Evolutionary Biology, Genetics and Genomics

Transposons are a major contributor to gene expression variability under selection in rice populations

Raúl Castanera ✉, **Noemia Morales-Díaz, Sonal Gupta, Michael Purugganan, Josep M. Casacuberta** ✉

Center for research in agricultural genomics, CRAG (CSIC-IRTA-UAB-UB), campus UAB, Cerdanyola del Vallès, 08193 Barcelona, Spain. • Center for Genomics and Systems Biology, 12 Waverly Place, New York University, New York, NY 10003, USA • Center for Genomics and Systems Biology, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

Abstract

Transposable elements are an important source of genome variability. Here, we analyze their contribution to gene expression variability in crops by performing a TE insertion polymorphism (TIP)-eQTL mapping using expression data from rice varieties from the *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica* subspecies. Our data shows that TE insertions are associated with changes of expression of many genes known to be targets of rice domestication and breeding. An important fraction of these insertions were already present in the rice wild ancestors, and have been differentially selected in indica and japonica rice populations. Taken together, these data show that small changes of expression in signal transduction genes induced by TE insertions accompany the domestication and adaptation of rice populations.

> **eLife assessment**
>
> This **valuable** study reports on the role of transposable elements in gene expression variation in rice and how TE-associated expression changes could have been selected during domestication. The combination of evidence from linkage studies and selection scans for a subset of insertions is **convincing**, although it is currently unclear how large the contribution of TEs to gene expression variability compared to other variants is, which should be reflected in the title. The work will be of interest to colleagues working on the role of transposable elements in adaptation and to biologists working on domestication.

Introduction

Transposable elements (TEs) are a major component of eukaryotic genomes, particularly in plants [1]. Their ability to move, creating mutations by insertion/excision, and amplify in

genomes - thus generating new copies that provide opportunities for recombination - make them a major source of genetic variability [2]. For this reason TEs are considered a major driver of plant genome evolution, both in the wild or under human selection [3]. TE insertions and other structural variants, however, have been frequently overlooked when using genome wide association studies (GWAS) to look for genetic variants linked to interesting phenotypes [4]. The recent development of efficient tools to identify transposon insertion polymorphisms (TIPs) [5] has allowed the incorporation of TIPs in these analyses, and several GWAS reports have shown that TIPs uncover additional variability linked to phenotypic traits [6]–[8]. Interestingly, TIPs often explain more phenotypic variance compared to SNPs, and can be used for genomic prediction [9]. The reason for this could be that, as compared with SNPs, TE insertions are more frequently the causal mutation.

A major fraction of the mutations linked to crop domestication and breeding are associated with changes of the expression of genes involved in signal transduction [10],[11]. TEs can alter gene transcription, activating or repressing them, by different means. On the one hand, insertion of TEs within a promoter can interfere with transcription, and the silencing of TEs inserted close to genes can result in repression of gene expression [12]–[14]. Alternatively, TE insertions can also result in gene overexpression [15],[16], as TEs bring their own transcriptional promoters that can induce expression of neighboring genes. This is particularly clear for LTR-retrotransposons which contain promoters in their 5' and 3' LTRs, and may provide nearby genes with alternative promoters. In addition, TEs can also contain transcription factor binding sites (TFBS) that can alter the expression of host genes [17]; it has been shown, for example, that MITEs have frequently amplified and redistributed TFBS in plant genomes [18].

Here we explore the impact of the movement of TEs during the recent evolution of rice on the variability of gene expression in this species. We took advantage of a recently published transcriptional analysis of 208 rice varieties grown under wet paddy and intermittent drought conditions [19] to perform an expression quantitative trait locus (eQTL) GWAS using TIPs as genetic markers. We show that TIPs are frequently associated with changes of expression of rice genes and that many TIPs altering the expression of regulatory genes have been positively selected in indica and japonica rice subspecies.
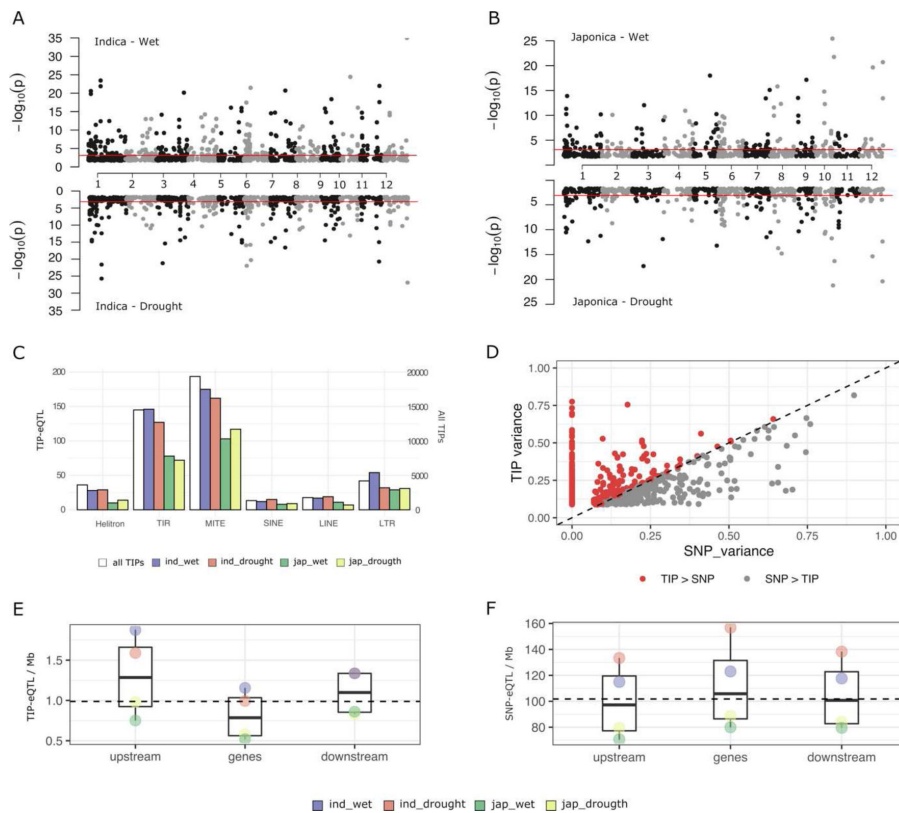
Characterization of rice TIP-eQTLs.

A-B, Association between TIPs and gene expression levels in *cis* (TIP-eQTLs) found in indica and japonica population along the twelve rice chromosomes. Horizontal lines represent 5% significance thresholds corrected for multiple testing using FDR method. C, Number of TIP-eQTLs found in each condition classified at the order TE level. White bars represent the number of total TIPs (secondary axis). D, Expression variance explained by leading TIP-eQTL and SNP-eQTL. Each point represents a gene. Indica and Japonica TIP-eQTLs are combined into a single plot. Number of TIP-eQTLs (E) or SNP-eQTL (F) per Mb of gene feature. Upstream and downstream regions are 1 Kb long. Gene region includes 5' and 3' UTRs. All the TIPs and SNPs with a significant association (FDR < 5%) were used in this analysis.

