# The replicative amplification of MITEs and their impact on rice trait variability.

Raul Castanera[1*], Pol Vendrell-Mir[1], Amélie Bardil[1], Marie-Christine Carpentier [2], Olivier Panaud[2], Josep M. Casacuberta[1*]

[1] Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Bellaterra, 08193 Barcelona, Spain.

[2] Laboratoire Génome et Développement des Plantes, UMR CNRS/UPVD 5096, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy., 66860, Perpignan Cedex, France.

* Correspondence should be addressed to J.M.Casacuberta (josep.casacuberta@cragenomica.es) or R. Castanera (raul.castanera@cragenomica.es)

## Abstract

Transposable elements (TEs) are a rich source of genetic variability. Among TEs, Miniature Inverted-repeat Transposable Elements (MITEs) are of particular interest as they are present in high copy numbers in plant genomes and are closely associated with genes. MITEs are deletion derivatives of class II transposons, and can be mobilized by the transposases encoded by the latters through a typical cut-and-paste mechanism. However, this mechanism cannot account for the high copy number MITEs attain in plant genomes.

We present here an analysis of the distribution of Transposon Insertion Polymorphisms (TIPs) in 1,059 *O. sativa* genomes representing the main rice population groups. We performed MITE TIP-GWAS to study the impact of these elements on agronomically important traits and found that these elements uncover more trait associations than SNPs on important phenotypes such as grain width. Finally, using SNP-GWAS and TIP-GWAS we provide evidences of the replicative amplification of MITEs, suggesting a mechanism of amplification uncoupled from the typical cut-and-paste mechanism of class II transposons.

## Introduction

Transposable Elements (TEs) are an essential component of plant genomes. Their capacity to amplify in genomes, their ability to create new genetic variability by insertion/excision, and the possibility that their multiple copies offer for recombination, make them a rich source of genomic variants that can be selected through evolution [1]. The mechansism by which TEs can create new gene variants resulting in selectable phenotypes, both in wild plants and in crops are manyfold. These include the generation of gene knock-outs, but also the induction of their epigenetic silencing or changes in gene regulation by inactivating enhancers or represors upon insertion or by adding new regulatory elements contributed by the TE [2].

Miniature Inverted-repeat Transposable Elements (MITEs) are short non-coding TEs thought to be deletion derivatives of class II cut-and-paste TEs [3]. However, differently to the class II TEs they derive from, MITEs are found in high-copy number in plant genomes [4]. It has been suggested that MITE high copy number may be the result of a highly efficient transposition due to their promiscuity in being mobilized by different transposases [5], to the evolution of particular characteristics of the related transposase in some genomes [6], to a higher mobilization efficiency with respect to the related autonomous transposons [7] or to the presence of particular transposase-producing elements [8]. However, as highlighted short after MITE discovery, there is an apparent contradiction between a conservative cut-and-paste mechanism of transposition and MITEs high copy number in genomes, which suggests the possibility of the uncoupling of transposition and amplification mechanisms [3,9]. In fact, replicative amplification of MITEs has long been proposed [10], although no evidence for such a mechanism has been obtained until now, and the amplification of MITEs still remains and unsolved question.

Whereas the mechanism by which MITEs amplify is still obscure, the work done in different plant species has clearly shown the potential of these elements to alter the coding capacity and the expression of plant genes. MITEs are strongly associated with plant genes [11–13], and examples of these elements potentially altering gene expression have accumulated over the years [11,12,14,15]. Moreover, it has recently been shown that MITEs frequently contain transcription factor binding sites (TFBS) in plants, which may allow their mobilization to alter transcriptional networks by rewiring new genes [16].

In the last few years, different methods for analyzing TE insertion polymorphisms (TIPs) in resequenced genomes has been developed and recent benchmarking studies show that these methods can be reliable on datasets of enough coverage [17]. These methods have recently been used to perform genome-wide association studies (GWAS) using LTR-retrotransposon (LTR-RT) TIPs to analyze both the LTR-RT dynamics in rice [18], and the genetic basis of agronomic traits in rice and tomato [19,20]. These recent publications show that LTR-RT TIPs can allow discovering associations not seen with conventional GWAS strategies based on SNPs, and highlight the importance of analyzing the important fraction of the genetic variability TEs account for. As already explained, MITEs are a type of TEs particularly suited to induce gene variability in plants. Here we used MITE TIPs to perform TIP-GWAS on different agronomic traits and show that, indeed, MITE TIPs reveal an additional fraction of the genetic variability associated to important crop traits, and may allow discovering the underlying causal genes for some traits. In addition, we used MITE TIP-GWAS to study MITE dynamics in rice, and confirmed that MITEs amplify by a replicative mechanism from a reduced number of MITE copies.

## Results and Discussion

### 1. TE annotation and selection of rice varieties

Most TE sequences within genomes are fragmented and degenerated remnants of active TEs [3] and, although an extensive annotation of TE copies is essential for properly characterizing a genome, an analysis of the TE dynamics can be better accomplished starting with a stringent library of complete and potentially active elements. For this reason, and as a first step towards a comprehensive analysis of the TE dynamics in rice, we performed a stringent TE annotation using the REPET package (Flutre et al, 2011). As MITEs are particularly prevalent in plants [4], we complemented this annotation with a targeted annotation of MITEs based on MITE-hunter (Han et al, 2010) and the already published annotations in the P-MITE database (Chen et al, 2014). We used this pipeline to annotate 3 assembled rice reference genomes belonging to the three major rice subgroups (indica, japonica and aus), and we clustered all TE consensuses to obtain a global TE library of the species. We identified a total of 821 complete TE families belonging to all major TE orders, excluding incomplete and chimeric elements (Supplementary Fig 1, Supplementary dataset S1). In spite of the stringency of the approach used to build the TE library, a RepeatMasker annotation with the 821 consensus sequences on the Nipponbare reference genome showed highly congruent results in comparison to recently published rice TE annotations. More specifically, 95% of the annotation overlaps rice 6.9.5.liban annotation [21] (78% overlap in the opposite direction), and the classification agreement at the order level between reciprocal consensuses of the two libraries was 99%. However, the annotation used here has fewer LTR-RT consensuses (112 vs 389), probably due to the high stringent approach that did not retained partial, low-copy number and degenerated LTR-RT families, and has much more MITE consensuses (400 vs 173) due to the dedicated tools used.

In order to study the TE dynamics in rice, and in particular that of MITEs, we took advantage of the availability of resequencing data of 3000 rice varieties to look for TE insertion polymorphisms (TIPs) in the different varieties of the TEs annotated in the reference genomes. A recent benchmarking exercise indicated that PopoolationTE2 [22] is a good tool for this purpose, but also that coverage is a key factor than can limit the detection of TIPs [17]. For this reason, we selected the 1,059 rice genomes sequenced at

15x or more (Supplementary Table S1), and subsampled all genomes to 15x in order to be able to perform an unbiased comparisons of TIP abundance between varieties.

We found an average of 40,568 insertions per variety. We filtered the insertions using a zygosity cut-off of 0.7 as recently recommended to avoid false positive calls [17]. After applying this filter, we obtained an average of 18,463 insertions per variety.

The number of TIPs per variety is variable (Fig. 1), but a first analysis indicated that CAAS varieties had a significantly lower number of TIPs than IRIS varieties (Supplementary Figure S2), despite the identical processing pipeline followed to sequence the selected varieties [23]. We confirmed this result using two independent additional approaches: i) by analyzing a single MITE family using a different TIP prediction tool, Trackposon [18] in the 3,000 rice genomes, and ii) by looking for the TE insertions identified among the structural variants of the 3,000 rice genomes described in (Fuentes et al, 2019)[24]. Both analyses showed a significantly higher number of insertions in IRIS than in CAAS varieties (Supplementary Figure S2). We didn´t find significant differences in sequencing and mapping coverage between CAAS and IRIS. Also, both sets contain a representative sample of varieties from the nine subpopulations described in (Wang et al, 2018)[25]. Although we could not identify the reason for such a difference, we warn that it could be a limitation for quantitative studies targeting TIPs in the 3,000 genomes dataset. In order to avoid any possible bias we decided to use only the IRIS data, which reduced the number of varieties analyzed to 738.

