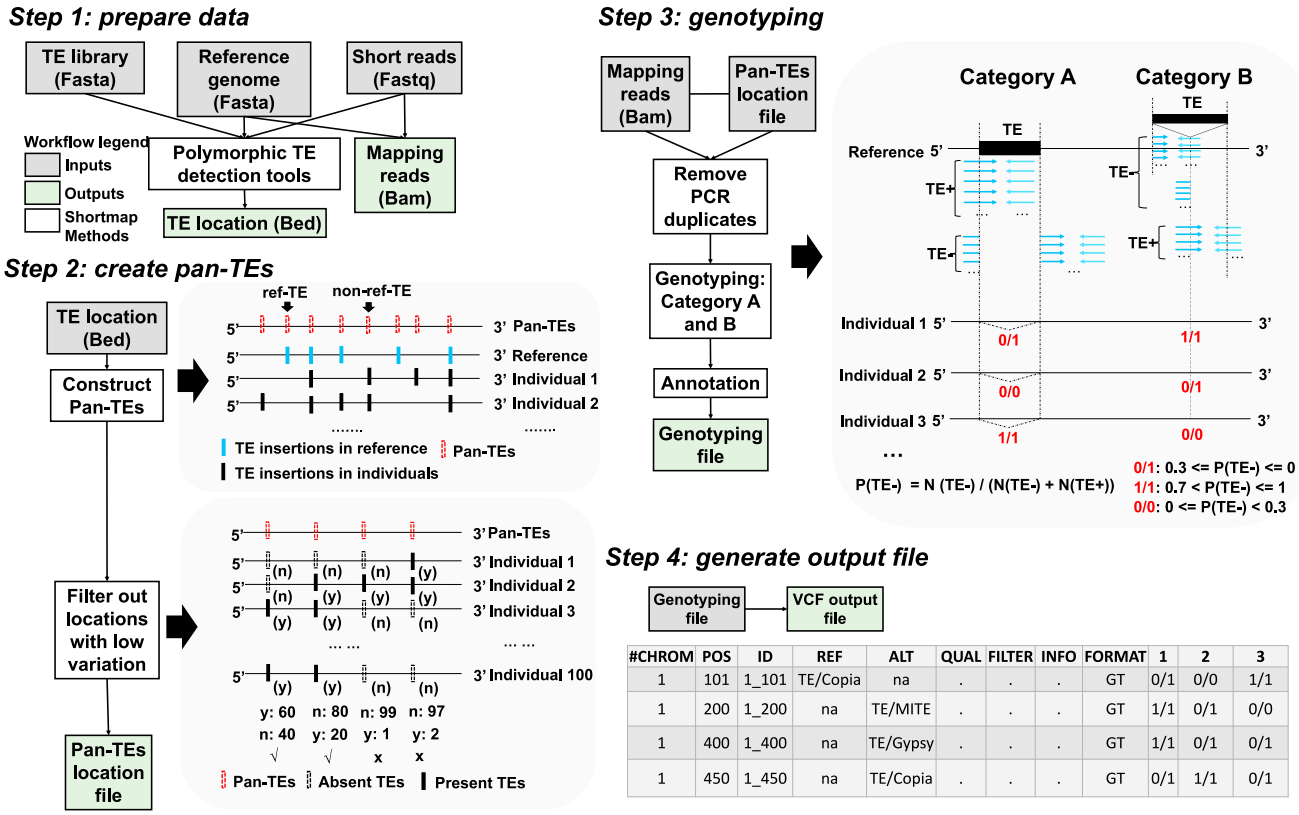
To identify transposon polymorphisms in the rice test dataset (Supplemental Tables 1 and 2) (Yano et al., 2016), we first developed a TEmarker pipeline based on TEs with genotype information across samples (Supplemental Note 1). This is achieved through four major steps (Figure 1). (1) TE location data are prepared using polymorphic TE detection tools wrapped in TEmarker. (2) The pan-TE insertions are constructed from the population using TE locations for each individual, and low/non-polymorphic locations are removed. (3) TE genotypes are generated based on the proportion of clipped reads that map to the border region of the TE insertion sites. The proportion is used to call heterozygous variants. This step also removes potential false-positive insertions where no clipped reads are found. (4) A variant call format (VCF) file is generated for downstream analyses (Figure 1; Supplemental Note 1: TEmarker pipeline).

To verify that TEmarker could accurately identify TE presence/absence polymorphisms, we simulated 350 TE insertion loci, 258 of which could be identified using six tools wrapped in McClintock (Nelson et al., 2017), as well as Jitterbug (Hénaff et al., 2015). TEmarker was used to perform genotyping across these loci and returned 252/258 (97.7%) of the simulated TE insertions (Supplemental Table 3). By comparison, the seven tools generated 790 false positives in total, predicting 787 insertion loci that were not part of the simulated set. TEmarker identified 787/790 (99.6%) of these sites as false positives (Supplemental Table 4). In general, TEmarker showed greater accuracy (AC), precision (PR), and specificity (SP) but marginally lower sensitivity (SE; 0.78 versus 0.81) than the six tools wrapped in McClintock (Supplemental Table 5). To test the ability of TEmarker to detect deletions, we simulated 400 TE deletion loci. TEmarker identified 374/400 (93.5%) of these deletion loci (Supplemental Table 6). To further check the performance of TEmarker, we analyzed one published dataset (PRJNA565484) with one assembled genome (IRGC132278) and associated Illumina short-read sequencing data (Zhou et al., 2020). We predicted 1,111 TE integration sites based on the short reads using TEmarker, and 97.30% (1,081/1,111) of them matched TE locations in the IRGC132278 genome (Supplemental Table 7). These results suggest that TEmarker achieves good performance for the reliable prediction of TE integration based on simulated and real datasets.

A total of 1,114,323 TEs were identified in the rice test dataset, including 953,247 non-polymorphic TEs and 372,247 polymorphic TEs. TEs for which no reads mapped to the region in more than 30% of individuals were removed, leaving 350,485 TEs. These TEs were further filtered by removing loci with a minor allele frequency (MAF) <0.05, and 6,073 TEs were retained (Supplemental Table 8). The average AF of the TEs was 0.188, and most had an AF between 0.05 and 0.1 (Figure 2A). A total of 426,337 SNPs in this population generated from Yano's study (Yano et al., 2016) were downloaded for comparison with the TE-derived markers. The average AF among these SNPs was 0.215, and their AF values showed a similar peak between 0.05 and 0.1; however, this peak exhibited a greater negative skew than did the TE peak (Figure 2A). Most TEs were ClassI_LTR (long terminal repeat) TEs (62.6%), and ClassII_DNA_MITE (MITE) TEs (19.4%) were twice as common as ClassII_DNA_not_MITE (nMITE) TEs (11.1%) (Figure 2B). More than half of the TE insertions (54.7%) were close to genes. Among the TE insertions located near genes, more than half were found in promoter regions, and TE insertions in untranslated (UTR) regions were the least-frequent category (Figure 2C). We compared linkage disequilibrium (LD) between the TEs and SNPs by checking how frequently the TEs were linked to nearby SNPs. We found that a higher number of TEs showed low LD with nearby SNPs, suggesting that they may represent genetic diversity not identified by SNPs (Figure 2D). In addition, TEs with higher MAFs were more frequently classified into the high LD group (Figure 2E), consistent with previous studies (Stuart et al., 2016; Yang et al., 2019).

SNP versus TE GWAS performance

To assess whether TEs are necessary for generating phenotypic variation, we performed GWAS for the 6,073 TEs and the 426,337 SNPs for seven traits (Yano et al., 2016) over 2 years (2013 and 2014) (Supplemental Figures 1–7). To define association peaks, the rice genome was binned into 50-kb regions, and all of the TE or SNP peaks within these regions were merged



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	1	2	3
1	101	1_101	TE/Copia	na	.	.	.	GT	0/1	0/0	1/1
1	200	1_200	na	TE/MITE	.	.	.	GT	1/1	0/1	0/0
1	400	1_400	na	TE/Gypsy	.	.	.	GT	1/1	0/1	0/1
1	450	1_450	na	TE/Copia	.	.	.	GT	0/1	1/1	0/1

Figure 1. Overview of TEmarker.