Results

TIPs are associated with gene expression variation in rice

We used a recently published dataset containing genome resequencing and transcription data for rice (*Oryza sativa*) varieties to perform TIP-eQTL mapping analyses. Rice is one of the most important food crops in the world, and varieties are generally found in two subspecies – *O. sativa* ssp. *japonica* and ssp *indica*. Other minor variety groups include aus varieties, which appear closely related to indica, and basmati rice varieties that appear to be a hybrid between ssp. *japonica* and aus. The data we analyzed consists of an 'indica' dataset for 126 *O. sativa* ssp. *indica* and some *circum*-aus accessions, and 84 'japonica' accessions comprising of *O. sativa* ssp. *japonica* and some *circum*-basmati varieties, described in Groen et al. (2020) [19]. We predicted TIPs using Popoolation TE2 [20], and identified 45,050 TIPs. Using the Minghui MH63 short read data and assembled draft reference genome [21] we estimated the performance of the TIP genotyping on this dataset to have 97.1% sensitivity and 92.2% precision (Supplementary text 1). Most of the TIPs detected are related to DNA transposons (84%), especially MITEs belonging to the *Stowaway* (24%) and *Tourist* (20%) superfamilies; LTR retrotransposons make up a minor fraction (9 %).

The 45,050 TIPs are distributed along the 12 chromosomes and 81.3 % of them are less than 5 kb away from an annotated gene (IRGSP RAP-DB gene models). This results in up to 81.4% of rice genes having a TIP less than 5 Kb distant. In order to analyze the potential of TIPs to

generate gene expression variability in rice, we used the 3' mRNA-seq expression data from leaves of adult plants grown in normally-watered soil ("wet" condition) and under intermittent drought-stress ("drought stress" condition) [19]. We selected data from 15,549 genes showing expression in more than 10% of the samples. Due to the strong population structure present in rice we separately analyzed data in the indica (126 varieties) and japonica (82 varieties) datasets. We found a total of 563 significant associations between TIPs and gene expression levels in the indica population and 356 in the japonica population (TIP-eQTLs with MAF > 3%, FDR adjusted p value < 0.05 in at least one replicate) (Figure 1A and 1B, Supplementary Table 2). These involve 477 and 317 genes, in indica and japonica respectively, representing 3.1% and 2% of the expressed genes. The TIP-eQTLs were related to the different TE types (orders [22]), and the proportion of TE orders found in the significant TIPs did not differ significantly from their representation across the genome (Figure 1C).

In order to estimate the relative contribution of TIPs and SNPs to gene expression changes, we also performed eQTL mapping analyses using 1 million of randomly chosen SNPs as markers [19]. We obtained 4,913 SNP-eQTLs in indica and 3,308 SNP-eQTLs in japonica associated with SNPs located less than 5 kb from the gene; each SNP-eQTL represents the leading SNP-gene association. This corresponds to ~31.6% of the genes in indica and ~21.3 % in japonica. For each gene with a significant association, we compared the proportion of variance (R^2) explained by the most significantly associated TIPs and/or SNPs. For most genes that have an associated TIP, either there is no associated SNP or the TIP explains more expression variance than the associated SNP (~62% of the indica TIP-eQTLs and ~73% of the japonica TIP-eQTLs) (Figure 1D). This demonstrates that TIPs uncover genetic associations with changes of expression that are not observed when using SNPs as molecular markers.

The SNP-eQTLs found in our study are associated to both positive (47%) and negative (53%) effects on gene expression levels, very close to an even distribution. In contrast, TIPs more frequently (59%) have a negative effect. The proportion varies among TE superfamilies, with TIR/MULEs being close to 50%, and LTR/Gypsy having the most frequently negative impact on gene expression (67%, Supplementary Table 1). This suggests that, as compared with SNP-eQTLs, TIP-eQTLs are more frequently the likely causal mutation responsible for the change in expression, and that the effect of the insertion is more generally negative, in line with recent analyses done in *Capsella*[23], especially for long elements such as LTR-retrotransposons.

An analysis of the TIP and SNP-eQTL location with respect to the associated genes shows that while SNP-eQTLs are more evenly distributed across the genic region (Figure 1F), TIP-eQTLs are over-represented in the 1 kb region upstream of the gene (Figure 1E), which is the region that most frequently contains promoter elements regulating transcription. This suggests that although some TIP-eQTLs could be in linkage disequilibrium with the causal mutation for the change of expression, an important fraction of TIP-eQTLs could be the actual causal mutation inducing expression variation.

It has been proposed that an important fraction of gene expression variation is deleterious, and therefore, alleles associated with major expression effects should be maintained at lower frequencies [24],[25]. An analysis of the effect size of the different TIP-eQTLs taking into account their frequency in the population shows that, indeed, low and high-frequency TIPs (corresponding to rare alleles) show the highest positive or negative effect on the expression of the associated gene (Figure 2 A-D, Supplementary Figure 1). This is consistent with what is described as the rare variant effect [25] where rare (and likely deleterious) mutations in a population are associated with greater effects on gene expression in a population.
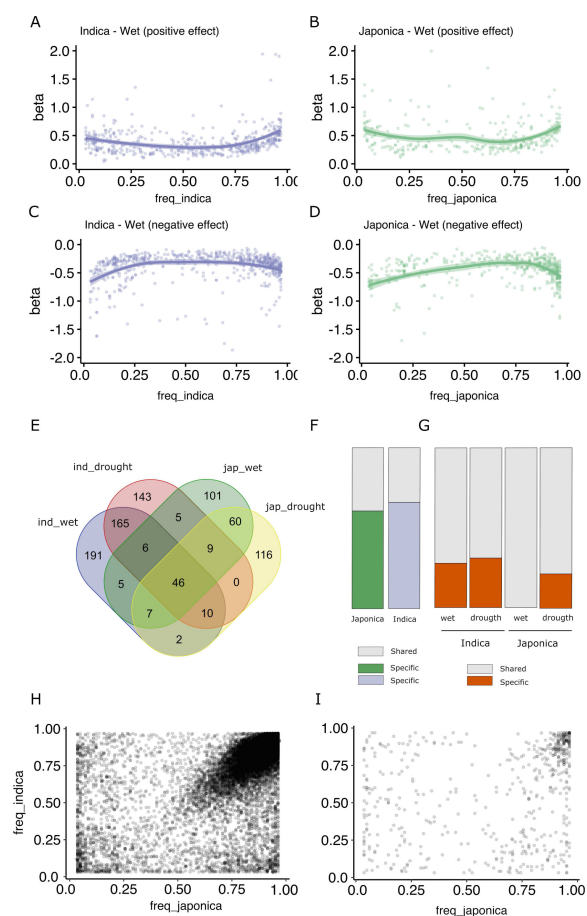
Figure 2

Effect size and population frequencies of wet and drought-stress TIP-eQTLs. representation of the effect size (beta) of the TIP-eQTL with respect to the frequency of the insertion TIP-eQTLs in indica (A, C) and japonica (B,D) for positive (A,B) and negative (C,D) effects.

Venn diagram illustrating the intersection between TIP-eQTLs detected in each condition of the two subspecies analyzed, indica (ind), and japonica (jap) (E). Percentage of subspecies-specific TIP-eQTLs, considering as shared the TIPs that fall in the intersection between all the FDR-corrected TIP-eQTLs found in a given population and all the associations of the other population using a relaxed cutoff ($p < 0.05$, no FDR correction) (F). Percentage of condition-specific TIP-eQTLs considering as shared the TIPs that fall in the intersection between all the FDR-corrected TIP-eQTLs found in a given population and all the associations of the other population using a relaxed cutoff ($p < 0.05$, no FDR correction. Relationship between the frequencies in indica and japonica populations of the 33,389 non-eQTL TIPs (H) and 829 TIP-eQTLs (I).