## 2. TE dynamics in rice varieties

We detected a total number of 103,109 TIPs in the 738 IRIS varieties corresponding to 32,449 retrotransposons and 70,660 DNA transposons of all superfamilies. PCA analysis and phylogenetic trees performed based on predicted TIPs are congruent with the previously defined groups of rice varieties based on SNPs [25] (Figure 1A, Figure 1B). Although the analyses using the TIPs of each type of TEs alone gave similar results, the resolution of the different groups varied among them (Supplementary Figure S3). For example, the analysis based on LTR-RT TIPs better resolved all the different rice groups as compared with MITEs, suggesting a higher number of non-fixed and non-private LTR-RT TIPs.

The analysis of TIPs in the different varieties shows that most TE families have a relatively constant number of TIPs per variety among the four main varietal groups

(japonica, indica, aus and aro), with a coefficient of variation between groups (CV) lower than 20% in 82% of the families, and with a median of 9.5% (Figure 2A). However, a few number of families are much more variable, the CV reaching a maximum of 128.2%. Up to 150 TE families had a coefficient of variation higher than 20% and have potentially experienced a varietal group-specific amplification (Figure 2A, 2B), suggesting that TEs may be at the origin of an important fraction of recent genomic variation in rice. Interestingly, the coefficient of variation as well as the proportion of private insertions, are higher for smaller families (Figure 2C,2D), suggesting that these families have been more active recently. Among the different TE orders, LTR-retrotransposons were the only order significantly enriched among the group-specific amplification families (Fisher test p-value = 0.0062, Table 1), which is in agreement with recent reports suggesting that rice LTR-RTs have been active in agro [18] .

In order to get more insight on the dynamics of the different types of TEs we analyzed the TIP frequency distribution for each type of TEs in a subset of 382 varieties that correspond to traditional varieties, which show sustained cultivation in specific geographic locations, show local adaptation and are a good representation of the variability of the species [26]. ,

As shown in Figure 3A the frequency distribution is different for different types of TEs. In particular, and as previously shown [18], our study reveals that most LTR-RT TIPs are present at very low frequency, suggesting both a recent activity and a high turnover of LTR-RT insertions. On the contrary, although there are an important number of MITE TIPs present at low frequency, there is a significant number of them that appear to be fixed in the population, presenting a “U-shaped" frequency spectrum with an excess of high-frequency derived mutations at the expense of middle-frequency variants. This suggests that, whereas most LTR-RT insertions are recent, an important fraction of MITE insertions is much older. This can also be seen when analyzing the fraction of TIPs corresponding to insertions present in rice (*Oryza sativa* ssp. *japonica* cv. Nipponbare) and 9 different *Oryza* genomes [27] (Figure 3B). Whereas less than 10% of the LTR-RT insertions present in Nipponbare rice are also present in *O. glumaepatula*, more than 30% of the MITE insertions are also present in this relatively distant *Oryza* genome (Figure 3B). This data clearly suggests that an important fraction of MITE insertions have been fixed during rice evolution. As MITEs tend to concentrate close to genes [3,9] these fixed insertions may have played an important role in the evolution of

rice genes. On the other hand, an analysis of the frequency within rice varieties of the insertions shared between Nipponbare rice and *O. glumaepatula*, shows that these relatively old insertions have been maintained in the whole population (Figure 3B), confirming that, as already proposed [3], MITE insertions are highly stable, in spite of the possibility of being excised by related transposases. A recent study that analyzed the population frequencies of rice TEs in a high-coverage dataset of 96 indica and 106 japonica accessions found lower fixation frequencies for all TE families in comparison to synonymous SNPs [28]. Nevertheless, they found higher population frequencies for LTR-retrotransposons in comparison to previously published data [18] and the data shown here. In line with what the authors discuss, we think that these differences are likely due to the inclusion of partial and truncated LTR-RT and probably old elements (not included in the dataset used here), likely found at higher population frequencies, An analysis of the chromosomal distribution of LTR-RT TIPs shows that, as expected, gypsy and copia LTR-RT insertions have opposite distributions. Whereas gypsy TIPs tend to concentrate in regions surrounding the centromeres, copia LTR-RTs insertions show a wider distribution in chromosome arms, paralleling that of genes (Figure 4A and Supplementary Figure S4 ). However, although these general trends are clear, different families show specific patterns of insertions. Among gypsy, family RLX_comp_IRGSP_B_R1329_Map20, which corresponds to the previously described CRR retrotransposon [29], concentrates in the centromeres of all chromosomes (Supplementary Figure S5), whereas other families have a wider distribution and a few are frequently inserted close to genes (Supplementary Table S2). Similarly, whereas most copia families are frequently found within or close to genes, others show a more random distribution and a few are preferentially found far from genes, with a pattern more typical for gypsy elements (Supplementary Table S2). These specific patterns of distribution probably reflect different target specificities for each subfamily that modulate the already described general pattern of distribution of plant gypsy and copia LTR-RTs.

The distribution of MITE insertions is similar to that of copia LTR-RTs and also follows that of genes (Figure 4A and Supplementary Figure S4). This is expected as Copia LTR-RTs and MITEs are known to be closely associated to plant genes [9]. However, a comparison of the distribution of TIPs present at low- and high-frequency, which allow starting to discriminate between the role of target specificity and selection in shaping TE distribution, shows clear differences between the two TE groups.

Whereas only low frequency copia LTR-RTs insertions follow gene distribution, MITE insertion association with genes is stronger for high frequency insertions (Figure 4, Supplementary Figure S6). This suggests that whereas in general both copia LTR-RTs and MITEs seem to target genic regions for integration, LTR-RTs are progressively cleaned away from these regions whereas MITEs inserted close to genes are more frequently maintained. As reported above, LTR-RT TIPs are found at low frequencies, suggesting that most of them are quickly eliminated. On the contrary, an important fraction of MITE TIPs are found at high frequencies and many of them are fixed in rice. The smaller size of MITEs, as compared with LTR-RTs, could make their insertions less deleterious and more easily tolerated, and the preferential retention of MITEs close to genes could simply be due to the difficulty of eliminating them without affecting neighboring genes. However, the fixation of MITEs close to genes could also be the result of a positive selection of some MITE insertions.

An analysis of the non-reference TIPs with respect to genes shows that MITEs concentrate in the 5' and 3' proximal regions (<500 nt) of genes (Figure 4B and 4C) and that this association is even more striking when analyzing MITE TIPs found at low frequency (Supplementary Figure S8).

Many examples of MITE insertions in 5' and 3' of genes that alter gene expression in different plants have accumulated over the years, including insertions in promoters enhancing [30–32] or repressing [33] transcription, or repressing translation [34]. Moreover, MITEs have been shown to frequently contain transcription factor binding sites in plants, including rice [16], suggesting a potential impact on promoter evolution. It is therefore possible, that some of the MITE insertions described here may actually modify gene expression, and that some of the insertions present at high frequency in rice may have been positively selected.

**3. Association of MITE insertions with trait variability.**

Genome-wide association studies (GWAS) using LTR-RT TIPs as a genotype, instead of SNPs, have recently been performed in rice and tomato [19,20]. These studies showed that these TIPs can reveal additional genetic associations with traits that are not seen with SNPs. Here we wanted to explore the potential of MITE TIPs for TIP-GWAS. We reasoned that MITE's small size, high copy number, and close association with genes, could make them particularly suited for this purpose. Moreover, whereas most MITE

TIPs are found at high frequency, a fraction of them is found at much lower frequencies, the frequency distribution having a U-shape, as reported above. Therefore, we have performed GWAS using the 66,368 MITE TIPs (including 6,000 TIPs that could also be related to TIR-TEs) as genotype and the phenotypic data available for the 451 indica varieties [35] We obtained MITE association peaks with 6 phenotypes (grain width, length and weight, salt injury at EC18, flowering time and leaf length, Supplementary Tables S3 and S4). Some of the peaks obtained where coincident with peaks using SNPs as genotype, suggesting that both were in LD with the causal mutation (that could also be the SNP or the MITE TIP themselves). However, in many cases the MITE-GWAS revealed different peaks allowing to explore additional genetic regions potentially linked to trait variability.