This pipeline contains four major steps. In step 1, users must provide whole-genome re-sequencing data, a reference genome, and a TE library as inputs. The TE locations of each individual are generated using polymorphic TE detection tools such as McClintock (Nelson et al., 2017) that are wrapped in TEmarker. In step 2, pan-TEs are built based on the collection of TE locations from all samples. Reference TEs and non-reference TEs are defined by the presence/absence of the TE in the reference genome. Pan-TEs with low variance or small numbers of individuals are removed. In step 3, mapping and Pan-TE files are merged to genotype. This step contains two major categories. For category A, when reads are clipped into two parts after mapping to a single location by a TE insertion in the reference genome, reads are labeled "TE–" when supporting the missing TE and are labeled "TE+" when supporting the existing TE. In category B, reads are labeled TE– when supporting a TE existing in one location without a TE insertion in the reference genome. After combining categories A and B, the clipped reads are blasted against the TE library to obtain annotations for the TEs from category B in order to obtain a final genotype file from all individuals. In step 4, a genotype information file in VCF format is generated for use in downstream analysis.

into a single peak. This resulted in the identification of 17 TE-derived (26 TEs) and 587 SNP-derived (6,465 SNPs) peaks (Figure 3A and 3B). More than half of the TE peaks (53%) overlapped with SNP peaks, and most of the TEs (69%) were contained within overlapping peaks. By contrast, only a few SNP-derived peaks (2%) overlapped with the TE peaks, but one-fourth (25%) of the SNPs were contained within these peaks (Figure 3B). We compared the significant SNPs from each bin between the overlapped and non-overlapped peaks, and the number of SNPs was significantly enriched ($p < 0.05$) in the overlapping peaks relative to the non-overlapping peaks (Figure 3C). Furthermore, a total of 74 genes were identified within the TE peaks, and more than half of them (54.1%; 40/74) were detected within the SNP-associated peaks (Figure 3D; Supplemental Table 9). For specific traits, all of the TE peaks for spikelet number, panicle number, and leaf blade width overlapped with SNP peaks. No TE peak was found for days to heading. The other traits showed partial overlap with the SNP peaks, except for panicle length in 2013 (Figure 3E; Supplemental Table 10). Some candidate genes that may control the related traits were found under overlapping peaks (Supplemental Table 9).

Candidate genes in overlapping peaks

A strong association with plant height was identified for both TE markers and SNPs on chromosome 11 (Figure 3F and 3G). Among the genes underlying this peak was one candidate gene (*LOC_Os11g08410*) in Bin_89 of SNP peak 11_88_98 that overlapped with TE peak 11_88_90 (Figure 3F and 3G; Supplemental Table 9). This same candidate gene was also found within the overlapping peak for panicle length (Supplemental Figure 8; Supplemental Table 9). The *LOC_Os11g08410* locus encodes a GATA zinc-finger-type transcription factor that has been shown to control plant height and panicle length in rice (Yano et al., 2016).

On chromosome 4, both TE and SNP peaks that overlapped exhibited a strong association with leaf blade width (Figure 3H and 3I; Supplemental Figure 1). One gene within the peak region, *LOC_Os04g52479*, encodes *Narrow Leaf 1* (*NAL1*), which has been shown to control flag leaf width in rice (Qi et al., 2008). Peaks were located in Bin_625 in SNP peak 4_624_629 that overlapped with TE peak 4_624_624 (Figure 3H and 3I). A TE

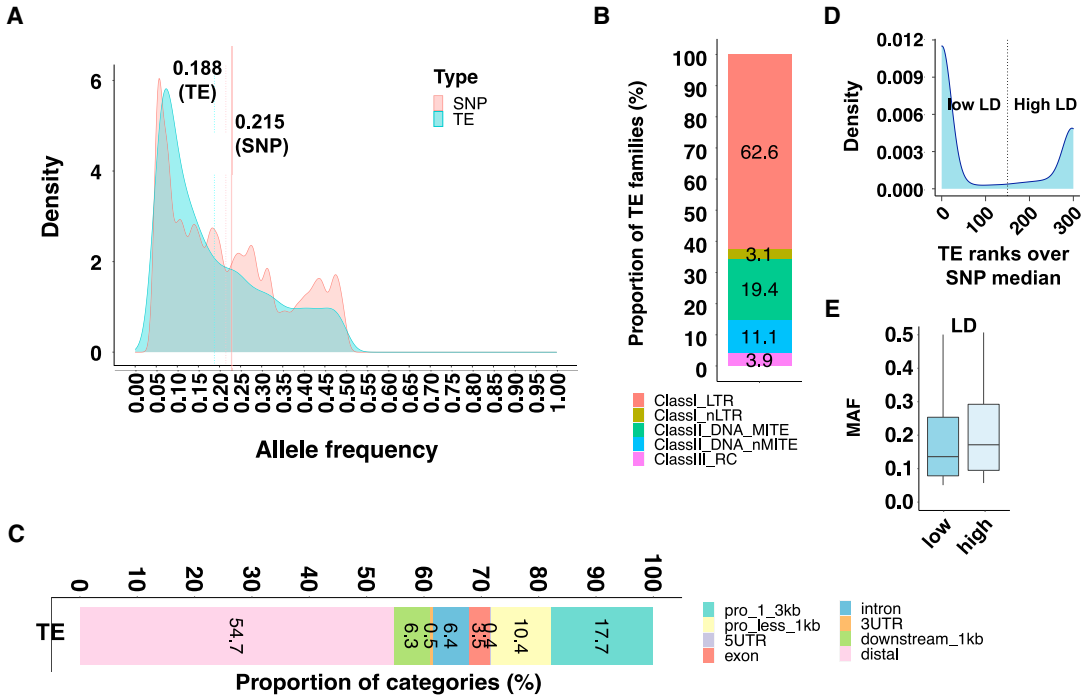


Figure 2. TE space landscape.

(A) Distribution of minor allele frequencies (MAFs) across TEs and SNPs. Dashed vertical lines indicate the mean MAF.

(B) Proportions of five TE families.

(C) Proportion of TEs by proximity to genes.

(D) Density of the number of TE r^2 ranks (0–300) that are above the SNP-based median r^2 value for common TE variants.

(E) Distribution of MAFs for each LD category. Boxes represent the interquartile range (IQR) from quartiles 1 to 3.

within this interval was situated about 17 kbp away from *NAL1* (Figure 3I; Supplemental Table 9).

Panicle number

On chromosome 4, TE peak 4_624_624 overlapped with SNP peak 4_624_625. Three candidate genes were identified in the overlapping region. Two of them (*LOC_Os04g52440* and *LOC_Os04g52450*) are homologous to *Arabidopsis POP2*, which encodes a transaminase that degrades γ-amino butyric acid (GABA) (Supplemental Table 9). Functionally, *pop2* flowers accumulate GABA, and the development of their pollen tubes is inhibited (Ma, 2003).

Spikelet number

The overlapping peak region associated with TE peak 7_35_35 and SNP peak 7_35_35 contained two candidate genes (Supplemental Table 9). One gene (*LOC_Os07g04050*) is orthologous to *AT4G32460*, which participates in pectin methyl esterase regulation during seed germination (Zúñiga-Sánchez et al., 2014). The other (*LOC_Os07g04040*) is orthologous to *AT3G53810*, and mutation of this gene leads to male sterility, resulting in a defect in pollen development in *Arabidopsis* (Wan et al., 2008).

Awn length

About one-third of the significant TEs (31%; 8/26) were associated with awn length, and four of them were under SNP peaks (Supplemental Table 10). Under the overlapping peaks, three rice genes may encode concanavalin-A-like lectin protein kinase (Supplemental Table 9). Lectin plays an important role in plant defense, which may induce variations in awn length (Koh et al., 2007; De Hoff et al., 2009; Amiri et al., 2013; Huang et al., 2020).

Positive selection on transposons associated with phenotypic changes

In this rice population, some domestication-related loci and nearby candidate genes have been identified based on SNPs (Yano et al., 2016). However, the effects of TEs on agronomic trait variations during domestication are less clear. To characterize potential selection on the frequency of TEs associated with phenotypic changes, we binned the population into two groups (in materials and methods: analysis of selective sweeps), the upper and lower 30%. Group 1 (G1) contained the top 30% of individuals with the highest trait values, and G2 contained the bottom 30% of individuals with the lowest trait values. In total, seven paired groups based on the seven traits were analyzed using the 6,073 TEs (MAF > 0.05). A total of 417 sweep regions (SRs) were identified, covering 547 TEs (SR-TEs). From non-SR-TEs to SR-TEs, there was an increasing proportion (44.8%–50.1%) of TEs near genes (Figure 4A). Compared with TEs overall, there was an increase in the frequency of SR-TEs at the ends of chromosomes, consistent with the SNP results (Figure 4B; Supplemental Figures 9 and 10).