A comparison of TIP-eQTLs associated with changes of expression in the two different growth conditions (wet and drought stress) shows that there is an important overlap of eQTLs between the two environments (38.5 % for indica TIP-eQTLs and 33.2% for japonica TIP-eQTLs) (Figure 2E). Such overlap increases to 90% (average of indica and japonica, wet) and 71 % (average of indica and japonica, drought stress) when we compare TIP-eQTLs of one condition with a non-FDR-corrected TIP-eQTLs in the other environment (Figure 2G). This result suggests that the majority of TIP-eQTLs are associated with changes of expression in both growth conditions, and that only a small number of TIP-eQTLs (81 in indica and 78 in japonica) are associated with stress-specific changes of expression. Nevertheless, among the TIP-eQTLs associated with changes of gene expression in both wet and drought stress, there are TIP associated with genes known to be involved in drought tolerance (Supplementary Table 2). As an example, TIP_49046 is associated with a reduction of the expression of the gene synaptotagmin-5 (*OsSYT-5*). *OsSYT-5* encodes a Ca^{2+} sensing protein with a C2 domain, is expressed in both stressed and non-stressed plants, and it has been recently shown that its silencing enhances drought tolerance in rice [26]. Other examples are TIP_23764 and TIP_52367, associated with an increased expression of the gene encoding the OsSAPK10 ABA-activated protein kinase and the S-type euonymus-related lectin gene OsEULS2 in respectively, two genes known to mediate drought stress in rice [27],[28]

TIP-eQTLs are present at different population frequencies in indica and japonica

Although the overlap of TIPs associated with changes in expression under wet and drought stress conditions is very high, the overlap between indica and japonica TIP-eQTLs is low, and

only 90 TIP-eQTLs out of the 563 from indica and 356 from japonica are significantly associated with variation of gene expression in both subspecies (Figure 2E). Even when we make the less-stringent comparison of the TIP-eQTLs of one subspecies with the non-FDR-corrected TIP-eQTLs of the other, as much as ~69% of indica and ~60% of japonica TIP-eQTLs appear as subspecies-specific (Figure 2F). This result may suggest differences in the transcriptional networks in indica and japonica or, alternatively, that different alleles showing expression level differences may be differentially represented in the two subspecies. An analysis of the frequencies in indica and japonica of all the TIP-eQTLs described here shows that approximately one-third of the TIPs are only found in one of the two subspecies. This suggests that TEs have been actively inserting (or have been excised or eliminated from the population) very recently during rice evolution, likely post-domestication and during the crop diversification phase, as has already been proposed for some rice TE families [29],[30]. Here we show that some of these recent TIPs are associated with changes in gene expression and may therefore have phenotypic consequences. We found 44 TIP-eQTLs corresponding to TIPs present only in japonica varieties and 64 TIP-eQTL corresponding to TIPs present only in indica varieties. Some of these insertions are found at high frequencies in its corresponding subspecies, which suggest that they may have been under positive selection (Supplementary Table 3).

However, two thirds (66%) of all TIP-eQTLs found in this study correspond to TIPs present in both subspecies, which suggests that these insertions are relatively old and were already present in the ancestor of indica and japonica rice. Remarkably, in most cases they are found at very different frequencies in the two subspecies (Figure 2I). This suggests that these insertions were already present in the ancestor of indica and japonica rice and have been differentially retained in the two subspecies. Interestingly, an analysis of all non-eQTL-TIPs shows that an important fraction is found at the same frequency in both genomes (Figure 2H), suggesting that the preferential retention of TIP-eQTLs in indica or japonica may be a specific phenomenon potentially linked to their impact on the expression of the associated genes.

We examined whether TIP-eQTLs present in both indica and japonica varieties are also found in either *Oryza rufipogon* and *Oryza nivara*, the wild rice relatives believed to be the ancestors of domesticated rice [31]. We looked for the presence of the TIPs identified in rice in a set of 72 accessions of *O. rufigogon* and 10 accessions of *O. nivara* (collectively called rufipogon/nivara from now on) (Supplementary Table 4). Up to 552 of the 829 TIP-eQTLs present in indica and/or japonica (66%), can be found in rufipogon/nivara varieties, confirming that an important fraction of TIP-eQTLs found in indica and japonica may be relatively old insertions already present in the wild ancestors of domesticated rice. Not surprisingly, in most cases the TIP frequencies in rufipogon/nivara are different from the frequencies of these insertions in indica and/or japonica (Supplementary Table 5); these may have arisen either due to selection or genetic drift accompanying the bottleneck associated with crop evolution.

In order to identify possible selection of TIPs in indica and japonica populations, we used the Population Branch Statistic (PBS) method [32] which examines strong differentiation in frequencies between populations. This approach consists in comparing the three pairwise FST values between indica, japonica and rufipogon/nivara to estimate the frequency change in TIPs that occurred since the divergence of the two rice subspecies from its wild ancestor. We calculated PBS values for a total of 11,698 TIPs present in the japonica, indica and rufipogon/nivara populations, which included 354 TIPs that are eQTLs in in indica and/or japonica populations. TIP-eQTLs (Figure 3A) showed much higher PBS values in comparison to those not identified as eQTLs (Figure 3B) (mean absolute PBS of 0.06 vs 0.02, respectively, $p < 0.01$, Wilcoxon test). Furthermore, TIPs with the most extreme PBS values (above 95 percentile) are enriched for TIP-eQTLs (Fisher's Exact test odds ratio = 4.87, $p < 0.01$),

suggesting that a high fraction of TIP eQTLs have been positively selected in indica or japonica. Note that some of the TIPs not identified as eQTLs in this particular dataset may actually be linked to variation of gene expression in organs, developmental stages or environmental conditions different from the ones analyzed here. Interestingly, up to 35% of the TIP-eQTLs identified here, associated with 119 genes, fell within the top 10% of absolute PBS values for all TIPs (Fig 3A), suggesting that TIPs associated with gene expression variation are likely under differential selection during rice evolution.
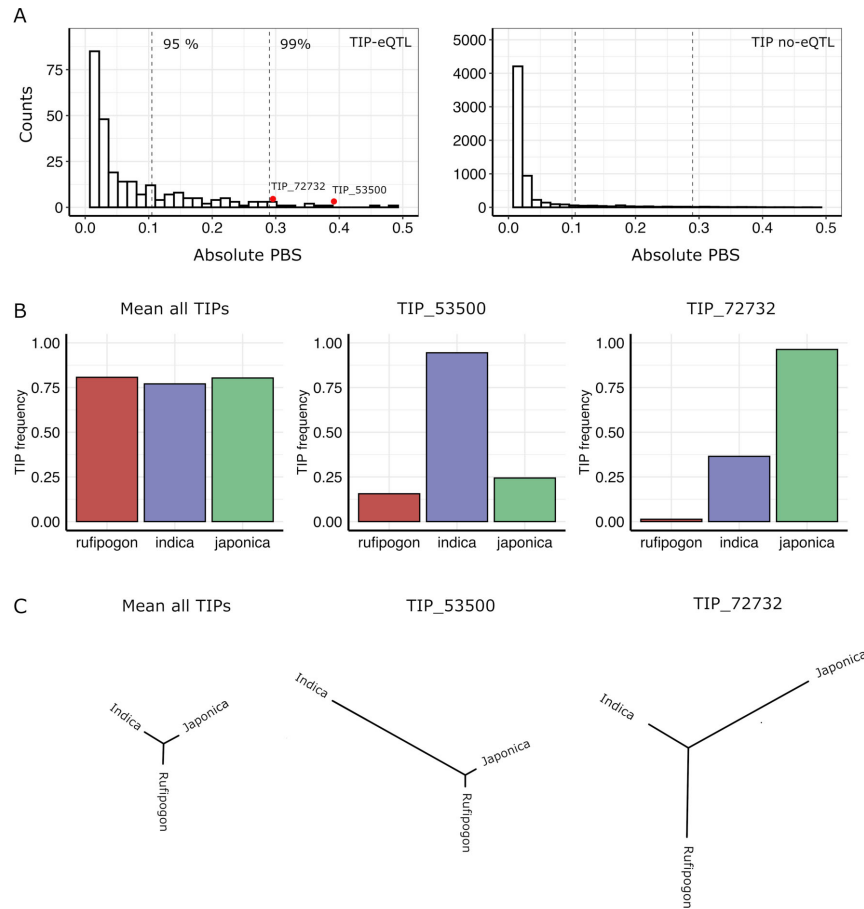

Figure 3

Signatures of positive selection on TIP-eQTLs.