As an example, whereas the SNP-GWAS revealed a single genomic locus linked to grain width, the GWAS produced with the MITE TIP data revealed 9 different significant association peaks (Figure 5, Supplementary Table S3). The only SNP association obtained was found at the well-characterized GSE5 gene of the GW5 locus controlling grain size, with the leading SNP present at 3.9 Kb of GSE5 [36]. Interestingly, the strongest association using MITE TIPs (MITE-1) was found with a MITE insertion present in the promoter region of the GSE5 gene (350 bp upstream, chr05:5364000-5365000, pvalue = 3.28e-25).This insertion is in strong linkage disequilibrium with a deletion located 4,500 bp upstream the GSE5 gene, and the co-occurrence of these two structural variants constitute one of the three main haplotypes of GSE5 in cultivated rice ($GSE5^{DEL1+IN1}$) [36]. However, in spite of the MITE insertion being closer to the GS5 gene than the deletion, it was suggested the deletion was the main responsible of the causal phenotype [36]. Indeed, our analysis also suggest that the MITE insertion is not the causal variant responsible of grain width, as the few varieties carrying the MITE but not the deletion have narrow grains (supplementary Figure S8).

In addition to the association related to the GSE5 gene, we detected 8 more peaks with MITE insertions associated with grain width, several of which are located close to genes or within QTLs already characterized as linked to grain phenotypes. The MITE-15 insertion is located at 33 Kb of the OsARG gene, which controls grain yield in rice [37], and five additional MITE TIPs are within a previously characterized QTLs linked to grain yield (MITE-11 at qGY-3 [38] and MITE-21 at QTARO QTL-450 [39]), embryo length (MITE-3 at qEML-2 [40]), grains per panicle (MITE-12 at gpp4 [41]), grain shape (MITE-21 at qLW-6 [42] ) and 1000 kernel weight (MITE-16 at QKw5 [43]), which

reinforces de relevance of the associations found. A detailed analysis of the MITE insertion positions showed that four out of these nine significant MITE TIPs were located inside genes or in their close vicinity (1000bp upstream or downstream) (Supplementary Table S3). We anticipate that these genes could be good candidates to be linked to the grain width trait.

The analysis of the potential implication of MITE insertions with trait variability will require further analysis, but the number of potential associations and the analysis of the grain width trait here provided highligh the interest of MITE TIP-GWAS for analyzing the genetic determinants of trait variability in plants.

## 4. GWAS as a tool to identify active replicative TEs

In addition to study the genetic basis of agronomic traits, GWAS can also be used to study the genetic determinants of TE activity, as recently done for LTR-RTs in rice [18]. Here we wanted to use this approach to study the genetic determinants of MITE amplification. This is particularly relevant as although MITEs are known to be mobilized by transposases of class II related elements, it is difficult to imagine how a cut-and-paste mechanism may allow them to reach the high-copy numbers that MITEs frequently attain in eukaryotes, and in particularly in plant genomes [4]. Although a replicative transposition mechanism independent of related transposases was proposed long-time ago [10], and the phylogenetic analyses of MITE populations would be consistent with such a mechanism [5,11,12,14,44], the direct evidence of the replicative amplification of MITEs is still lacking.

As a first step towards analyzing the genetic determinants of MITE amplification, and in order to validate the approach, we looked for the genetic determinants of well-known replicative TEs such as LTR-RTs and LINEs. To do so, we performed a GWAS analysis on 451 rice varieties belonging to indica subspecies using a SNP matrix of 173,260 SNPs as genotype, and the TIP copy number of 216 rice TE families as phenotype (104 LINEs and 112 LTR-RT), running one association analysis per TE family. We used indica population as they are present in a higher number in our dataset, and this higher number provide us with more power for the GWAS analyses. We found up to 189 significant peaks (p value cutoff = 3.2E-07, Bonferroni adjusted significance threshold) corresponding to 130 TE families (60% of the total families). The SNPs associated with family amplification were distributed along the twelve chromosomes

without a particular enrichment in any genomic region (Supplementary Table S5). 64% of the significant peaks (122 peaks from 75 families) had a TIP of the same family at less than 100kb of the corresponding SNP (median distance to SNP = 24.9 Kb), suggesting that the TE insertion corresponding to that TIP could be at the origin of the amplification of the family. In 97% of these cases, the varieties with the TIP have higher copy number than varieties without the TIP ($p < 0.05$, Wilcoxon test), suggesting that, indeed, the presence of this particular copy of the TE could be at the origin of the increase in TE copy number.

Replicative transposons, such as LTR-RTs and LINEs are usually amplified from a small number of complete and active elements that act as "master copies". This is particularly clear for LINEs which are transcribed from an internal promoter located in it's 5' end (Swergold, 1990), absent from the vast majority of elements due to 5' deletions that make them inactive (Farley et al., 2004). This is the consequence of a transposition mechanism that frequently leads to the integration of elements truncated in 5', due in part to the low processivity of the RT and to the microhomology-facilitated recombination during integration [45]. Therefore, most newly transposed elements are incapable of expression and transposition, and their structure is different from that of the few "master" elements that are at the origin of each LINE family. Therefore, and in order to validate that the TIPs we found associated with the increase in copy number of particular TE families, we compared the structure of the LINE copies identified as potentially responsible for the amplification of a LINE family with that of the rest of the copies. As an example, Figure 6A (top) shows the Manhattan plot corresponding to the family RIX_comp_MH63_B_R3107_Map6. The three major peaks in chromosomes 8, 11 and 12 identify positions contain a TIP of the same family at positions chr08:5,711,756, chr11:23,993,235 and chr12:1,525,695, respectively. These TIPs (hereafter referred as genetic factors, GF) are in strong linkage disequilibrium (LD) with the minor SNP allele (ie, GF1 and GF2) or with the major SNP allele (ie, GF3), which may explain the association of the SNP with the LINE family copy number. The association of these three LINE insertions with the copy number of this LINE family can also be seen in a TIP-GWAS performed using LINE TIPs as genotype (Figure 6A bottom). Moreover, the presence of the three LINE copies is associated with an increase of copy number of this LINE family (Figure 4B), suggesting that indeed, the three LINE copies may correspond to "master" LINE elements at the origin of this LINE family. In order to validate the presence of these insertions and study the structure of these three

potential "master" LINEs, we used the sequence of the indica varieties from the recently published platinum genomes [27]. The element corresponding to GF-1 is present in the LIMA::IRGC 81487-1 genome and both GF-1 and GF-3 are present in the IR 64 genome. A comparison of these elements with all the other copies of the same family present in these genomes shows that they are the only complete ones, the other copies being truncated at 5´ (Figure 6D). All these results suggest that the associations detected point to the active (or "master") elements that are at the origin of the most recent replicative amplification of these families.

## 5. The replicative amplification of MITEs

We found significant association SNP peaks for 175 out of the 400 MITE families and up to 37% of these peaks (66) have a MITE-TIP of the same family at less than 100kb (Supplementary Table S6, see Figure 7 for an example). The presence of the MITE copy is linked to an increase of MITE copy number in 96% of the cases. We used the Indica varieties from the recently published platinum genomes [27] to analyze the insertion corresponding to the TIP located close to the 66 SNP peaks associated to MITE copy number and in the 42 cases where the TIP is predicted to be present in one of the platinum genomes, we confirmed the presence of a MITE in the genome assembly.