We next investigated the genes underlying the SRs. A total of 346 TE and 697 SNP SRs were identified, 58 and 52 of which were shared for the TEs and SNPs, respectively. About 102 genes were identified in these regions, and some were associated with

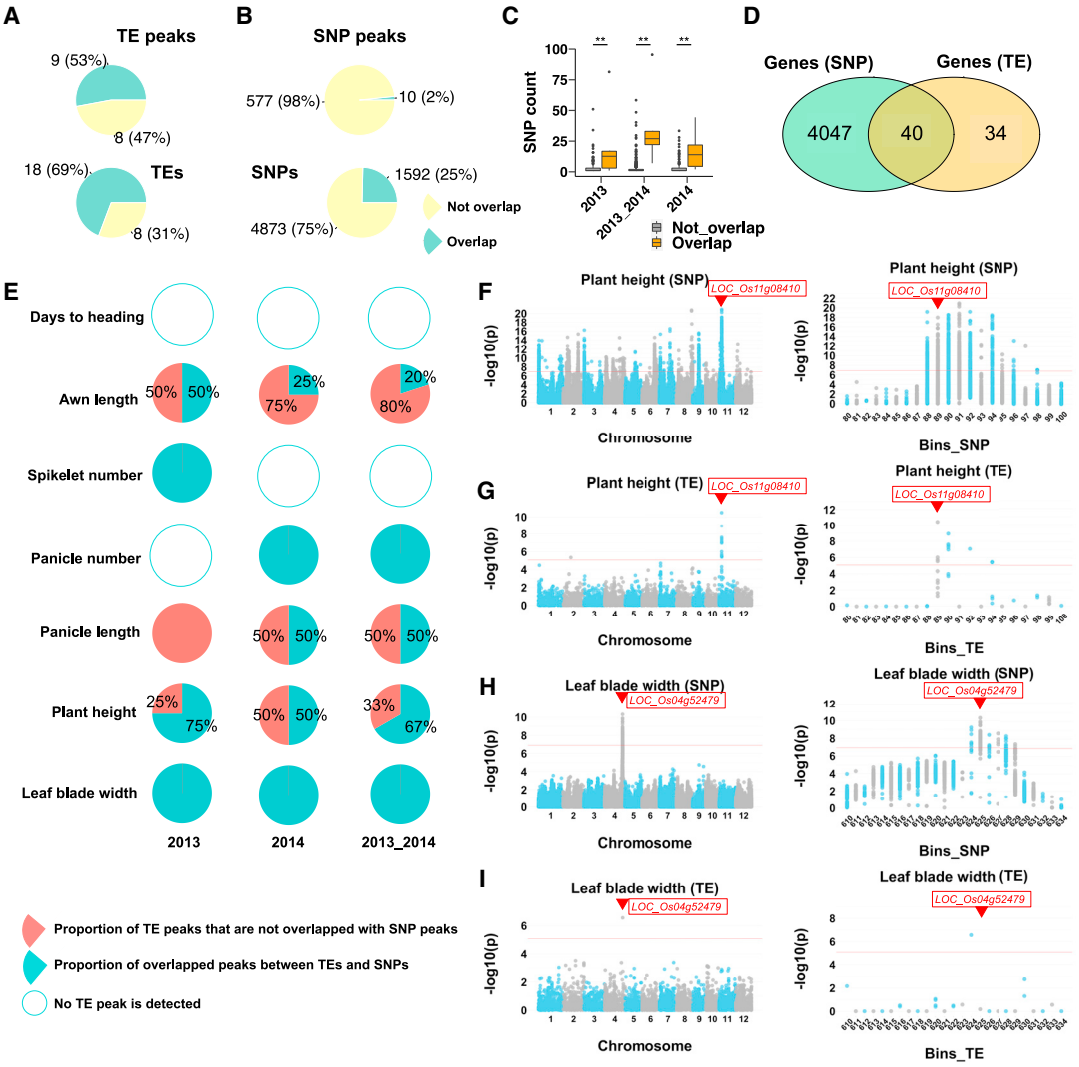


Figure 3. Comparisons of TEs and SNPs associated with phenotypes through GWAS.

(A) Proportion of overlapping TE versus SNP peaks and TEs contained within overlapping versus non-overlapping peaks.

(B) Proportion of overlapping SNP versus TE peaks and SNPs contained within overlapping versus non-overlapping peaks.

(C) SNP count binned by overlap (yes/no) across data collection years 2013, 2014, and 2013 & 2014.

(D) The number of shared and specific candidate loci under the SNP and TE peaks.

(E) Overlap of TE and SNP association peaks across seven phenotypic traits.

(F–I) Manhattan plots between SNPs and TEs for two traits: plant height and leaf blade width. The red line represents the significance threshold based on Bonferroni correction. The right plot shows a bin-level region derived from the highlighted peak in the left plot. Each bin contains 50 kb. Arrowheads in plant height and leaf blade width indicate the positions of *LOC_Os11g08410* (Bin_89) and *LOC_Os04g52479* (Bin_625), respectively.

phenotypic traits that were under selection (Supplemental Tables 11 and 12). For example, two candidate genes (*LOC_Os11g37520* and *LOC_Os04g01990*) for plant height were identified. *LOC_Os11g37520* encodes *Ethylene overproducer 1 (ETO1)-like protein 1*, which is involved in ethylene signaling regulation to influence plant growth (Iqbal et al., 2017). The *LOC_Os04g01990* locus is orthologous to *Slow growth1* (*SLO1*) in *Arabidopsis*, and *slo1* plants are smaller than the wild type (Grennan, 2011). Interestingly, under non-shared SRs of TEs, one candidate gene (*LOC_Os11g08410*) was located on chromosome 11. This gene encodes a GATA zinc-finger-type transcription factor that has been shown to control plant height in rice (Yano et al., 2016) (Figure 4C). Several candidate genes for other traits were also detected under overlapping regions. For awn length, *LOC_Os06g32970* is orthologous to *AT3G05975*, which belongs to the late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family. For days to heading, *LOC_Os03g03100* (OsMADS50) is an important activator of flowering in rice (Ryu et al., 2009) (Supplemental Tables 11 and 12; Supplemental Figure 11). These observations are consistent with the potential for positive selection to act on TEs and result in phenotypic changes among rice populations.

Transposons potentially relate to the presence of indels in nearby gene bodies

It is not surprising that transposon insertions can influence phenotypes, but work remains to be done in tracking the history of

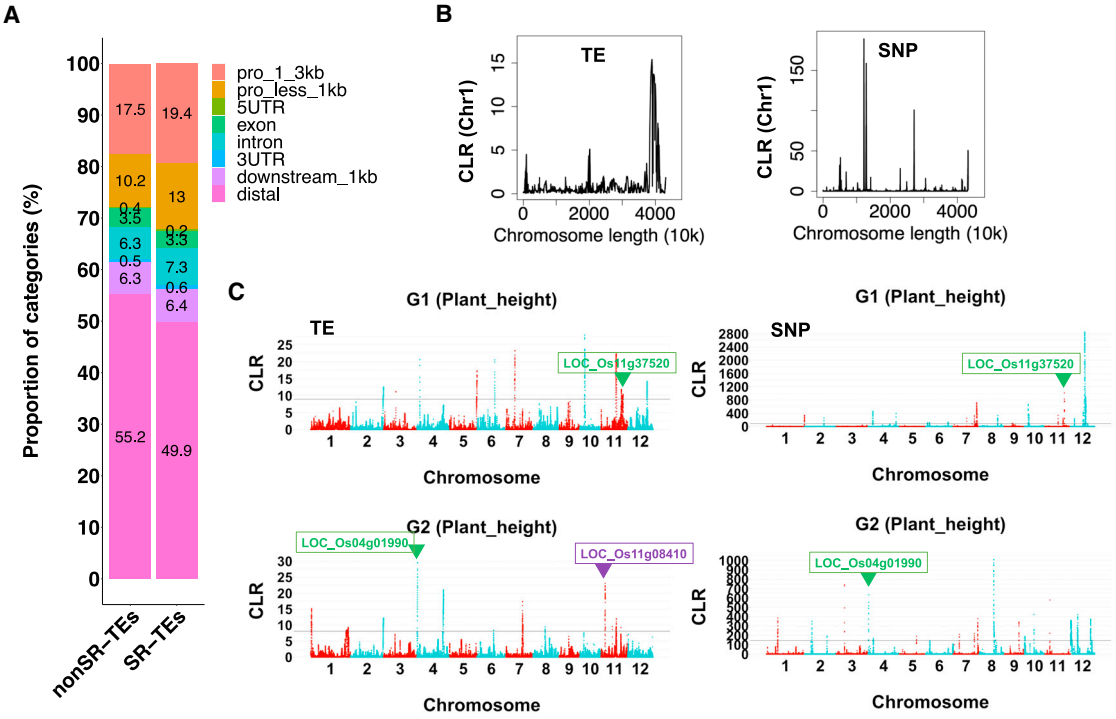


Figure 4. Positive selection on TEs associated with phenotypic changes.