A, Absolute PBS of 354 TIP-eQTLs (left) and 11,344 TIPs (no-eQTL, right) present in indica, japonica and rufipogon populations. Dotted vertical lines represent the 95th an 99th percentile of the PBS values of the whole dataset (11,698 TIPs). Red dots represent two examples of TIP-eQTLs with extreme PBS values. B) Population frequency of the two TIPs with extreme PBS values, marked as red dots in the left panel A, as well as of the whole TIP dataset. C) Fst-based tree of the two TIPs with extreme PBS values, as well as of the whole TIP dataset (average Fst).

We observe that TIPs can show evidence of selection in japonica, indica or both subspecies. As examples, Figure 3B and C shows the representation of the PBS analysis, and the frequency in the populations, of two TIP-eQTL whose frequency has greatly increased in indica (TIP_53500), or japonica (TIP_72732), compared with the mean PBS for all TIPs.

Gene variants selected during rice evolution: some examples

Genes linked with TIP-eQTLs showing extreme indica or japonica PBS metrics are good candidates for genes underlying adaptation between different rice subspecies. Among the genes with TIP-eQTLs that have extreme PBS values in indica or japonica we can find several examples that are known to regulate plant architecture, plant and grain development or abiotic stress responses.

Some of the most extreme PBS values are for TIPs associated with changes of expression of genes involved in the signal transduction of hormones, including brassinosteroids, ABA, ethylene, jasmonic acid (JA) and auxin (Supplementary Table 5). We found four different TIP-

eQTLs with high PBS values associated with the EG2/OsJAZ1 gene (Os04g0653000), a locus encoding a JA signaling repressor [33]; the four insertions are related to MITEs of the Tourist transposon superfamily. Three of the insertions (TIP_32891, TIP_32892 and TIP_32894), located upstream (3.4 kb and 2.7 kb) and downstream (150 bp) of the gene, are associated with an increase of expression of *EG2*, and are physically linked to one another (mean r^2 value = 0.88). In contrast, the fourth insertion (TIP_32893), is not linked to them, is within the first intron of the gene and is associated with expression reduction.

The recent advances in characterizing the pangenome of rice and the super-pangenome that includes its wild ancestors [34]–[36] has allowed us to characterize the locus in indica, japonica and rufipogon/nivara. This analysis showed that *EG2* is primarily found in two different haplotypes (Figure 4A). One haplotype, Hap A, contains the three MITE insertions associated with higher expression of EG2 (Figure 4A, 4B, and Supplementary Figure 2A), while the second haplotype, Hap B, contains the fourth MITE insertion and is associated with reduced EG2 expression (Figure 4A, 4B and Supplementary Figure 2A). Hap A was identified to be at high frequency in rufipogon/nivara (48%), whereas Hap B is at lower frequency (16%). The frequency of both Hap A and Hap B is slightly increased in japonica (65% and 28%, respectively). In indica, however, there is a clear reduction in the frequency of Hap A and a concomitant strong frequency increase of Hap B, associated with reduced expression in indica (16% and 83%, respectively). Interestingly, EG2/OsJAZ1 is a repressor of spikelet development [33] and the number of differentiated spikelets per panicle tends to be lower in japonica as compared with indica [37].

Figure 4

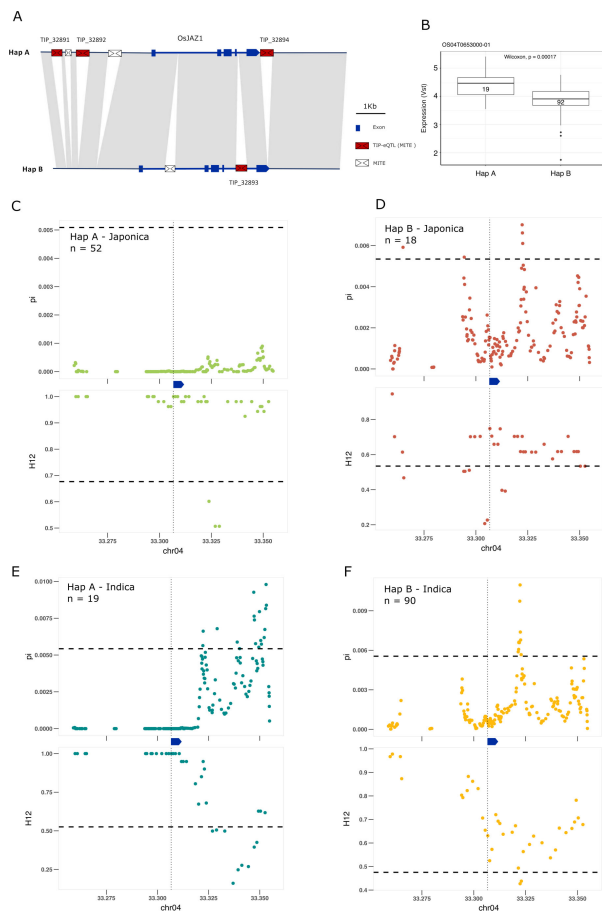## Selection on TIP-eQTLs associated with *EG2* expression.

A, Representation of the two main *EG2* haplotypes present in rice and rufipogon/nivara populations identified in the rice super-pangenome. Conserved nucleotide regions are connected by gray marks. Structural variants longer than 50bp are shown as white spaces. TIP-eQTLs are shown as red boxes. Additional TIPs are shown as white boxes. B, Boxplot representation of the expression of the two *EG2* haplotypes in the indica population. Numbers inside boxplots represent the number of accessions in each group. Analysis of nucleotide diversity (π) and diversity in haplotypes homozygosity tracts (H12) for Hap A (C, E) and Hap B (D, F) in japonica (C, D) and indica (E, F) populations. The vertical dotted black line shows the position of the TIP insertion. the EG2 gene is schematically shown in blue. The horizontal dashed black line represents the mean of 1000 random permutation pulls (for p-value see Supplemental Table 6).

We looked more closely for signs of selection at this specific locus, by examining levels of nucleotide diversity (π[38]) and diversity in haplotypes homozygosity tracts (H12 [39]).

Consistent with the expression (and possible spikelet phenotypic differences), we found that in japonica the haplotype associated with increased *EG2* expression (Hap A) was under positive selection, whereas the haplotype associated with decreased expression of *EG2* (Hap B) did not show any sign of selection; this indicates that positive selection acts on individuals carrying the haplotype for high expression of *EG2* in japonica (Figure 4C-D). Specifically, we observe significantly lower π and higher H12 in Hap A, associated with increased *EG2* expression, while this pattern was not significant for Hap B (supplementary Table 6). However, in indica the opposite pattern appears to prevail, wherein there is a stronger evidence of selection for Hap B, but less so for Hap A (Figure 4E-F). This suggests positive selection towards higher expression of *EG2* in japonica, and lower expression of *EG2* in indica, which could potentially explain the higher number of spikelets per panicle observed in the latter[37].