As an alternative approach, we performed GWAS using MITE TIPs instead of SNPs as genotype, and we detected the association of the copy number of a MITE family with the insertion of a specific MITE of the same family for 187 families, including 44 of the 66 associations found previously by SNP-GWAS. The high number of associations with MITE insertions of the same family strongly suggests that the associated MITE is not in LD with another genetic factor influencing MITE amplification but that this particular MITE copy is probably at the origin of the increase in copy number of the MITE family.

As already explained, MITEs can be mobilized by the transposases of related TIR-TEs, and it has been proposed that their colonization success could be due to their ability to be mobilized by the transposases of both closely and distantly related class II elements [5], to particular the evolution of characteristics of the transposase in some genomes [6] or to a higher mobilization efficiency with respect to the related autonomous transposons [7].

In order to check whether the presence of particular transposases related to the MITE families could be at the origin of the amplification of some MITE families we looked for TIPs of transposase encoding class II elements (TIR-TEs) close to (less than 100kb)

the significant SNP association peaks, and we also performed GWAS using the TIR-TE TIPs (arising from TE consensuses coding for transposases) as genotype instead of SNPs. Although we found TIR-TEs close to SNP association peaks, none of them had significant similarity (even limited similarity at the TIR level) to the corresponding MITE family. On the other hand by using TIP-GWAS we detected only 51 associations between the presence of a TIR-TE TIP (39 different elements) and a MITE family copy number. Only two of the 39 TIR-TE elements have significant similarity at the TIR level with the corresponding MITE family, but in the two cases, the region also contains a MITE copy of the same family, suggesting that this TIP is likely a MITE and not a transposase. These results suggest that the TIR-TE TIPs detected potentially associated to MITE family copy number are not causal and probably are in LD with another genetic determinant of MITE amplification.

By comparing the sequences of MITEs and TIR-TEs present in our TE library we identified 18 MITE families showing significant sequence similarity with TIR-TEs from which they probably derive (Supplementary Table S7). The analysis of these families by SNP-GWAS and TIP-GWAS did not reveal any association between the presence of this TIR-TE with the copy number of the corresponding MITE family. As an example Figure 7 shows the case of the hAT family DTX_comp_IRGSP_B_R1932, and the related MITE family SE260300235fam318_632. It can be clearly seen that none of the TIR-TEs, inclusing those of closely related elements, is associated with the amplification of the MITE family (Figure 7D, in green). By contrast, MITE TIPs displayed the highest association with family activity and were located in the closest vicinity of the leading SNPs (Figure 7A,B), and are in strong LD with the alleles identified by the SNP-GWAS (Figure 7E).

All these results suggest that, even if transposases of related class II transposons could mobilize MITEs by a cut-and-paste mechanism, they may not be the main genetic determinant for their amplification. The fact that particular MITE copies show a positive association with MITE copy number suggests, that as it is the case for LINEs and LTR-RTs, one or few MITE copies are amplified by a replicative mechanism to generate a MITE family.

The replicative origin of MITE families was proposed short after the discovery of these elements [10]. Izsvak and co-workers proposed that the high secondary structure of MITEs could allow them to fold back to form a stem-loop ssDNA molecule that might detach from the chromosome during replication, in a mechanism similar to that of certain

bacterial transposons that move trough single-strand transposition associated with DNA replication [46]. The identification here of particular copies at the origin of the amplification of MITE families, would explain the star-shaped phylogenetic trees obtained in genome-wide analyses of MITEs [5,11,12,14,44], and suggest that some MITE copies may have particular characteristics (e.g. sequence, chromatin, locus,...) that make them more prone to amplification. In order to look for possible characteristics of the MITE copies involved in MITE amplification, we compared the sequence of 20 of these MITEs present in the assembled indica rice LIMA::IRGC 81487-1, as well as of their flanking sequences, with that of the rest of the members of the family for the 20 different MITE families. We analyzed potential differences in GC content and Minimum Free Energy Structure of these elements including flanks of different length (see methods), as well as TE content of the regions, but failed to detect significant differences between the MITE copies involved in MITE amplification and the rest of the copies of their respective families (Supplementary Table S8).

In order to analyze the possible influence of different chromatin characteristics using the data available for the Nipponbare reference genome, we look for the presence in this genome of the MITE TIPs characterized here as associated to the copy number of their respective families. We were able to identify 43 of these MITEs in Nipponbare and we compared them with all the annotated copies of the corresponding MITE families in the whole genome. The characteristics of the chromatin associated to these MITEs seems different than that of the rest of the copies of their respective families. In particular, they show a potential enrichment in transcription activation marks, with 33% more H3K4me3 and 14% more DNAse hypersensitive peaks coinciding with these specific insertions, as compared with the rest of the copies of the corresponding families. On the other hand, a recent analysis has characterized the meiotic recombination hotspots in indica rice mapping them to the Nipponbare genome [47]. An analysis of the overlap of these spots with MITEs TIPs shows that 18.6 % the 43 MITEs characterized here as associated to the copy number of their respective families overlap with these recombination hotspots whereas only 12.2 % of all MITEs overlap with these sites. On the other hand, origins of replication are closely associated with genes and are enriched in active transcription epigenetic marks [48], and it has been proposed that the presence in certain Arabidopsis MITEs of binding sites for E2F/RBR1 transcription factors [49] may

promote DNA replication and/or repair and thus facilitate their transposition and amplification [50,51].

Although the precise mechanism of MITE amplification remains to be elucidated, the results presented here show that few particular MITE copies are at the origin of their replicative amplification. The enrichment of chromatin marks associated with active transcription on these elements as well as some enrichment in recombination hotspots compared with the rest of the elements of the family, may help designing experiments to start characterizing this mechanism.

# Methods

## Data source

We used the sequencing data released from the 3000 rice genomes project [23] and three Platinum indica assemblies (LIMA::IRGC 81487-1, KHAO YAI GUANG::IRGC 65972-1 and LARHA MUGAD::IRGC 52339-1) [27]. We obtained the fastq files from the 1,059 varieties sequenced at a minimum coverage of 15X.

## Reconstruction of TE family consensuses

We used TEdenovo from the REPET package [52] to build TE consensus sequences from the Japonica Nipponbare genome [53], Indica MH63 [54]and Aus N22 [55], representing the three main rice subspecies. TE consensuses were classified at the order level using PASTEC [56], and only those classified as “complete“ (based on structure and coding potential) elements were retained. We concatenated the three datasets and removed redundancy by obtaining a library of centroids at 80% identity using VSEARCH [57]. In parallel, we run MITE-Hunter [58] in the three rice genome assemblies. We combined the predictions with the high-quality MITE dataset from PMITE [4] and followed the same clustering approach to obtain a comprehensive MITE library. Finally, we concatenated the two libraries (generic TEs and MITEs) to obtain a definitive set of complete TE consensuses capturing the diversity of the rice. Repeatmasker (http://www.repeatmasker.org/) was used to identify genomic regions with similarity to TE consensuses of our TE library.

**Detection of TE insertions from sequencing reads**

We used PopoolationTE2 [22] to detect TE insertions using Nipponbare as reference genome and our TE consensuses (Supplementary Dataset S1). The mode "separate" was used in order to be able to quantitatively compare the number of TIPs per family. We obtained a high-confidence set of insertions by zygosity filtering at 0.7 as previously recommended [17]. In order to avoid the bias caused by differential sequencing depth, we randomly sampled every accession to 15X prior running the tool using seqtk (https://github.com/lh3/seqtk) . To identify non-reference insertions, we intersected all detected insertions with the regions annotated by RepeatMasker and selected all the insertion points that were further than 25bp to any annotated TE in the rice Nipponbare reference genome used here as a reference.

**Estimation of TIP insertion frequencies**

Nipponbare genome was split into 500bp (MITEs) and 1Kb (rest of TE) windows. The results of PopoolationTE2 were transformed into bed format. As a first step, filtered insertions from different TE orders were separated and intersected with the genome windows to obtain TIPs. TIPs from the 1,059 varieties were concatenated in one file and identical windows were collapsed to remove redundancy. Finally, TIPs of the same order arising from adjacent windows were merged. This procedure was followed to obtain a set of high-confidence TIPs. In a second iteration, we intersected all the insertions of the 1,059 varieties (excluding the zygosity filter and splitting by order) with the positions of the high-confidence TIPs dataset to score the presence/absence status of each TIP in each variety and obtain the final TIP matrices. TIP insertion frequencies were calculated by obtaining the proportion of varieties that contain each insertion.