(A) Proportion of TEs under selective sweep regions (SRs) or non-SRs and the proximity to gene relative abundances.

(B) Average of CLR from all seven traits on chromosome 1.

(C) Manhattan plots of CLR of group 1 (G1) and G2 in plant height from 2013. G1 and G2 contain individuals from the upper and lower 30% of trait values, respectively. The horizontal black line indicates the CLR value significance threshold. Arrowheads with gene names suggest candidate genes that control the relative traits. Green genes are shared by the TE and SNP SRs, whereas the purple genes are found only under the TE SRs.

their movements and understanding the significance of their transposition. Transposons leave footprints after moving out, and these may provide clues to help trace back their movements (Muñoz-López and García-Pérez, 2010). To analyze the TE jumping that resulted in phenotypic changes, we focused mainly on possible footprints in candidate genes under SNP association peaks in order to find evidence of transposons that may have affected these genes.

TE insertions or deletions (indels) were associated with phenotypes in this rice population via GWAS. A total of 1,090 indels were significantly associated with the studied traits (Supplemental Table 13); 246 of them were in gene body regions and were also identified in SNP association peaks (Supplemental Table 14). To identify possible TE footprints derived from these 246 indels, co-occurrence of indels and TEs was assessed. Because this could occur through random associations, we tested it against a binomial model (Supplemental Table 15; Materials and methods: Co-occurrence of indels and TEs). Subsequently, two scenarios were examined: TEs within overlapping TE-SNP GWAS peaks and TEs outside of these co-localized GWAS peaks (Figure 5A and 5B). Overall, TEs under overlapping peaks displayed greater correlations with indels than did TEs outside of overlapping peaks (Supplemental Figures 12 and 13). Indels were highly correlated ($|rho| > 0.8$; $p < 0.05$) with the presence/absence of 14 TEs within overlapping peaks and seven TEs outside these peaks (Supplemental Table 15).

Under the overlapping peaks, an indel (indel_h1; indel_11_4433083) identified within the gene body of *LOC_Os11g08410* was significantly correlated ($|rho| = 0.918$, $p < 0.05$) with a MITE TE (TE_h1; TE_11_4558551) approximately 120 kbp upstream of the gene (Supplemental Table 16; Figure 5C). The presence of the insertion in the indels (1/1) and the absence of TEs (1/1) were associated with a significant increase in plant height (Figure 5D). Three cases were identified under non-overlapping peaks, including *mPing* (TE_h2; TE_2_13161937), which was significantly correlated ($|rho| = 0.509$, $p < 0.05$) with indel_h2 (indel_11_4439639) on a different chromosome. In the second case, an LTR TE (TE_h3; TE_11_4458534) was significantly correlated ($|rho| = 0.721$, $p < 0.05$) with indel_h3 (indel_11_5565303) approximately 1 Mbp away from the TE (Supplemental Figure 14). In the third case, one MITE MuDR TE was significantly correlated ($|rho| = 0.460$, $p < 0.05$) with indel_h4 (indel_3_859132) on a different chromosome (Supplemental Figure 15). TE_h2 and TE_h4 belonged to the deletion status, whereas TE_h5 was in the insertion status (Figure 5E; Supplemental Figures 14A and 15A; Supplemental Table 16), and their alternative genotype (1/1) was associated with significantly ($p < 0.01$) higher trait values. Results for indel_h2, indel_h3, and indel_h4 were similar (Figure 5F; Supplemental Figures 14B and 15B).

To further understand the relationships between these three TEs and indels, footprints of the TEs were identified (Materials and methods: Identification of TE footprints). The alternative sequences at indel_h1, indel_h3, and indel_h4 and the reference

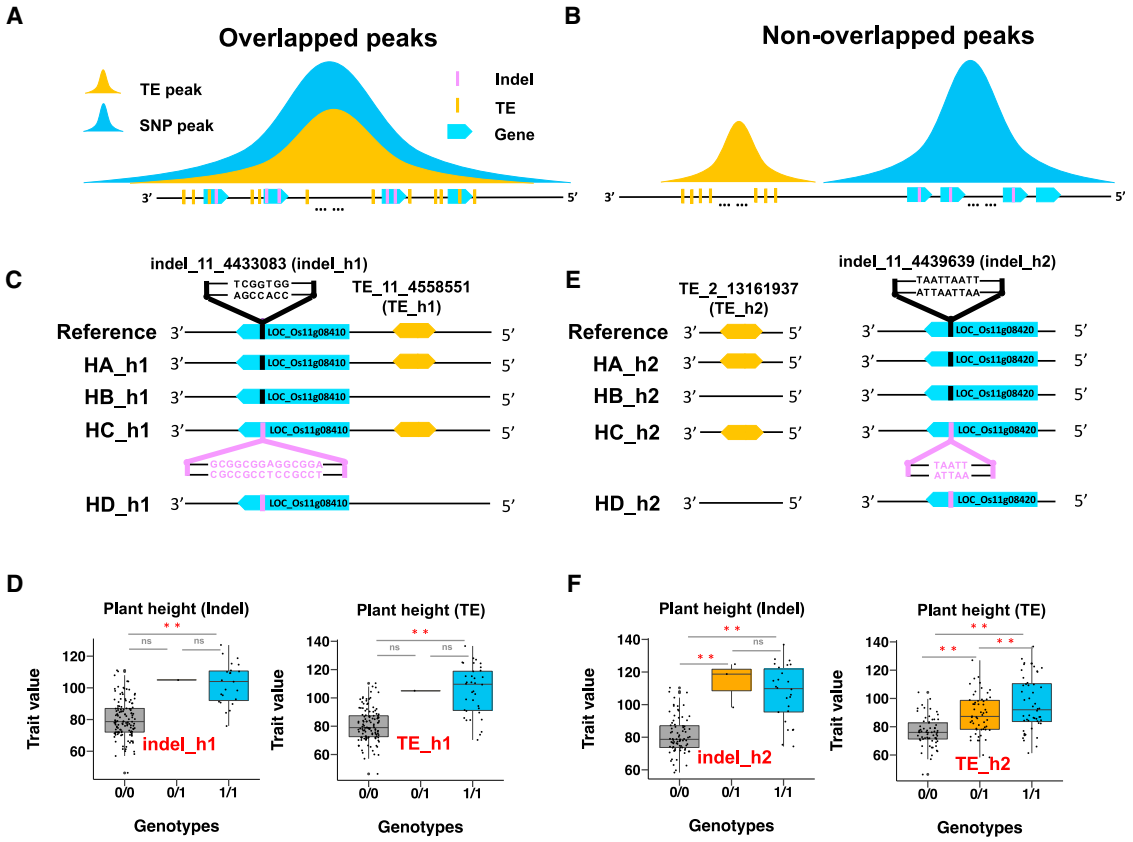


Figure 5. Association of TEs and indels with phenotypic traits.

(A) In overlapping peaks, TEs are inserted around candidate genes, and indels are contained in gene body regions.

(B) Under non-overlapping peaks, TEs are found far from candidate genes, and indels are contained in gene body regions.

(C) Four haplotypes were identified based on the indel_h1 and TE_h1 loci (here, the letter-number designation indicates trait and analysis position, e.g., "h1" is an abbreviation for plant height for the first analyzed location).

(D) Comparisons of trait values among different genotypes at the TE_h1 and indel_h1 loci, respectively. *$p < 0.01$ and no significance (ns) $p > 0.05$. 0 and 1 of the genotypes indicate the reference allele and alternative allele.

(E) Similar to (C), but with different TEs (TE_h2) and indels (indel_h2).

(F) Analyses similar to those in (D) for the h2 analyzed locations.

sequence at indel_h2 were consistent with the footprints of these four TEs (Figure 6A; Supplemental Figures 14C and 15C, and 16–18), indicating that the TEs are probably related to these indels. We next compared the sample numbers of individuals for nine combinations of haplotypes. For indel_h1&TE_h1 and indel_h4&TE_h4, most samples had haplotype HAA, in which indels were not found in the bodies of genes near TE_h1 and TE_h4. By contrast, for indel_h2&TE_h2 and indel_3&TE_h3, the TEs co-occurred with indels in nearby gene bodies. These two cases suggest that TEs may be connected to the occurrence of indels (Figure 6B).

Taken together, our results suggest that TEs near or far from the candidate genes are both likely to be correlated with indels found in these gene bodies. The detection of TE footprints within these indels suggests the potential for TE-mediated regulation at large genetic distances.