Another good candidate for genes underlying adaptation of rice is the *OsGAP* gene. There are two different TIP-eQTLs with high PBS values associated with changes of expression of the *OsGAP* gene (Os07g0500300) in japonica, whereas they do not correlate with changes in expression in indica (Supplementary Figure 2B). *OsGAP* encodes a putative GTPase activating protein, similar to CAR proteins (C2-domain abscisic acid-related proteins), that play an important role in ABA signal transduction in Arabidopsis [40]. The first TIP (TIP_50057) is located ~4 kb upstream of the OsGAP gene and is associated with an increase of expression of the gene in japonica whereas the second insertion (TIP_50059) is located within the first intron and is associated with a decrease in expression in japonica (Figure 5A-B). The analysis of the locus using the super-pangenome data shows that *OsGAP* is found in two different haplotypes, one containing TIP_50057 (Hap A) and the other containing TIP_50059, together with some additional structural differences (Hap B; Figure 5A). The two TIPs defining the two haplotypes are present in the rufipogon/nivara population at complementary frequencies (73% and 19%, respectively). Interestingly, the proportion of the two haplotypes in cultivated rice seem to be very different, with an increased frequency of the haplotype associated with a reduced expression of *OsGAP* (Hap B) in both japonica (~50%) and indica (~95%) were it reaches near fixation.
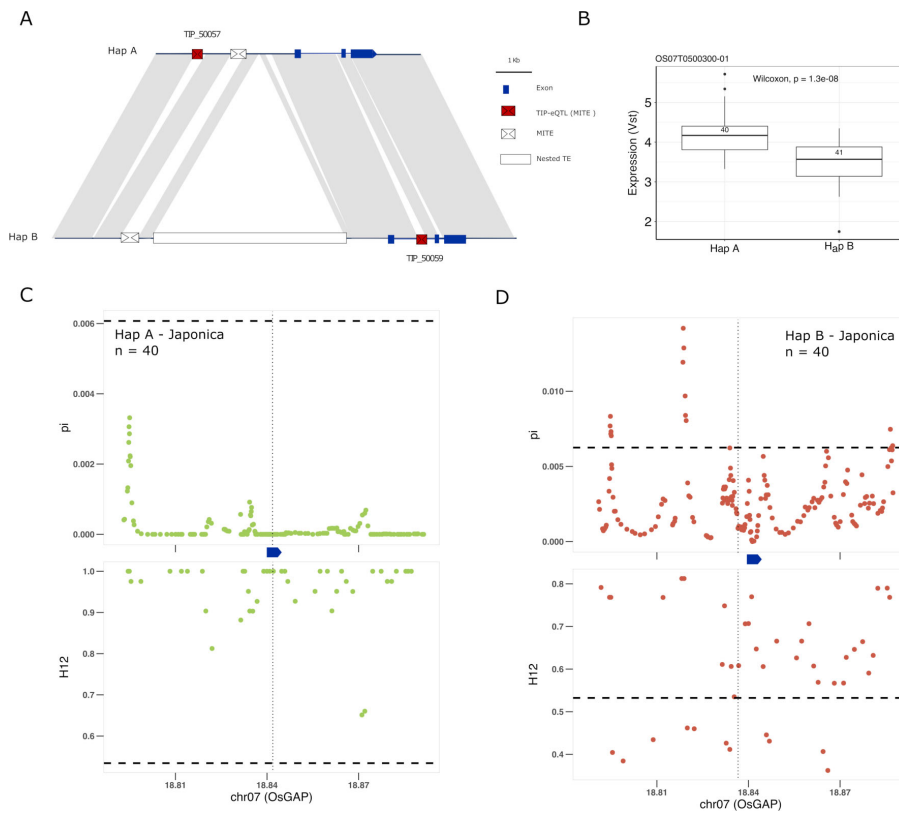
Selection on TIP-eQTLs associated with *OsGAP* expression.

A, Representation of the two main *OsGAP* haplotypes present in rice and rufipogon/nivara populations identified in the rice super-pangenome. Conserved nucleotide regions are connected by gray marks. Structural variants longer than 50bp are shown as white spaces. TIP-eQTLs are shown as red boxes. Additional TIPs are shown as white boxes. B, Boxplot representation of the expression of the two *OsGAP* haplotypes in the japonica population. Numbers inside boxplots represent the number of accessions in each group. Analysis of nucleotide diversity (π) and diversity in haplotypes homozygosity tracts (H12) for Hap A (C) and Hap B (D) in japonica population. The vertical dotted black line shows the position of the TIP insertion. The OsGAP gene is schematically shown in blue. The horizontal dashed black line represents the mean of 1000 random permutation pulls (for p-value see Supplemental Table 6).

It has been proposed that *OsGAP* is a negative regulator of ABA signaling in seed germination and dormancy, and that reduced expression of *OsGAP* may prevent rice pre-harvest sprouting (PHS) [40]. Reduced seed dormancy is a common target for selection of cultivated varieties, but this trait may come at a cost of a higher PHS risk, which may be a problem, especially in regions where heavy rain is common during the harvest season. Therefore, the appropriate degree of dormancy may depend on the agroecological conditions in which a particular variety is grown. We see here that two independent TIPs are associated with changes of *OsGAP* expression in japonica but not in indica (Figure 5A, Supplementary Figure 2B). Interestingly, there are clear and strong signs of selection for increased expression of *OsGAP* in japonica, wherein Hap A was under positive selection, and Hap B was selectively neutral, indicating positive selection acts on only those individuals carrying haplotype for high expression of *OsGAP* in japonica (Figure 5C,D; Supplementary Table 6). In contrast, there is no evidence for selection in indica varieties despite the near fixation of one haplotype. This suggests that an increased level of *OsGAP* may be relevant for the control of dormancy in japonica but not in indica (despite its near-fixation in the latter), which would be in line with recent data suggesting that seed dormancy is regulated by different genes/alleles in indica and japonica [41].

Finally, we find a TIP-eQTL whose frequency has greatly increased in cultivated rice with respect to rufipogon/nivara. TIP_45706, corresponding to an insertion of a gypsy-like LTR-RT at ~1 kb upstream of the *OsMPH1* gene (Figure 6A). This insertion is associated with a decrease in the expression of *OsMPH1* in indica and possibly japonica varieties (Figure 6B,