**Population structure analyses**

Principal component analysis was performed using prcomp function from the *stats* R package [59]. TIP-based Neighbor-joining tree was built using *ape* library from the R package [60].

**SNP and TIP-GWAS**

We used the LFMM 2 R package [61] to obtain genotype-phenotype associations applying latent factor mixed models in indica varieties to correct for population structure (K = 4).

For the SNP-WGAS we used the 170 K SNP matrix obtained by Carpentier et al. [18] in lfmm format, derived from the 404k CoreSNP dataset available at SNP-seek [35] (after filtering for minor allele frequency >5% and missing data < 20%). For TIP-GWAS, we used TIP matrices as genotype and considered TIPs with minor allele frequencies > 1%. Bonferroni correction was used to set significance p value thresholds. Phenotypic traits were downloaded from SNP-seek database.

**Identification of genetic factors**

The results of every SNP-GWAS using TE family copy number as phenotype and SNPs as genotype (one SNP-GWAS per TE family) were manually inspected to identify clear peaks with significant associations (Bonferroni-adjusted p value threshold = 3.2E-07). In each case, we looked for TIPs of the same family located at < 100 Kb, which is a reasonable cutoff given the patterns of linkage disequilibrium in indica rice [62]. If a TIP of the same family was found (putative genetic factor), phenotypes of varieties with and without the genetic factor were collected to produce violin plots. Statistical significance between family copy number between varieties carrying or not the genetic factor was assessed using Wilcoxon two-tailed test (p value cutoff = 0.05).

**Validation of Genetic Factors**

Each of the windows containing the genetic factor were extended 4000bp on each flank and extracted from the reference genome (IRGSP, Nipponbare). The regions were mapped to three Platinum indica assemblies present in our dataset (LIMA::IRGC 81487-1, KHAO YAI GUANG::IRGC 65972-1 and LARHA MUGAD::IRGC 52339-1) using minimap2 [63]. By screening the best mapping hits we identified the orthologous regions in the three assemblies. Using the family consensus as query for a BLASTN search [64] (cutoff 1E-20), we verified the presence of the corresponding TIP in each orthologous region. Schematic representation of all genomic BLAST hits obtained using the consensus sequence as query was performed using Sushi R library [65].

**Sequence analysis of Genetic factors**

The sequence of every MITE in LIMA::IRGC 81487-1 genome was extracted including 100, 250 and 500 bp upstream and downstream. Sequence GC content was determined with bbmap [66] and the minimum free energy with ViennaRNA package [67].

**Data availability**

The datasets generated during and/or analysed during the current study are available in the following repositories: TE consensuses, raw insertions detected by PopoolationTE2 and TIP matrices are available in Zenodo (10.5281/zenodo.4058696); Raw sequencing data of the 1059 rice genomes is available as part of the 3,000 rice genomes project in SRA accession PRJEB6180; H3K4me3 and DNAse hipersensitivity sites were obtained from GEO database (accession numbers GSM489083 and GSM655033 ).

## References


1. Tenaillon, M. I., Hollister, J. D. & Gaut, B. S. A triptych of the evolution of plant transposable elements. *Trends in Plant Science* **15**, 471–478 (2010).
2. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013).
3. Feschotte, C., Jiang, N. & Wessler, S. R. Plant transposable elements: Where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329–341 (2002).
4. Chen, J., Hu, Q., Zhang, Y., Lu, C. & Kuang, H. P-MITE: A database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res.* **42**, D1176–D1181 (2014).
5. Feschotte, C., Swamy, L. & Wessler, S. R. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with Stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* **163**, 747–758 (2003).
6. Guermonprez, H., Loot, C. & Casacuberta, J. M. Different strategies to persist: The pogo-like Lemi1 transposon produces miniature inverted-repeat transposable elements or typical defective elements in different plant genomes. *Genetics* **180**, 83–92 (2008).
7. Yang, G., Nagel, D. H., Feschotte, C., Nathan Hancock, C. & Wessler, S. R. Tuned for transposition: Molecular determinants underlying the hyperactivity of a stowaway MITE. *Science (80-. ).* **325**, 1391–1394 (2009).
8. Chen, J. *et al.* Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nat. Commun.* **10**, 641 (2019).
9. Casacuberta, J. M. & Santiago, N. Plant LTR-retrotransposons and MITEs: Control of transposition and impact on the evolution of plant genes and genomes. *Gene* **311**, 1–11 (2003).
10. Izsvák, Z. *et al.* Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J. Mol. Evol.* **48**, 13–21 (1999).
11. Santiago, N., Herráiz, C., Ramón Goñi, J., Messeguer, X. & Casacuberta, J. M. Genome-wide analysis of the Emigrant family of MITEs of Arabidopsis thaliana. *Mol. Biol. Evol.* **19**, 2285–2293 (2002).
12. Lu, C. *et al.* Miniature inverted-repeat transposable elements (MITEs) have been

accumulated through amplification bursts and play important roles in gene expression and species diversity in oryza sativa. *Mol. Biol. Evol.* **29**, 1005–1017 (2012).

13. Benjak, A., Boué, S., Forneck, A. & Casacuberta, J. M. Recent amplification and impact of MITEs on the genome of grapevine (Vitis vinifera L.). *Genome Biol. Evol.* **1**, 75–84 (2009).

14. Naito, K. *et al.* Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461**, 1130–1134 (2009).

15. Yang, G. *et al.* A two-edged role for the transposable element Kiddo in the rice ubiquitin2 promoter. *Plant Cell* (2005). doi:10.1105/tpc.104.030528

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

17. Vendrell-Mir, P. *et al.* A benchmark of transposon insertion detection tools using real data. *Mob. DNA* **10**, 53 (2019).

18. Carpentier, M. C. *et al.* Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nat. Commun.* **10**, 24 (2019).

19. Akakpo, R., Carpentier, M. C., Ie Hsing, Y. & Panaud, O. The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytologist* (2020). doi:10.1111/nph.16356

20. Domínguez, M. *et al.* The impact of transposable elements on tomato diversity. *Nat. Commun.* (2020). doi:10.1038/s41467-020-17874-2

21. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

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

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

24. Fuentes, R. R. *et al.* Structural variants in 3000 rice genomes. *Genome Res.* **29**, 870–880 (2019).

25. Wang, W. *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).

26. Gutaker, R. M. *et al.* Genomic history and ecology of the geographic spread of

rice. *Nat. Plants* **6**, 492–502 (2020).

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

28. Kou, Y. *et al.* Evolutionary genomics of structural variation in Asian rice (Oryza sativa) domestication. *Mol. Biol. Evol.* (2020). doi:10.1093/molbev/msaa185

29. Nagaki, K. *et al.* Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol. Biol. Evol.* **22**, 845–855 (2005).

30. Zheng, X. *et al.* Natural Variation in CCD4 Promoter Underpins Species-Specific Evolution of Red Coloration in Citrus Peel. *Mol. Plant* **12**, 1294–1307 (2019).

31. Shimada, T. *et al.* MITE insertion-dependent expression of CitRKD1 with a RWP-RK domain regulates somatic embryogenesis in citrus nucellar tissues. *BMC Plant Biol.* **18**, 166 (2018).

32. Yin, S. *et al.* Transposon insertions within alleles of BnaFLC.A10 and BnaFLC.A2 are associated with seasonal crop type in rapeseed. *J. Exp. Bot.* **71**, 4729–4741 (2020).

33. Mao, H. *et al.* A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. *Nat. Commun.* **6**, 8326 (2015).

34. Shen, J. *et al.* Translational repression by a miniature inverted-repeat transposable element in the 3′ untranslated region. *Nat. Commun.* **8**, 14651 (2017).