DISCUSSION

The importance of generating informative genetic markers for agronomic research, human health, and evolutionary inference cannot be overstated (Qiu et al., 2013). Here, the generation of hundreds to thousands of SNP variants within and among populations generates the power needed to detect loci that are associated with population- and trait-level differences. Importantly, the abundance of SNPs often identifies many candidate genes, and the linkage between these markers can generate spurious associations, obfuscating inferences about the underlying genetic mechanisms (Hirschhorn and Daly, 2005; Wang et al., 2020). We demonstrated that the development of functional TE-derived markers supplements the identification of loci associated with rice phenotypic traits through genome scans and GWAS. We found that changes in the TE landscape were correlated with allelic variance, driving changes in trait values over short (<120 kbp) and large (~1 Mbp) genetic distances. Finally, we developed a start-to-finish pipeline (TEmarker; https://github.com/yanhaidong1/TEmarker) for the development and implementation of TE-derived markers from GWAS data.

Recent work has identified TIPs in several major crops, including maize, rice, and tomato (Akakpo et al., 2020; Carpentier et al., 2019; Lai et al., 2017; Roy et al., 2015). These studies performed GWAS of TE insertions associated with important

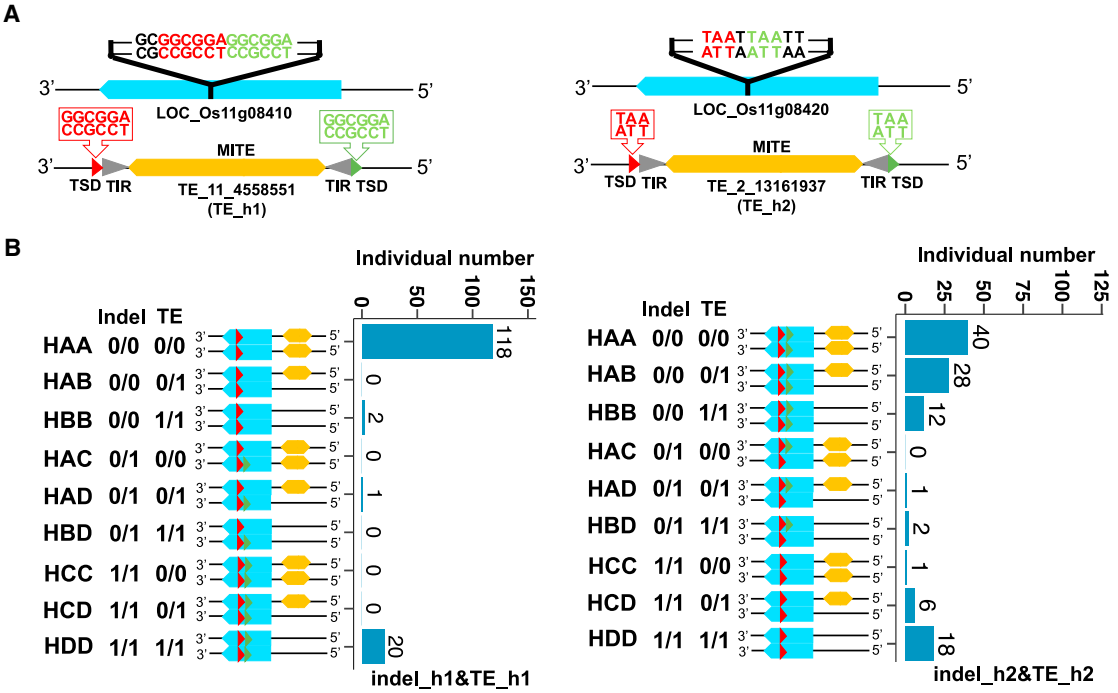


Figure 6. Two cases show possible relationships between TEs and indels.
(A) Footprints of TE_h1 and TE_h2. TSD, tandem site duplication; TIR, tandem inverted repeat; LTR, long terminal repeat.
(B) Individual numbers of nine haplotype combinations in the diploid rice population.

agronomic traits, validated insertions associated with nearby gene expression, and showed that TEs could lead to the generation of multiple transcript isoforms. Here, we focused more on how TEs influence phenotypic changes in rice. We also analyzed positive selection on TEs associated with phenotypic changes. By correlating TE insertions with indels in genes, we proposed two possible ways that TEs may produce allelic variations associated with agronomic traits. Previous TIP tools have leveraged different algorithms to identify TE insertions, but the predicted TIPs are highly divergent based on simulated testing data (Supplemental Figure 19). Thus, TEmarker was developed to combine results from different tools, remove potential false-positive TE insertions (Supplemental Tables 4–6) and perform high-throughput genotyping. Ideal genetic markers usually have the following properties: (1) they have high levels of polymorphism that demonstrate measurable differences among individuals in a population; (2) they can discriminate heterozygotes and homozygotes (co-dominance); (3) they can provide polymorphism per character (bi- to multi-allelic); (4) they are useful for assessing genetic diversity within/among populations; and (5) they can be used to identify alleles associated with traits (Bader, 2001; Mammadov et al., 2012). In this study, 6,073 TEs (MAF > 0.05) were identified and characterized for functional utility (Figure 2). In GWAS, TEs were able to identify the same association peaks as SNPs (Figure 3). Also, genome scans identified loci under positive selection that had previously been identified as genes under selection in rice (Figure 4). Taken together, these results demonstrate the utility of TEs as genetic markers. In addition, understanding the effect size of markers in GWAS is important for characterizing the basic properties of the data and determining the significance of markers (Bukszár and van den Oord, 2010). A previous study of tomato found that the effect size of TIPs was much larger than that of SNPs within the same population (Dominguez et al., 2020). By contrast, our results in rice showed the opposite pattern: SNPs still had a greater effect on the phenotypes of interest (Supplemental Table 17). Therefore, in practice, it is unclear whether the effect of TE genetic markers is system-dependent; nonetheless, TE genetic markers certainly have a complementary role relative to SNP markers, and it is therefore important to use both markers for GWAS.

In genetic association studies, LD estimates the degree of correlation among markers (Bush and Moore, 2012). LD creates a scenario in which a marker that is directly connected to the trait of interest cannot be distinguished from a marker that is linked to a diagnostic marker. In a GWAS, the second outcome can result in a significant association with a SNP that may or may not be the influential SNP, and additional studies are required to identify the causal variant (Hirschhorn and Daly, 2005; Wang et al., 2020). Therefore, prevalent LD among SNP variants often leads to the loss of a majority of markers and to wasted genotyping effort, and causal variants are still difficult to pinpoint. We found a greater number of TEs that showed low LD with nearby SNPs, suggesting that they might represent a source of genetic variation not detected by SNPs (Figure 2D and 2E). In practice, we found that TE GWAS identified many fewer associations than SNP GWAS for each trait (26 versus 6,465). Intriguingly, although the majority of TE association peaks overlapped with SNP association peaks, the inverse was not true, with just a 2% overlap. Yet, about one-fourth of the SNP markers were contained within the TE overlap peaks. This suggests that reduced LD among TE markers is winnowing

spurious SNP associations. In fact, among the co-localized peaks, we were able to detect previously identified candidate genes associated with the studied phenotypic traits. Thus, from the same data used to generate SNPs, TE markers can help to decrease spurious GWAS associations once both datasets identify the same candidate genes, without the need for additional experimentation.

TEs have been shown to operate as long-distance enhancers of gene expression, particularly in primates (Jacques et al., 2013; Trizzino et al., 2017). For example, in mouse, a TE-derived SINE is an enhancer that lies 488 kb downstream of *ISL1*, which encodes a transcription factor that is essential for motor neuron differentiation (Bejerano et al., 2006). In plants, the most elegant example is long-distance regulation by the LTR retrotransposon *Hopscotch* in maize. This TE is located 60 kbp upstream of the gene *tb1* and acts as a long-distance enhancer by regulating *tb1* transcription (Lisch, 2013). In the present study, a TE marker associated with variance in leaf blade width was about 17 kbp away from *NAL1*, a gene previously shown to control this trait (Figure 3H and 3I). In another case, a TE marker significantly associated with plant height was identified about 120 kbp from a candidate gene (*LOC_Os11g08410*) known to control this phenotype (Figure 5C). Although more work needs to be done to document such interactions, TE markers provide the raw evidence needed to unravel complicated regulatory mechanisms.

The movement of TEs is another challenging aspect of this mechanism. TE_h1 is located within 120 kbp of the indel_h1 locus (Figure 5C), and it is plausible that TEs can jump from indel_h1 to the nearby region. By contrast, TE_h2 and TE_h3 are located 1 Mb away and on different chromosomes from indel_h2 and indel_h3 (Figure 5E and Supplemental Figure 14A), which is harder to resolve. One hypothesis is that there is a decreased functional distance, either through a loop or chromatin packing, that creates less separation and affords an opportunity for TE transposition. Although plausible, these hypothesized mechanisms require more evidence for evaluation. In summary, however, it is clear that the use of TE markers enables the identification of candidate loci associated with phenotypic variance and that these markers can serve as anchors for downstream analyses to elucidate mechanisms or identify stable traits for crop improvement.