), and is found at low frequency in rufipogon/nivara (~11%) but at high frequency in both indica (70%) and japonica (95%). *OsMPH1* encodes a MYB-like transcription factor that has been shown to regulate plant height, its reduced expression resulting in shorter plants [42]. A reduction in plant height due to a mutation in the *SD1* gene, which encodes the gibberellin biosynthesis gene GA-20ox, was at the origin of the Green Revolution in the 1960s, but it has been shown that alleles of *SD1* resulting in shorter culm length were also selected during the domestication of japonica rice [43]. Our results suggest possible parallel selection of alleles in other genes that may lead to similar phenotypes. This is further reinforced by strong selection acting on the haplotype containing the TIP_45706 insertion in both indica and japonica (Figure 6C and 6DSupplementary table 6), leading to reduced expression of *OsMPH1* and thus shorter plants.
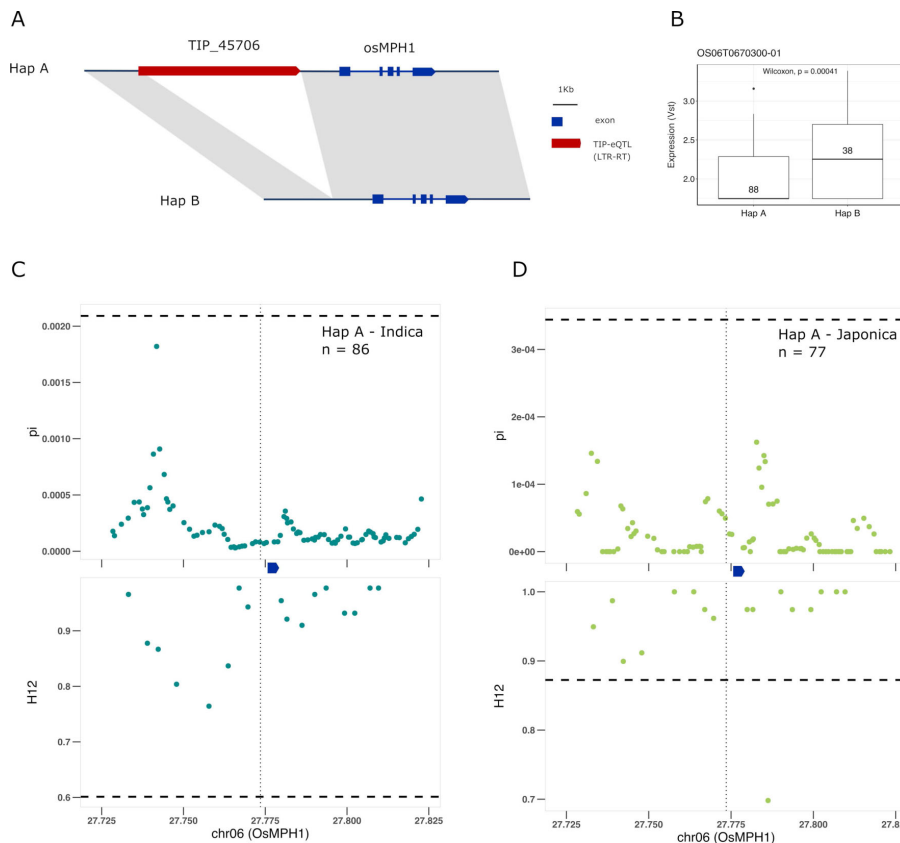

Figure 6

Selection on TIP-eQTLs associated with *OsMPH1* expression.

Representation of the two main *OsMPH1* haplotypes present in rice and rufipogon/nivara populations identified in the rice super-pangenome. Conserved nucleotide regions are connected by gray marks. Structural variants longer than 50bp are shown as white spaces. TIP-eQTLs are shown as red boxes. B, Boxplot representation of the expression of the two *OsMPH1* haplotypes in the japonica population. Numbers inside boxplots represent the number of accessions in each group. Analysis of nucleotide diversity (π) and diversity in haplotypes homozygosity tracts (H12) for Hap A in indica population (C) and Hap A (D) in japonica population. The vertical dotted black line shows the position of the TIP insertion. The OsMPH1 gene, which is schematically shown in blue. The horizontal dashed black line represents the mean of 1000 random permutation pulls (for p-value see Supplemental Table 6).

Discussion

Transposons are a major source of genome variability, and are known to affect gene expression in numerous ways. The importance of rare variants and unfixed TE insertions for gene expression variation in humans [44] and plants [23] has already been demonstrated. A recent pangenome analysis in tomato has shown that the phenotypic variation of important crop traits is linked to structural variants present in the species, most of which are actually TIPs, which are associated to subtle changes of transcription of genes involved in signal

transduction [45]. In the present study we evaluated the importance of TIP-related expression variability in the recent evolution of rice. To this end we performed a TIP-eQTL mapping using expression data from rice varieties from the *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica* subspecies. We show here that using TIPs in addition to SNPs as genetic information allows the uncovering of additional genomic associations to changes in gene expression, and that when both TIPs and SNPs are both associated, TIPs often explain more of the variance in expression. This is in line with recent reports that have also used TIPs for GWAS with different phenotypes in tomato and rice [7],[8], and suggests that TIPs may be more frequently the causal mutation of a particular phenotype than SNPs. The concentration of TIP-eQTLs in upstream regions of genes, where most transcriptional regulatory elements usually reside, also suggest that a high fraction of the TIP-eQTLs here described are the actual mutation underlying gene expression variation.

The close association of some TEs with genes could also partially explain the frequent association of more than one TIP with changes of expression of particular genes. Indeed, among the 718 genes with TIP-eQTLs described here, 18% have two of more TIPs associated with changes of expression. However, not all genes have the same number of TIPs. Among the 30 genes with TIPs in indica and japonica that have the highest population branch statistics (PBS) – an indicator of strong differential selection - 13 have more than one TIP-eQTL. This suggests that TIPs are a major source of gene expression variability in rice, in particular in genes that may have been strongly selected during the recent evolution of rice. In some cases, the different TIPs linked to a gene are associated with opposite effects on its expression and are present in different haplotypes. In these instances, as shown for the *OsJAZ1* and *OsGAP* genes, the two haplotypes have strong signs of selection, although one positively and the other negatively, pointing to role these TE insertions may play in adaptive evolution of these crop subspecies.

The results presented here show that most TIPs associated with variation of gene expression in indica and/or japonica are relatively old insertions that are also present in rufipogon/nivara, the wild ancestors of rice. This is in line with recent data showing that a high number of the structural variants associated with changes of expression in tomato were already present in its wild ancestor [45] and highlights the importance of the standing variation, already present in the wild ancestors, for crop adaptation. It is assumed that most TE insertions are selectively neutral or slightly deleterious [46], and the presence of TIPs associated with expression variation of genes in the wild ancestors of rice, suggests that this variation can be well tolerated in wild rice. Interestingly, we show here that many of these TIPs have been positive or negative selected in rice populations, which shows that they translate into selectable phenotypic differences in the agroecological conditions of cultivated rice. Indeed, we show examples of selected variants with modified expression of genes known to be linked to important traits that were targets of selection. It has been already proposed that a major fraction of the mutations linked to crop domestication and breeding are associated with changes of gene expression involved in signal transduction [10],[11]. Here we show that specific expression variants of genes involved in signal transduction have been differentially selected in indica and japonica rice populations. In addition, our results also point to TEs as a major driver of gene expression variation selected during crop adaptation and breeding.

Methods

TIP detection

Re-sequencing data for 126 indica and 82 japonica rice accessions was obtained from Groen et al [19] (Bioproject accessions PRJNA557122, PRJNA422249 and PRJEB6180). BBDuK (https://sourceforge.net/projects/bbmap/) was used for adaptor and quality trimming. Clean reads were aligned to the Nipponbare reference genome [47] using BWA [48]. PoPoolationTE2 [20] was used to detect TIPs in the mode "joint" using Nipponbare TE annotation described by Ou et al [47]. TIPs with a zygosity lower than 0.25 in all samples were excluded to avoid false positives. TIPs were further filtered using the parameters —min-count 5, --max-otherte-count 2, --max-structvar-count 2, and only those having MAF higher than 3% and no missing data were kept. Finally, the TIP matrix was transformed to binary form using zygosity cutoff of 0.05 to define an insertion as present.

TIP and SNP-eQTL mapping

Transcriptome data for the 208 rice accessions (wet and drought stress conditions; tree independent replicates per condition) was obtained from Groen el al [19] and separated into individual replicate matrices. Transcripts expressed in more than 10% of the samples on each replicate were extracted and counts were normalized using the *yst* function of DESeq2 [49]. TIP and SNP-eQTL mapping were performed using Matrix eQTL software [50], applying the simple linear regression model and including subpopulation groups as covariates. We used a cut-off distance of 5,000bp to identify *cis*-eQTLs and a 5% False Discovery Rate threshold for multiple testing corrections.