35. Mansueto, L. *et al.* Rice SNP-seek database update: New SNPs, indels, and queries. *Nucleic Acids Res.* **45**, D1075–D1081 (2017).

36. Duan, P. *et al.* Natural Variation in the Promoter of GSE5 Contributes to Grain Size Diversity in Rice. *Mol. Plant* **10**, 685–694 (2017).

37. Ma, X. *et al.* OsARG encodes an arginase that plays critical roles in panicle development and grain production in rice. *Plant J.* (2013). doi:10.1111/j.1365-313x.2012.05122.x

38. Mao, B. B. *et al.* Characterization of QTLs for harvest index and source-sink characters in a DH population of rice (Oryza sativa L.). *Yi Chuan Xue Bao* **30**, 1118–1126 (2003).

39. Yonemaru, J. ichi *et al.* Q-TARO: QTL annotation rice online database. *Rice* **3**, 194–203 (2010).

40. Dong, Y., Tsuzuki, E., Kamiunten, H., Terao, H. & Lin, D. Mapping of QTL for embryo size in rice. *Crop Sci.* **43**, 1068–1071 (2003).

41. Xiao, J., Li, J., Yuan, L. & Tanksley, S. D. Identification of QTLs affecting traits of agronomic importance in a recombinant inbred population derived from a subspecific rice cross. *Theor. Appl. Genet.* **92**, 230–244 (1996).

42. Yan, C. J. *et al.* Mapping quantitative trait loci associated with rice grain shape based on an indica/japonica backcross population. *Acta Genet. Sin.* **30**, 711–716 (2003).

43. Li, Z., Pinson, S. R. M., Park, W. D., Paterson, A. H. & Stansel, J. W. Epistasis for three grain yield components in rice (Oryza sativa L.). *Genetics* **145**, 453–465 (1997).

44. Xin, Y., Ma, B., Xiang, Z. & He, N. Amplification of miniature inverted-repeat transposable elements and the associated impact on gene regulation and alternative splicing in mulberry (Morus notabilis). *Mob. DNA* **10**, 27 (2019).

45. Martin, S. L., Li, W. L. P., Furano, A. V. & Boissinot, S. The structures of mouse and human L1 elements reflect their insertion mechanism. *Cytogenet. Genome Res.* **110**, 223–228 (2005).

46. Lavatine, L. *et al.* Single strand transposition at the host replication fork. *Nucleic Acids Res.* **44**, 7866–7883 (2016).

47. Marand, A. P. *et al.* Historical meiotic crossover hotspots fueled patterns of evolutionary divergence in rice. *Plant Cell* **31**, 645–662 (2019).

48. Costas, C. *et al.* Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nat. Struct. Mol. Biol.* **18**, 395–400 (2011).

49. Hénaff, E. *et al.* Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of Brassica species. *Plant J.* **77**, 852–862 (2014).

50. Bouyer, D. *et al.* Genome-wide identification of RETINOBLASTOMA RELATED 1 binding sites in Arabidopsis reveals novel DNA damage regulators. *PLoS Genet.* **14**, e1007797 (2018).

51. Qiu, Y. & Köhler, C. Mobility connects: transposable elements wire new transcriptional networks by transferring transcription factor binding motifs. *Biochem. Soc. Trans.* **48**, 1005–1017 (2020).

52. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, (2011).

53. Sasaki, T. & Project, I. R. G. S. The map-based sequence of the rice genome.

*Nature* **436**, 793–800 (2005).

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

55. Stein, J. C. *et al.* Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat. Genet.* **50**, 285–296 (2018).

56. Hoede, C. *et al.* PASTEC: An automatic transposable element classification tool. *PLoS One* **9**, (2014).

57. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**, (2016).

58. Han, Y. & Wessler, S. R. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, (2010).

59. The R Foundation for Statistical Computing. R Development Core Team. *R A Lang. Environ. Stat. Comput.* (2011).

60. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

61. Caye, K., Jumentier, B., Lepeule, J. & François, O. LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies. *Mol. Biol. Evol.* **36**, 852–860 (2019).

62. Mather, K. A. *et al.* The extent of linkage disequilibrium in rice (Oryza sativa L.). *Genetics* **177**, 2223–2232 (2007).

63. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

64. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

65. Phanstiel, D. H., Boyle, A. P., Araya, C. L. & Snyder, M. P. Sushi.R: Flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**, 2808–2810 (2014).

66. Bushnell, B. *BBMap: a fast, accurate, splice-aware aligner*. *Joint Genome Instritute, department of energy* **13**, (2014).

67. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, (2011).

**Table 1. Number of TE-families per TE order showing group-specific amplification.** The number of families showing a coefficient of variation among varietal groups (CV) of more than 20% are shown with respect to the total number of families for each specific TE order. Only the LTR-RT order is significantly enriched (Fischer exact test) in families that show group-specific amplification (shown in bold)

| **Order** | **CV > 20 %** | **N. families** | **%** | **pvalue** |
|---|---|---|---|---|
| Helitrons | 1 | 29 | 3,45 | 0.0708 |
| TIR-TEs | 34 | 168 | 20,24 | 0.5843 |
| MITEs | 62 | 400 | 15,50 | 0.1094 |
| LINEs | 19 | 104 | 18,27 | 1,00 |
| LTR-RTs | 34 | 112 | 30,36 | **0.0062** |
| SIINEs | 0 | 8 | 0,00 | 0.6171 |