Footprints within indels and the presence/absence of nearby TEs suggest a possible connection between TEs and indel frequency (Figure 6B). On the basis of this result, we hypothesize two possible ways type A and type B (Supplemental Figure 20A), in by which TEs may produce these alleles can be hypothesized. For type A, a TE (TE_h1) in a homozygous state close to the target hits the gene where an indel will be produced (indel_h1). Three different potential scenarios can be proposed based on the haplotypes found in this analysis. In the first scenario, only one of the alleles is hit by one of the TE copies, leaving an indel in one allele when the TE jumps away. In this scenario, the other TE copy may stay or may be removed during the process. In a second scenario, both TE insertions reach a homozygous state before they are removed. An ancestral TE copy remains close to the gene. In the last scenario, no close TE copy remains when the TE insertion becomes homozygous (Supplemental Figure 20). For type B, a target gene is hit by a TE located far from it in the genome. When the TE is removed, an indel is produced (indel_h2). Two different scenarios can be proposed in this case, depending on whether the TE insertion reaches the homozygous state. Six different types of haplotype were found, depending on the final position of TE_h2 (Supplemental Figure 20A). The indel_h3&TE_h3 pair shows a situation similar to type B, except that the reference allele does not contain a TE (Supplemental Figure 14D), whereas indel_h4&TE_h4 exhibits a pattern similar to type A (Supplemental Figure 15D). To further investigate whether these two jumping paths are related to phenotypic changes, we explored the previously described genomic scenarios in two groups: group1, which contains a haplotypic combination before the TE hit, and group 2, which contains haplotypic combinations after the TE hit (Supplemental Figures 14E, 15E, and 20B). Significant differences ($p < 0.01$; two-tailed Welch's t test) (Ahad and Yahaya, 2014) were detected between these two groups, suggesting that the TE- and indel-derived alleles may be associated with phenotypic variation.

Concluding remarks

In this study, we evaluated the utility of TE-derived markers for use in genetic mapping and identified several features of TE markers. (1) Positively selected regions contain not only genes but also TEs that may contribute to phenotypic changes in rice. (2) TE markers detect association peaks that are equivalent to those produced by SNP markers through GWAS. (3) The production of new alleles responsible for new phenotypes may occur from the action of TEs, not only in the vicinity of the gene but also at greater distances. Nevertheless, further experiments are necessary to test this hypothesis. Finally, we developed a community tool, TEmarker (https://github.com/yanhaidong1/TEmarker), which can serve as a standalone tool or as a complement to SNP-marker pipelines. Importantly, this tool can be helpful for narrowing the number of candidate genes predicted from SNP markers, thereby potentially accelerating crop breeding. In addition, this tool may provide an accessible way to identify potential relationships between candidate loci and TEs.

MATERIALS AND METHODS

Dataset collection

One set of whole-genome shotgun sequencing data was obtained from one rice population (Supplemental Table 1). This dataset (DRA004358) includes measurements of seven traits in 176 *Oryza sativa* subsp. *japonica* accessions over 2 years (2013 and 2014): heading time, plant height, panicle length, panicle number, awn length, leaf blade width, and spikelet number (Yano et al., 2016) (Supplemental Table 2).

Transposon library generation

TEs in test reference genomes were annotated by combining *de novo* and homology-based approaches. For the 176 rice genome sequences, Os-Nipponbare-Reference-IRGSP-1.0 was downloaded from https://rapdb.dna.affrc.go.jp/download/irgsp1.html (Sakai et al., 2013). For the *de novo* approach, we used RepeatModeler (v.1.0.4) (http://www.repeatmasker.org/RepeatModeler/) (Smit and Hubley, 2018) and LTR_retriever (v.2.8.6) (Ou and Jiang, 2018) to build a *de novo* repeat library with default parameters. For the homology-based approach, we used RepeatMasker (http://www.repeatmasker.org, v.3.3.0) against Repbase (v. 20170127) (Price et al., 2005; Xu and Wang, 2007; Chen, 2009) with default parameters. The insertion time of LTRs in the TE library was obtained using the LTR_retriever tool.

Validation of TEmarker

To further confirm that TEmarker could accurately identify TE presence/absence polymorphisms, we obtained re-sequencing data from four *Oryza sativa* subsp. *japonica* accessions (SRR063607, SRR063630, SRR063631, and SRR063637) (Xu et al., 2012). To simulate TE presence, we first used the built TE library (Materials and methods: Transposon library generation) to mask TEs on chromosome 1 of the *japonica* genome using RepeatMasker (Chen, 2009), and we randomly selected identified TEs with at least 10× read coverage after removing all clipped reads on the TEs. The selected TEs were removed to construct a simulated genome. Loci where TEs were removed were the simulated TE insertions when we mapped the re-sequencing data to the simulated genome. Next, six tools from McClintock (Nelson et al., 2017) (ngs_te_mapper [Linheiro and Bergman, 2012], RelocaTE2 [Chen et al., 2017], TEMP [Zhuang et al., 2014], RetroSeq [Keane et al., 2013], PoPoolationTE [Kofler et al., 2012], and TE-locate [Platzer et al., 2012]), as well as Jitterbug tools (Hénaff et al., 2015), were used to generate combined TE location information based on re-sequencing data from the four samples. If the predicted TE loci were within 200 bp of the simulated TE insertions, these predicted loci were regarded as true-positive TE loci. We used the same re-sequencing data from the four accessions to simulate TE absence. TEs from the library were randomly selected and inserted into chromosome 1 of the rice genome to produce a simulated genome. These simulated insertions are true-positive TE absences if reads are mapped to these regions. SE, SP, AC, and PR are defined as follows:

$$SE = TP/(TP + FN);$$

$$SP = TN/(FP + TN);$$

$$PR = TP/(TP + FP); \text{ and}$$

$$AC = (TP + TN)/(TP + TN + FP + FN).$$

To validate TEmarker using biological data, we analyzed one dataset (PRJNA565484) with one genome assembly (IRGC132278) and Illumina short-read data (Zhou et al., 2020). We first used an EDTA tool (Ou et al., 2019) to identify TEs defined as known TEs in the IRGC132278 genome. Next, we mapped the short reads to the IRGSP-1.0 genome (Sakai et al., 2013) to predict TE integrations using TEmarker. About 1 kb up- and downstream, sequence pairs of the integration sites were extracted from the IRGSP-1.0 genome and then blasted against the IRGC132278 genome. The intermediate sequences between the two blasted regions from the sequence pairs were the predicted TE locations in the IRGC132278 genome. Finally, we checked whether the predicted TE sites overlapped with known TEs in the IRGC132278 genome.

Analysis of selective sweeps

SweepFinder2 was used to calculate a composite likelihood ratio (CLR) to search for signs of selective sweeps in populations based on SNP and TE markers. SweepFinder2 was run with a grid size of 10 kb (DeGiorgio et al., 2016). In the analyses of SNP and TE markers, the top 1% of windows were selected as SRs, and the CLR of a window at the end of the top 1% was defined as the threshold score. The CLR scores were merged into SRs if the neighboring scores exceeded the threshold score.

LD analysis

To analyze LD of TEs, we referred to a method from a previous study (Yang et al., 2019). In brief, for each TE integration site, we selected 150 up- and downstream SNPs. TEs without enough SNPs were filtered out (300), leaving 5,545 TEs. PLINK (v.1.07) (Purcell et al., 2007) was used to calculate pairwise genotype LD (r^2 value) for all complete cases for SNP-TE and SNP-SNP pairs. A median SNP-SNP r^2 value was obtained after ordering all of the values. For each of the 300-ranked surrounding positions for the TE-SNP sets, we calculated the number of times that the r^2 value of TE-SNP pairs was higher than the median SNP-SNP r^2 value. If this value was over 150 ranks, we classified the TEs as low-LD TEs, whereas the TEs with values less than 150 ranks were defined as high-LD TEs.