Selection analyses

For the Population Branch Statistics (PBS) analysis we looked for the presence of rice TIPs in the 82 accessions belonging to the rufipogon/nivara population (Supplementary Table 4) and retained only those present in the three populations (indica, japonica, rufipogon/nivara), resulting in a matrix of 13,622 TIPs and 288 accessions. The TIP matrix was filtered to remove TIPs with more than 5% missing data, or with MAF < 5%. The remaining missing data (1.7%) was imputed using the "wright" algorithm implemented in SNPready. The clean matrix (11,698) was used to calculate the individual PBS values per TIP, following the formulas described in Yi et al., (2010)[32].

Along with PBS, to evaluate whether TIP-eQTLs show signs of selection, we estimated nucleotide diversity (pi [38]), which is a SFS (site-frequency spectrum) measure to test for the presence of selection. This was done using a custom script in a window of 25SNPs with a sliding window of 5. Further, we also estimated the H12 homozygosity statistic [39] that can detect the presence of both hard and soft sweeps associated with selection. Since we expect long runs of homozygosity in the regions around the selected loci, H12 was calcuted in windows of 50 SNPs, with a sliding window of 25. Both these statistics were estimated separately for indica and japonica, and for the samples with TIP insertions present and absent. This was for a range of 100kb (50kb up-and downstream the TIP insertion), using SNPs identified from Groen et al [19] but without the 1000bp linkage thinning. To test for the significance of the statistics, we performed a permutation test (N=1000) with random sampling with the same number of individuals that were used to estimate the selection statistics. The statistics were estimated for these 1000 random pulls in the same regions, and using the same window sizes as the original statistics.

Acknowledgements

Author contributions

RC and JMC conceived the project. RC and JMC planned the experiments, together with SG and MDP for the selection analyses. RC performed most experiments with the help of NMD, who also solved the structure of the different haplotypes. SG performed the analyses of selection. All authors discussed the results JMC and RC drafted the article. All authors contributed to the paper.

References

1. Wendel J. F. , Jackson S. A. , Meyers B. C. , Wing R. A. (2016) **Evolution of plant genome architecture** *Genome Biol* **17:**37

2. Tenaillon M. I. , Hollister J. D. , Gaut B. S. (2010) **A triptych of the evolution of plant transposable elements** *Trends in Plant Science* **15:**471–478

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

4. Voichek Y. , Weigel D. (2020) **Identifying genetic variants underlying phenotypic variation in plants without complete genomes** *Nat. Genet* **52:**534–540

5. Vendrell-Mir P. (2019) **A benchmark of transposon insertion detection tools using real data** *Mob. DNA* **10:**53

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

7. Domínguez M. (2020) **The impact of transposable elements on tomato diversity** *Nat. Commun* **11:**4058

8. Castanera R. (2021) **The amplification dynamics of MITEs and their impact on rice trait variability** *Plant J* **107:**118–135

9. Vourlaki I.-T. , Castanera R. , Ramos-Onsins S. E. , Casacuberta J. M. , Pérez-Enciso M. (2022) **Transposable element polymorphisms improve prediction of complex agronomic traits in rice**

10. Swinnen G. , Goossens A. , Pauwels L. (2016) **Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement** *Trends in Plant Science* **21:**506–515

11. Meyer R. S. , Purugganan M. D. (2013) **Evolution of crop species: Genetics of domestication and diversification** *Nat. Rev. Genet* **14:**840–852

12. Wang Q. , Wang Y. , Sun H. , Sun L. , Zhang L. (2020) **Transposon-induced methylation of the RsMYB1 promoter disturbs anthocyanin accumulation in red-fleshed radish** *J. Exp. Bot* **71:**2537–2550

13. Martin A. (2009) **A transposon-induced epigenetic change leads to sex determination in melon** *Nature* **461:**1135–1138

14. Quadrana L. (2020) **The contribution of transposable elements to transcriptional novelty in plants: the FLC affair** *Transcription* **11:**192–198

15. Tian Y. (2022) **Transposon insertions regulate genome-wide allele-specific expression and underpin flower colour variations in apple (Malus spp.)** *Plant Biotechnol. J* **20:**1285–1297

16. Studer A. , Zhao Q. , Ross-Ibarra J. , Doebley J. (2011) **Identification of a functional transposon insertion in the maize domestication gene tb1** *Nat. Genet* **43:**1160–1163

17. Rebollo R. , Romanish M. T. , Mager D. L. (2012) **Transposable elements: An abundant and natural source of regulatory sequences for host genes** *Annu. Rev. Genet* **46:**21–42

18. Morata J. , Marín F. , Payet J. , Casacuberta J. M. (2018) **Plant lineage-specific amplification of transcription factor binding motifs by miniature inverted-repeat transposable elements (MITEs)** *Genome Biol. Evol* **10:**1210–1220

19. Groen S. C. (2020) **The strength and pattern of natural selection on gene expression in rice** *Nature* **578:**572–576

20. Kofler R. , Gómez-Sánchez D. , Schlötterer C. (2016) **PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq** *Mol. Biol. Evol* **33:**2759–2764

21. Zhang J. (2016) **Building two indica rice reference genomes with PacBio long-read and Illumina paired-end sequencing data** *Sci. Data* **3:**160076

22. Wicker T. (2007) **A unified classification system for eukaryotic transposable elements** *Nat. Rev. Genet* **8:**973–982

23. Uzunović J. , Josephs E. B. , Stinchcombe J. R. , Wright S. I. (2019) **Transposable Elements Are Important Contributors to Standing Variation in Gene Expression in Capsella Grandiflora** *Mol. Biol. Evol* **36:**1734–1745

24. Josephs E. , Lee Y. , Stinchcombe J. , Wright S. (2015) **Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression** *Proc Natl Acad Sci U S A* **112:**15390–15395

25. Lye Z. , Choi J. Y. , Purugganan M. D. (2022) **Deleterious mutations and the rare allele burden on rice gene expression**

26. Shanmugam S. , Boyett V. A. , Khodakovskaya M. (2021) **Enhancement of drought tolerance in rice by silencing of the OsSYT-5 gene** *PLoS One* **16:**

27. Gao W. (2022) **miR2105 and the kinase OsSAPK10 co-regulate OsbZIP86 to mediate drought-induced ABA biosynthesis in rice** *PLANT Physiol* **189:**889–905

28. Lambin J. (2020) **OsEUL Lectin Gene Expression in Rice: Stress Regulation, Subcellular Localization and Tissue Specificity** *Front. Plant Sci* **11:**185

29. Lu L. (2017) **Tracking the genome-wide outcomes of a transposable element burst over decades of amplification** *Proc Natl Acad Sci U S A* **114:**

30. Carpentier M. C. (2019) **Retrotranspositional landscape of Asian rice revealed by 3000 genomes** *Nat. Commun* **10:**24

31. Choi J. (2017) **The Rice Paradox: Multiple Origins but Single Domestication in Asian Rice** *Mol. Biol. Evol* **34:**969–979

32. Yi X. (2010) **Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude** *Science (80-.)* **329:**75–78

33. Cai Q. (2014) **Jasmonic acid regulates spikelet development in rice** *Nat. Commun* **5:**3476

34. Qin P. (2021) **Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations** *Cell* **184:**3542–3558

35. Zhou Y. (2020) **A platinum standard pan-genome resource that represents the population structure of Asian rice** *Sci. Data* **7:**113

36. Shang L. (2022) **A super pan-genomic landscape of rice** *Cell Res* **32:**878–896

37. Ansari T. H. , Yamamoto Y. , Yoshida T. , Miyazaki A. , Wang Y. (2003) **Cultivar Differences in the Number of Differentiated Spikelets and Percentage of Degenerated Spikelets as Determinants of the Spikelet Number per Panicle in Relation to Dry Matter Production and Nitrogen Absorption** *Soil Sci. Plant Nutr* **49:**433–444