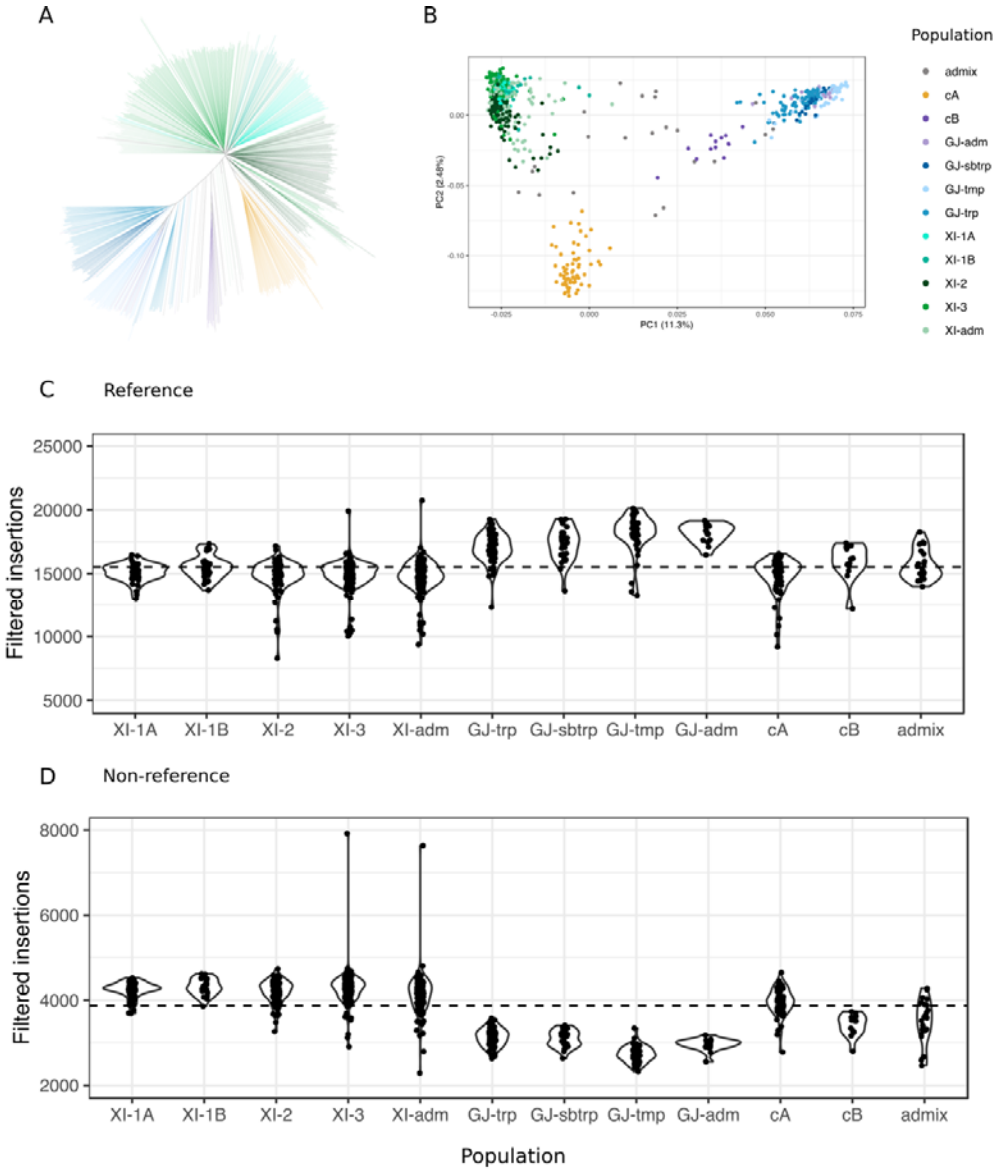



**Figure 1. TIP-based population structure and TIP content in 738 rice varieties.** A) Neighbor-joining tree. B) PCA based on a presence/absence matrix of 103,103 insertion polymorphisms. Colors denote the nine SNP-based sub-populations described by Wang et.al (2018). Reference (C) and non-reference (D) insertions per variety grouped by subpopulation.

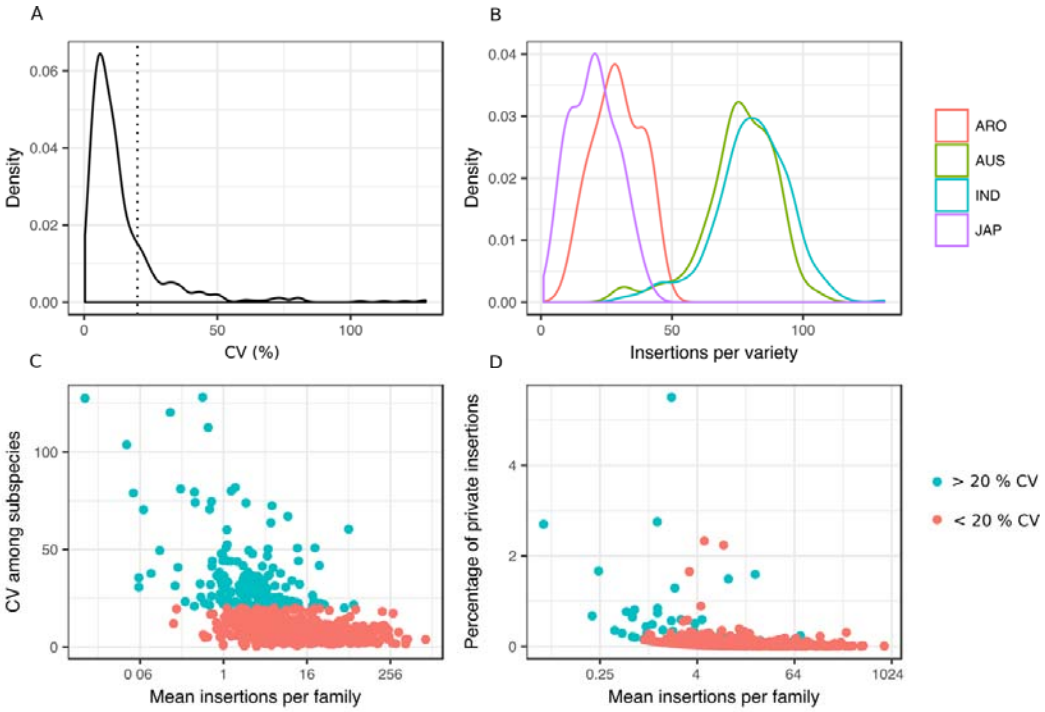



**Figure 2. Signatures of subspecies-specific activity in rice TE families**.
A) Distribution of inter-subspecies Coefficient of Variation in the amount of insertions. Dotted line represents a CV of 20%. B) Example of a Gypsy LTR-retrotransposon family (RLX_comp_MH63_B_R175_Map20) showing signs of amplification in Indica and Aus subspecies after their split with Japonica and Aro. C) Mean insertion per family *versus* CV. D) Mean insertion per family *versus* the number of private insertions.

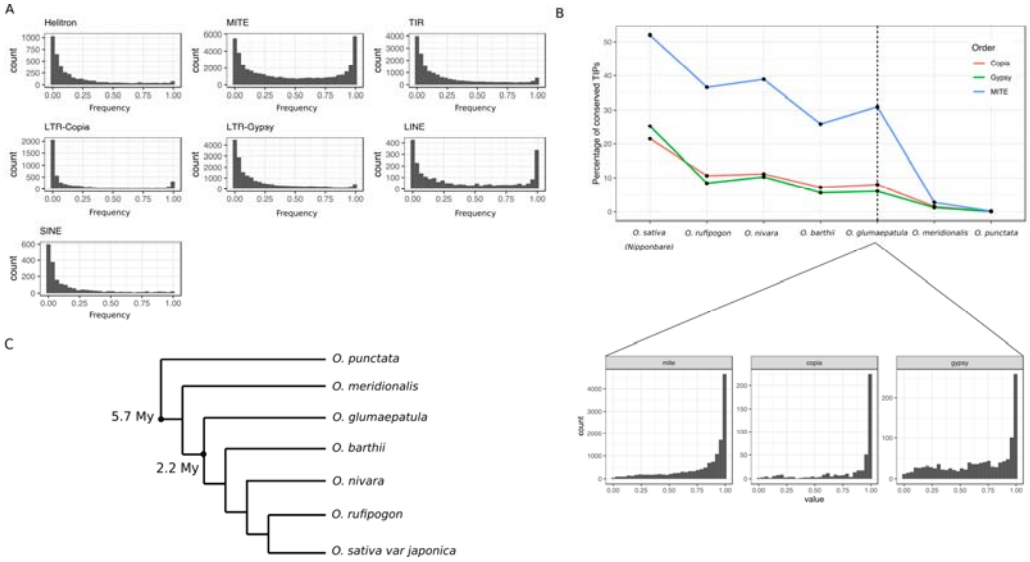



**Figure 3. TIP population frequencies and their conservation in wild *Oryzas*.** A) TIP frequencies in 382 rice traditional varieties. Number of TIPs per order are: 4,113 Helitrons, 45,837 MITEs, 18,592 TIRs, 5026 LTR-Copia, 18,742 LTR-Gypsy, 2,402 LINEs and 2,000 SINEs. B) Percentage of Population TIPs present in *O. sativa* Nipponbare and in 6 wild *Oryzas*. Lower panels show the frequency of the TIPs conserved between Nipponbare and *O. glumaepatula* in the traditional rice varieties. C) Dendrogram representation of the phylogenetic relationships between the *Oryza* species analyzed. Branch lengths do not represent real phylogenetic distances. Divergence times were obtained from http://www.timetree.org/.