GWAS analysis

GWAS analysis was performed using a linear mixed model (LMM) for rice (Yano et al., 2016) and grape (Liang et al., 2019) datasets. The LMM assumes the model $Y = X\beta + Zu + \varepsilon$, where X is a matrix of fixed effects including marker polymorphisms, β is a vector of fixed effects that can model both environmental factors and population structure, Z is an incidence matrix that shows a relationship between Y and u, u is a random effect with $u \sim N(0, K\sigma^2_G)$, and ε has a matrix of residual effects with $\varepsilon \sim N(0, I\sigma^2_e)$. More detailed information about this model can be found in Yano's study (Yano et al., 2016). The R package rrBLUP (v.4.6) was used to perform GWAS (Endelman, 2011). Two statistical thresholds were established, a suggested p value threshold ($p < 1 \times 10^{-5}$) and an adjusted p value ($p < 0.05$) for associations. Given that TEs could be 60 kb upstream of the gene (tb1) and regulate transcription (Lisch, 2013), association peaks were defined by dividing the genome into multiple 50-kb bins. In order not to miss TEs that may influence genes farther than 50 kb away, we manually checked TEs within 3 bins (150 kb) of some potentially important genes that control corresponding traits. A bin was regarded as a peak when it contained an associated TE or SNP, and nearby peak regions were merged into one peak region. The peaks were named with an "AA_BB_CC" format in which AA indicated chromosome number, BB indicated start-bin number, and CC indicated end-bin number.

Co-occurrence of indels and TEs

Indel calling was performed on the rice re-sequencing data using FreeBayes (v.0.9.21) (Garrison and Marth, 2012) with the following parameter settings: –min-mapping-quality 20 –min-base-quality 20 –min-coverage 5 –no-snps. The lengths of sequence strings from alternative and reference indel loci were compared, and the longer one was assumed to be the TE footprint based on the criteria of TE movement (Wicker et al., 2007). If the alternative indel had a longer sequence than the reference, we assigned "1" to the indel genotype "1/1" or "0/1" and "0" to the indel genotype "0/0," where the value "1" indicates that there is an assumed footprint in the indel locus. If the alternative indel had a shorter sequence than the reference, we assigned "1" to the indel genotype "0/0" or "0/1" and "0" to the indel genotype "1/1." Because there are two major ways that TEs can leave footprints in indels, regardless of TE presence or absence, we generated two plans (A and B) to assign values to the genotypes. In A, "1" was assigned to "1/1" and "0/1," where "1" indicates that the TEs were assumed to have left a footprint in the indels, and "0" was assigned to "0/0." In B, "1" was assigned to "0/0" and "0/1," and "0" was assigned to "1/1." Pearson correlation analysis was performed between the values of indels and TEs under each of the plans, and we selected the plan with the higher correlation rate, which represents the level of co-occurrence of indels and TEs.

To further evaluate how often TEs and indels co-occurred, we randomly selected 24 TEs that were significantly associated with phenotypes of interest and were correlated with indels under the SNP association peaks. This process was repeated 50 times; each time, we calculated the proportion of identified TE-indel pairs relative to all possible TE-indel pairs that could be obtained by $y = a \times b$, where a is 24 TEs, b is 246 indels, and y is 5,904 possible pairs. Finally, we obtained an average proportion of 0.021 from all 50 simulations. In our real data, 1,373 unique pairs were found (Supplemental Table 15). A binomial test was used to evaluate the probability that these were random correlations by setting 0.021 as the probability of success, 1,373 as the number of successes, and 5,904 as the number of trials based on a one-tailed test ("greater"). The returned p value was 0.

Identification of TE footprints

We used different approaches to identify footprints at the TE_h1, TE_h2, TE_h3, and TE_h4 loci. For the MITE TE at TE_h1, we extracted flanking sequences within 15 bp of all copies and used MUSCLE alignment (Madeira et al., 2019) (https://www.ebi.ac.uk/Tools/msa/muscle/) to manually check the possible tandem-site duplication (TSD) sequence and length. We next calculated the frequency of letters for the sequences with potential TSD information based on the Biostrings R package (Pages et al., 2016) and used the seqLogo R package (Bembom, 2007) to plot the TSD in RStudio (v.1.1.383) (Allaire, 2012) (Supplemental Figure 16). The TSD for the mPing at TE_h2 has been well documented (Robb et al., 2013; Chen et al., 2019). For the LTR TE (ClassI_LTR_Gypsy_TE688) at TE_h3, because most of its copies in the genome are fragmented and it is difficult to detect their true flanking sequences to make alignments, we manually checked the sequences surrounding the insertion position of this TE and identified the TSD (Supplemental Figure 17). For the MITE TE at TE_h4, we extracted sequences of all copies of this TE in the genome. Given that most of the copies did not have a complete TIR at two flanking regions, we manually checked each TE and found a copy with the complete TIR and possible TSD (Supplemental Figure 21).

DATA AND CODE AVAILABILITY

We developed a start-to-finish pipeline (TEmarker) that can be accessed and downloaded from GitHub (https://github.com/yanhaidong1/TEmarker). The simulated data and scripts used for validation can be downloaded at https://de.cyverse.org/dl/d/6885D6CC-3BDD-4F86-9982-CA57DCD79A31/Simulated_data_scripts_presence_validation.tar.gz and https://de.cyverse.org/dl/d/3074AC4B-141C-425B-879C-83D596EE79B8/Simulated_data_scripts_absence_validation.tar.gz. The re-analyzed dataset from Yano's study (DRA004358) can be downloaded from https://www.ncbi.nlm.nih.gov/sra?term=DRA004358 (Yano et al., 2016).

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

AUTHOR CONTRIBUTIONS

A.B. and H.Y. conceived and supervised this study. H.Y. performed the data analyses. H.Y. and A.B. annotated the results. H.Y. and D.C.H. wrote the manuscript. S.L. and L.H. revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors want to acknowledge the School of Plant and Environmental Sciences at Virginia Tech for the support of H.Y. They also want to acknowledge the Advanced Research Computing unit (www.arc.vt.edu) for the use of computational resources. Finally, the authors would like to acknowledge the reviewers for their useful suggestions that helped to improve the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 26, 2021
Revised: October 29, 2021
Accepted: December 16, 2021
Published: December 20, 2021

REFERENCES

Ahad, N.A., and Yahaya, S.S.S. (2014). Proceedings of the 21st National Symposium on Mathematical Sciences (SKSM21). AIP Conf. Proc. **1605**:888–893. https://doi.org/10.1063/1.4887707.

Akakpo, R., Carpentier, M.C., Hsing, Y.I., and Panaud, O. (2020). The impact of transposable elements on the structure, evolution and function of the rice genome. New Phytol. **226**:44–49.

Allaire, J. (2012). RStudio: integrated development environment for R. Kaleidoscope lc **770**:394.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., and Ciren, D. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell **182**:145–161. e123.

Amiri, R., Bahraminejad, S., and Jalali-Honarmand, S. (2013). Effect of terminal drought stress on grain yield and some morphological traits in 80 bread wheat genotypes. Int. J. Agric. Crop Sci. **5**:1145.

Bader, J.S. (2001). The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics **2**:11–24.

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature **441**:87–90.

Bembom, O., and Ivanek, R. (2021). seqLogo: Sequence logos for DNA sequence alignments (R package version 1.60.0.). https://bioconductor.org/packages/release/bioc/html/seqLogo.html.

Bevan, M.W., Uauy, C., Wulff, B.B., Zhou, J., Krasileva, K., and Clark, M.D. (2017). Genomic innovation for crop improvement. Nature **543**:346–354.

Braz, C.U., Taylor, J.F., Bresolin, T., Espigolan, R., Feitosa, F.L., Carvalheiro, R., Baldi, F., Lucia, G., and De Oliveira, H.N. (2019). Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. BMC Genet. **20**:1–12.

Bukszár, J., and van den Oord, E.J. (2010). Estimating effect sizes in genome-wide association studies. Behav. Genet. **40**:394–403.

Bush, W.S., and Moore, J.H. (2012). Genome-wide association studies. PLoS Comput. Biol. **8**:e1002822.

Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am. J. Hum. Genet. **86**:6–22.

Carpentier, M.-C., Manfroi, E., Wei, F.-J., Wu, H.-P., Lasserre, E., Llauro, C., Debladis, E., Akakpo, R., Hsing, Y.-I., and Panaud, O. (2019). Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nat. Commun. **10**:1–12.

Chen, J., Lu, L., Benjamin, J., Diaz, S., Hancock, C.N., Stajich, J.E., and Wessler, S.R. (2019). Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. Nat. Commun. **10**:1–10.

Chen, J., Wrightsman, T.R., Wessler, S.R., and Stajich, J.E. (2017). RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. PeerJ **5**:e2942.

Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics **Chapter 4**, Unit 4.10.

De Hoff, P.L., Brill, L.M., and Hirsch, A.M. (2009). Plant lectins: the ties that bind in root symbiosis and plant defense. Mol. Genet. Genomics **282**:1–15.

De Wit, P., Pespeni, M.H., and Palumbi, S.R. (2015). SNP genotyping and population genomics from expressed sequences–current advances and future possibilities. Mol. Ecol. **24**:2310–2323.