38. Nei M. , Li W. H. (1979) **Mathematical model for studying genetic variation in terms of restriction endonucleases** *Proc Natl Acad Sci U S A* **76:**5269–5273

39. Garud N. R. , Rosenberg N. A. (2015) **Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps** *Theor. Popul. Biol* **102:**94–101

40. Xu F. (2019) **Control of rice pre-harvest sprouting by glutaredoxin-mediated abscisic acid signaling** *Plant J* **100:**1036–1051

41. Magwa R. A. , Zhao H. , Xing Y. (2016) **Genome-wide association mapping revealed a diverse genetic basis of seed dormancy across subpopulations in rice (Oryza sativa L.)** *BMC Genet* **17:**28

42. Zhang Y. (2017) **OsMPH1 regulates plant height and improves grain yield in rice** *PLoS One* **12:**

43. Asano K. (2011) **Artificial selection for a green revolution gene during japonica rice domestication** *Proc Natl Acad Sci U S A* **108:**11034–11039

44. Goubert C. , Zevallos N. A. , Feschotte C. (2020) **Contribution of unfixed transposable element insertions to human regulatory variation** *Philos. Trans. R. Soc. B Biol. Sci* **375:**20190331

45. Alonge M. (2020) **Major impacts of widespread structural variation on gene expression and crop improvement in tomato** *Cell* **182:**145–161

46. Arkhipova I. R. (2018) **Neutral theory, transposable elements, and eukaryotic genome evolution** *Mol. Biol. Evol* **35:**1332–1337

47. Matsumoto T. (2005) **The map-based sequence of the rice genome** https://doi.org/10.1038/nature03895

48. Li H. , Durbin R. (2010) **Fast and accurate long-read alignment with Burrows-Wheeler Transform** *Bioinformatics* **26:**589–595

49. Love M. I. , Huber W. , Anders S. (2014) **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2** *Genome Biol* **15:**550

50. Shabalin A. A. (2012) **Matrix eQTL: ultra fast eQTL analysis via large matrix operations** *Bioinformatics* **28:**1353–1358

Author information

Raúl Castanera
Center for research in agricultural genomics, CRAG (CSIC-IRTA-UAB-UB), campus UAB, Cerdanyola del Vallès, 08193 Barcelona, Spain.

For correspondence:

raul.castanera@cragenomica.esraul.castanera@cragenomica.es
ORCID iD: 0000-0002-3772-7727

Noemia Morales-Díaz
Center for research in agricultural genomics, CRAG (CSIC-IRTA-UAB-UB), campus UAB, Cerdanyola del Vallès, 08193 Barcelona, Spain.
ORCID iD: 0000-0001-5258-486X

Sonal Gupta
Center for Genomics and Systems Biology, 12 Waverly Place, New York University, New York, NY 10003, USA

Michael Purugganan
Center for Genomics and Systems Biology, 12 Waverly Place, New York University, New York, NY 10003, USA, Center for Genomics and Systems Biology, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates
ORCID iD: 0000-0002-9197-4112

Josep M. Casacuberta
Center for research in agricultural genomics, CRAG (CSIC-IRTA-UAB-UB), campus UAB, Cerdanyola del Vallès, 08193 Barcelona, Spain.

For correspondence:

josep.casacuberta@cragenomica.esjosep.casacuberta@cragenomica.es

Editors

Reviewing Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Germany

Senior Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Germany

Reviewer #1 (Public Review):

For many years it has been understood that transposable elements (TEs) are an important source of natural variation. This is because, in addition to simple knockouts of genes, TEs

carry regulatory sequences that can, and sometimes do, affect the expression of genes near the TEs. However, because TEs can be difficult to map to reference genomes, they have generally not been used for trait mapping. Instead, single nucleotide polymorphisms are widely used because they are easy to detect when using short reads. However, improvements in sequencing technology, as well as an increased appreciation of the importance of TEs to both linked to favorable alleles and are more likely to be causing the changes that make those alleles beneficial in a given environment. Further, because TE activity can occur after bottlenecks, they can provide polymorphisms in the absence of variation in point mutations.

In this manuscript, the authors carefully examine insertion polymorphisms in rice and demonstrate linkage to differences in expression. To do this, they used expression quantitative trait locus (eQTL) GWAS using TIPs as genetic markers to examine variation in 208 rice accessions. Because they chose to focus on genes that were expressed in at least 10% of the accessions, presumably because more rare variants would end up lacking statistical power. This is an understandable decision, but it says that recent insertions, such as the MITE elements detailed in a previous paper, would not be included. Importantly, although TIPs associated with differentially expressed genes are far less common than SNPs' traditional eQTLs, there were a significant number of eQTLs that showed linkage to TIPs but not to QTL.

The authors then show that of the eQTLs associated with both TIPs and SNPs, TIPs are more tightly linked to the eQTL, and are more likely to be associated with a reduction in expression, with variation in the effects of various TEs families supporting that hypothesis. Here and throughout, however, the distance of the TEs could be an important variable. It is also worth noting the relative numbers in order to assess the claim in the title of the paper. The total number of eQTL-TIPs is ten-fold less than the number of eQTL-SNPs, and, of the eQTLs that have both, there are a significant number of eQTL-TIPs that are not more tightly linked to the expression differences than the eQTL.

The authors show that eQTL-TIPs are more likely to be in the promoter-proximal region, but this may be due to insertion bias, which is well documented in DNA-type elements. Here and throughout the authors are careful to state that the data is consistent with the hypothesis that TEs are the cause of the change, but do not claim that the data demonstrate that they are.

Throughout the rest of the manuscript, the authors systematically build the case for a causal role for TEs by showing, for instance, that eQTL-TIPs show much stronger evidence for selection, with increased expression being more likely to be selected than decreased expression. The authors provide examples of genes most likely to have been affected by TE insertions.

Overall, the authors build a convincing case for TEs being an important source of regulatory information. I don't have any issues with the analysis, but I am concerned about the sweeping claims made in the title. Once you get rid of eQTLs that could be altered by either SNPs or TIPs and include only those insertions that show strong evidence of selection, the number of genes is reduced to only 30. And even in those cases, the observed linkage is just that, not definitive evidence for the involvement of TEs. Although clearly beyond the scope of this analysis, transgenic constructs with the TEs present or removed, or even segregating families, would have been far more convincing.

The fact that many of the eQTL-TIPs were relatively old is interesting because it suggests that selection in domesticated rice was on pre-existing variation rather than new insertions. This may strengthen the argument because those older insertions are less likely to be purged due to negative effects on gene expression. Given that the sequence of these TEs is likely to have

diverged from others in the same family, it would have been interesting to see if selection in favor of a regulatory function had caused these particular insertions to move away from more typical examples of the family.

Reviewer #2 (Public Review):

In this manuscript, Castanera et al. investigated how transposable elements (TEs) altered gene expression in rice and how these changes were selected during the domestication of rice. Using GWAS, the authors found many TE polymorphisms in the proximity of genes to be correlated to distinct gene expression patterns between O. sativa ssp. japonica and O. sativa ssp. indica and between two different growing conditions (wet and drought). Thereby, the authors found some evidence of positive selection on some TE polymorphisms that could have contributed to the evolution of the different rice subspecies. These findings are underlined by some examples, which illustrate how changes in the expression of some specific genes could have been advantageous under different conditions. In this work, the authors manage to show that TEs should not be ignored when investigating the domestication of rise as they could have played an important role in contributing to the genetic diversity that was selected. However, this study stops short of identifying causations as the used method, GWAS, can only identify promising correlations. Nevertheless, this study contributes interesting insights into the role TEs played during the evolution of rice and will be of interest to a broader audience interested in the role TEs played during the evolution of plants in general.