DeGiorgio, M., Huber, C.D., Hubisz, M.J., Hellmann, I., and Nielsen, R. (2016). SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics **32**:1895–1897.

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M.B., and Hirsch, C.N. (2021). How the pan-genome is changing crop genomics and improvement. Genome Biol. **22**:1–19.

Doebley, J., Stec, A., and Hubbard, L. (1997). The evolution of apical dominance in maize. Nature **386**:485–488.

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J.M., Colot, V., and Quadrana, L. (2020). The impact of transposable elements on tomato diversity. Nat Commun. **11**:4058, Erratum in: Nat Commun. 2021 May 21;12(1):3203. PMID: 32792480; PMCID: PMC7426864. https://doi.org/10.1038/s41467-020-17874-2.

Endelman, J.B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome **4**:250–255.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint, arXiv:1207.3907.

Grennan, A.K. (2011). To thy proteins be true: RNA editing in plants. Plant Physiol. **156**:453–454.

Gross, B.L., and Olsen, K.M. (2010). Genetic perspectives on crop domestication. Trends Plant Sci. **15**:529–537.

Hayashi, K., and Yoshida, H. (2009). Refunctionalization of the ancient rice blast disease resistance gene pit by the recruitment of a retrotransposon as a promoter. Plant J. **57**:413–425.

Hénaff, E., Zapata, L., Casacuberta, J.M., and Ossowski, S. (2015). Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. BMC Genomics **16**:768.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. **6**:95–108.

Huang, D., Zheng, Q., Melchkart, T., Bekkaoui, Y., Konkin, D.J., Kagale, S., Martucci, M., You, F.M., Clarke, M., and Adamski, N.M. (2020). Dominant inhibition of awn development by a putative zinc-finger transcriptional repressor expressed at the B1 locus in wheat. New Phytol. **225**:340–355.

Iqbal, N., Khan, N.A., Ferrante, A., Trivellini, A., Francini, A., and Khan, M. (2017). Ethylene role in plant growth, development and senescence: interaction with other phytohormones. Front. Plant Sci. **8**:475.

Jacques, P.-E., Jeyakani, J., and Bourque, G. (2013). The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. **9**:e1003504.

Keane, T.M., Wong, K., and Adams, D.J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics **29**:389–390.

Kiyosawa, S. (1972). Inheritance of blast resistance transferred from some indica varieities in rice. Tokyo Nat. Inst. Agr. Sci. Bull. Ser. D.

Kofler, R., Betancourt, A.J., and Schlötterer, C. (2012). Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. **8**:e1002487.

Koh, S., Lee, S.-C., Kim, M.-K., Koh, J.H., Lee, S., An, G., Choe, S., and Kim, S.-R. (2007). T-DNA tagged knockout mutation of rice OsGSK1, an orthologue of Arabidopsis BIN2, with enhanced tolerance to various abiotic stresses. Plant Mol. Biol. **65**:453–466.

Lai, X., Schnable, J.C., Liao, Z., Xu, J., Zhang, G., Li, C., Hu, E., Rong, T., Xu, Y., and Lu, Y. (2017). Genome-wide characterization of non-reference transposable element insertion polymorphisms reveals genetic diversity in tropical and temperate maize. BMC Genomics **18**:702.

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., Liu, C., Nick, P., Du, F., and Fan, P. (2019). Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. Nat. Commun. **10**:1–12.

Linheiro, R.S., and Bergman, C.M. (2012). Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. PLoS One **7**:e30008.

Lisch, D. (2013). How important are transposons for plant evolution? Nat. Rev. Genet. **14**:49–61.

Liu, H.J., and Yan, J. (2019). Crop genome-wide association study: a harvest of biological relevance. Plant J. **97**:8–18.

Ma, H. (2003). Plant reproduction: GABA gradient, guidance and growth. Curr. Biol. **13**:R834–R836.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S.C., and Finn, R.D. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. **47**:W636–W641.

Mammadov, J., Aggarwal, R., Buyyarapu, R., and Kumpatla, S. (2012). SNP markers and their impact on plant breeding. Int. J. Plant Genomics **2012**:728398.

Moose, S.P., and Mumm, R.H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiol. **147**:969–977.

Muñoz-López, M., and García-Pérez, J.L. (2010). DNA transposons: nature and applications in genomics. Curr. Genomics **11**:115–128.

Nelson, M.G., Linheiro, R.S., and Bergman, C.M. (2017). McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. G3: Genes, Genomes Genet. **7**:2763–2778.

Ou, S., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. **176**:1410–1422.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., and Peterson, T. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. **20**:1–18.

Pages, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2016). Biostrings: string objects representing biological sequences, and matching algorithms. R. Package Version **2**, 10.18129.

Platzer, A., Nizhynska, V., and Long, Q. (2012). TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. Biology **1**:395–410.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics **21**:i351.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., and Daly, M.J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81**:559–575.

Qi, J., Qian, Q., Bu, Q., Li, S., Chen, Q., Sun, J., Liang, W., Zhou, Y., Chu, C., and Li, X. (2008). Mutation of the rice Narrow leaf1 gene, which encodes a novel protein, affects vein patterning and polar auxin transport. Plant Physiol. **147**:1947–1959.

Qiu, L.-J., Xing, L.-L., Guo, Y., Wang, J., Jackson, S.A., and Chang, R.-Z. (2013). A platform for soybean molecular breeding: the utilization of core collections for food security. Plant Mol. Biol. **83**:41–50.

Robb, S.M., Lu, L., Valencia, E., Burnette, J.M., Okumoto, Y., Wessler, S.R., and Stajich, J.E. (2013). The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. G3: Genes Genomes Genet. **3**:949–957.

Roy, N.S., Choi, J.-Y., Lee, S.-I., and Kim, N.-S. (2015). Marker utility of transposable elements for plant genetics, breeding, and ecology: a review. Genes Genomics **37**:141–151.

Ryu, C.H., Lee, S., Cho, L.H., Kim, S.L., Lee, Y.S., Choi, S.C., Jeong, H.J., Yi, J., Park, S.J., and Han, C.D. (2009). OsMADS50 and OsMADS56 function antagonistically in regulating long day (LD)-dependent flowering in rice. Plant Cell Environ. **32**:1412–1427.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.-C., Iwamoto, M., and Abe, T. (2013). Rice

Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. **54**:e6.

Smit, A., and Hubley, R.R. (2018). Open-1.0. 2008–2015 (Seattle: Institute for Systems Biology).

Stuart, T., Eichten, S.R., Cahn, J., Karpievitch, Y.V., Borevitz, J.O., and Lister, R. (2016). Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. elife **5**:e20777.

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. **43**:1160.

Tilman, D., Reich, P.B., Knops, J., Wedin, D., Mielke, T., and Lehman, C. (2001). Diversity and productivity in a long-term grassland experiment. Science **294**:843–845.

Trizzino, M., Park, Y., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G.H., Lynch, V.J., and Brown, C.D. (2017). Transposable elements are the primary source of novelty in primate gene regulation. Genome Res. **27**:1623–1633.

Wan, J., Patel, A., Mathieu, M., Kim, S.-Y., Xu, D., and Stacey, G. (2008). A lectin receptor-like kinase is required for pollen development in Arabidopsis. Plant Mol. Biol. **67**:469–482.

Wang, H., Cimen, E., Singh, N., and Buckler, E. (2020). Deep learning for plant genomics and crop improvement. Curr. Opin. Plant Biol. **54**:34–41.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., and Panaud, O. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**:973–982.

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., and Huang, L. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. **30**:105.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**:W265–W268.

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., Zhou, H., Yu, S., Xu, C., and Li, X. (2008). Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. Nat. Genet. **40**:761.

Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L., Huang, J., Deng, T., Luo, J., and He, L. (2019). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat. Genet. **51**:1052–1059.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.-C., Hu, L., Yamasaki, M., Yoshida, S., Kitano, H., and Hirano, K. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat. Genet. **48**:927.

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., and Parakkal, P. (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. Sci. Data **7**:1–11.

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., and Gaut, B.S. (2019). The population genetics of structural variants in grapevine domestication. Nat. Plants **5**:965–979.

Zhuang, J., Wang, J., Theurkauf, W., and Weng, Z. (2014). TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res. **42**:6826–6838.

Zúñiga-Sánchez, E., Soriano, D., Martínez-Barajas, E., Orozco-Segovia, A., and Gamboa-deBuen, A. (2014). BIIDXI, the At4g32460 DUF642 gene, is involved in pectin methyl esterase regulation during Arabidopsis thalianaseed germination and plant development. BMC Plant Biol. **14**:338.