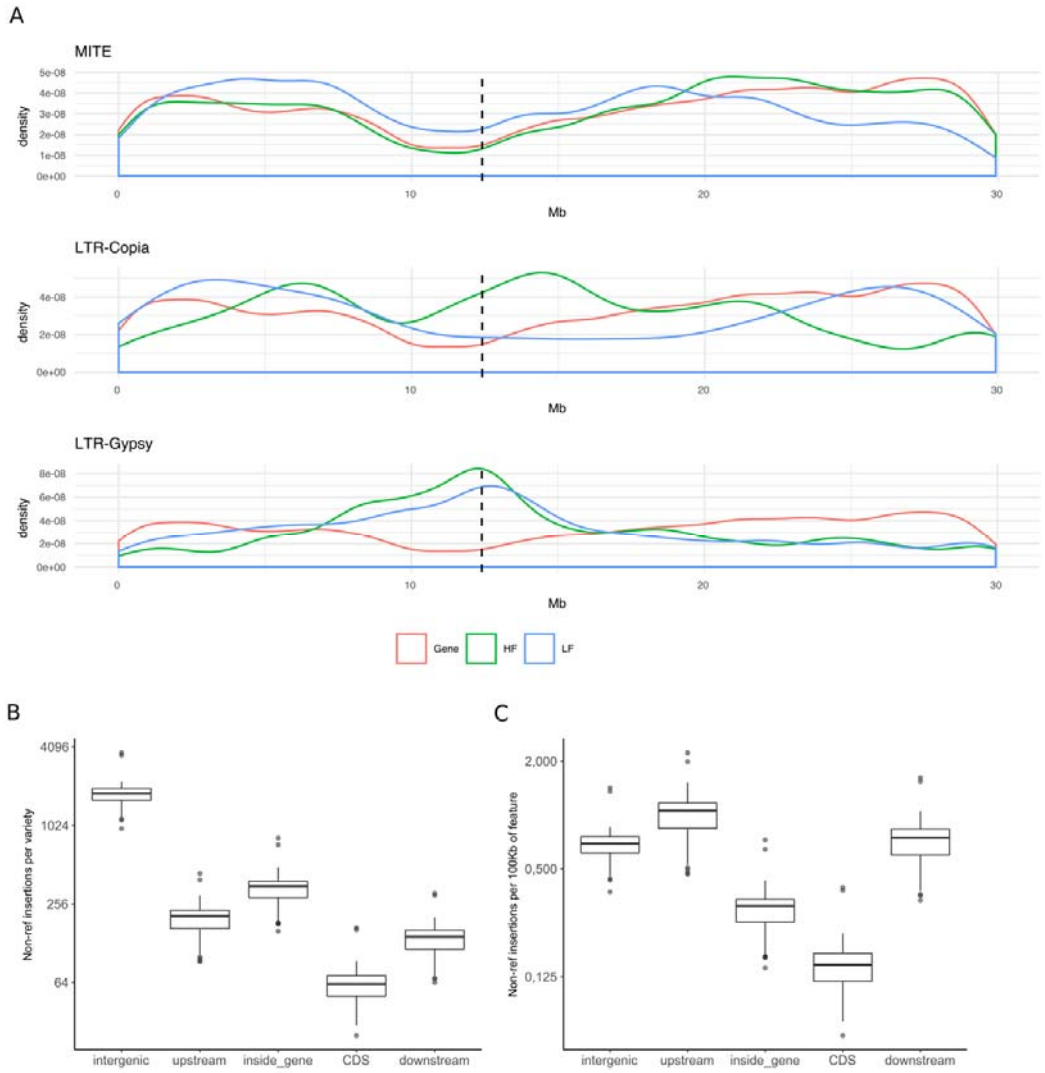



**Figure 4. Genome-wide distribution of TIPs and their impact on genes.**
A) Density plot showing the chromosomal distribution of High-frequency (HF) and Low-Frequency (LF) TIPs of MITEs, Gypsy and Copia LTR-retrotransposon and genes (Chr04). B) Number of non-reference MITE insertions per variety. C) Number of non-reference MITE insertions per 100Kb of genomic features.

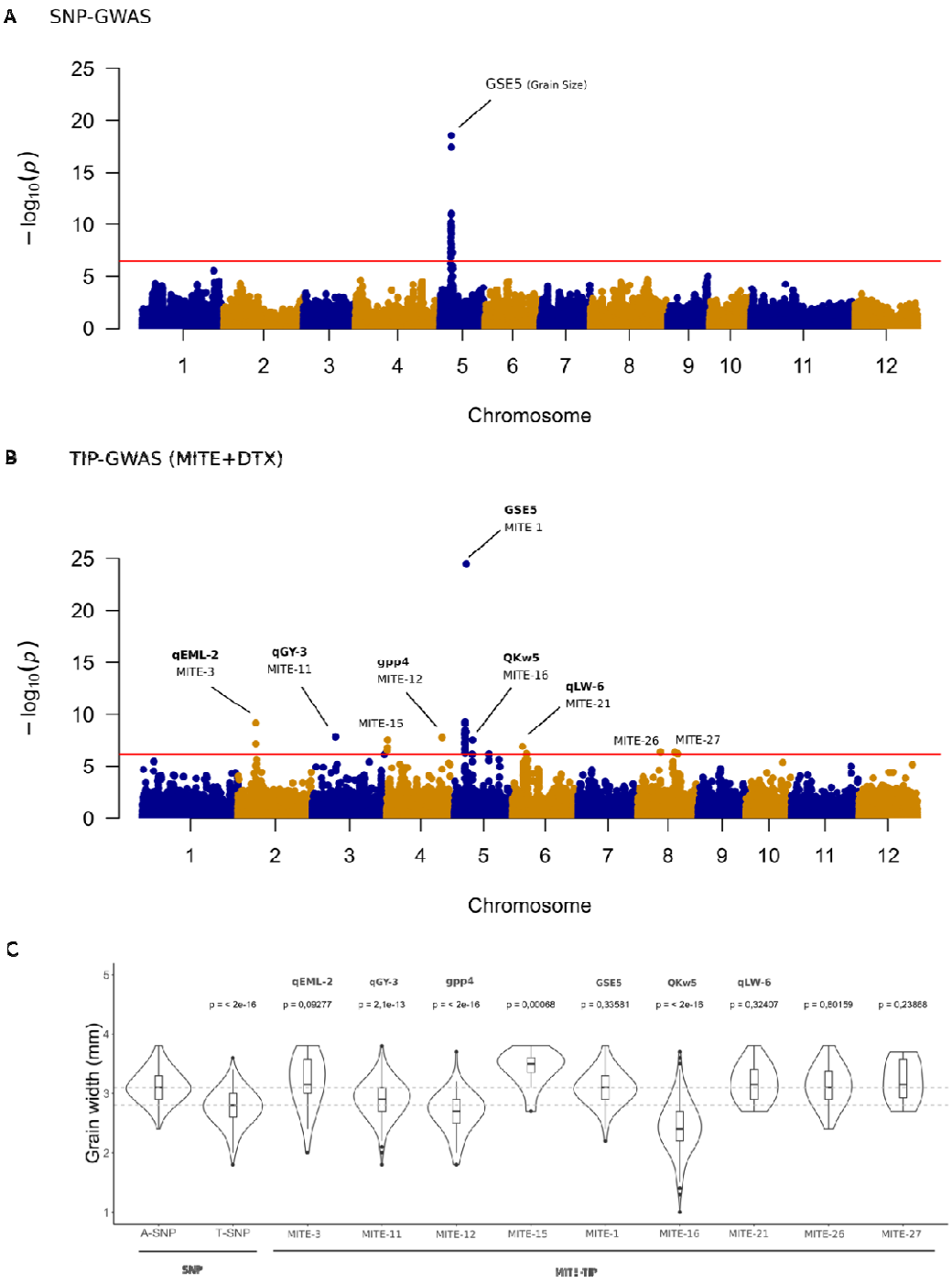



**Figure 5. Association of MITE transposon insertions with higher rice grain width.** A) Rice grain width SNP GWAS using 451 rice indica varieties. B) Rice grain width TIP GWAS using MITEs and DTX insertion polymorphisms as genotype. Red line represents the Bonferroni-adjusted significant threshold (3.2E-07). Significant peaks overlapping known seed QTLs are marked in the manhattan plots. C) Violin plots showing the rice grain width in the different subsets of varieties carrying each variant of the leading SNP at position chr05:5361195 (left panel), coinciding with the upstream region of the seed gene GSE5. Rice grain width of varieties carrying every significant

TIP are shown in the right panel. Differences between each distribution and the control (Genotype “A” at chr05: 5361195) were tested using a Wilcoxon two-tailed test (p value cutoff = 0.05). More information about the significant MITE-TIPs is shown in Supplementary table S3.

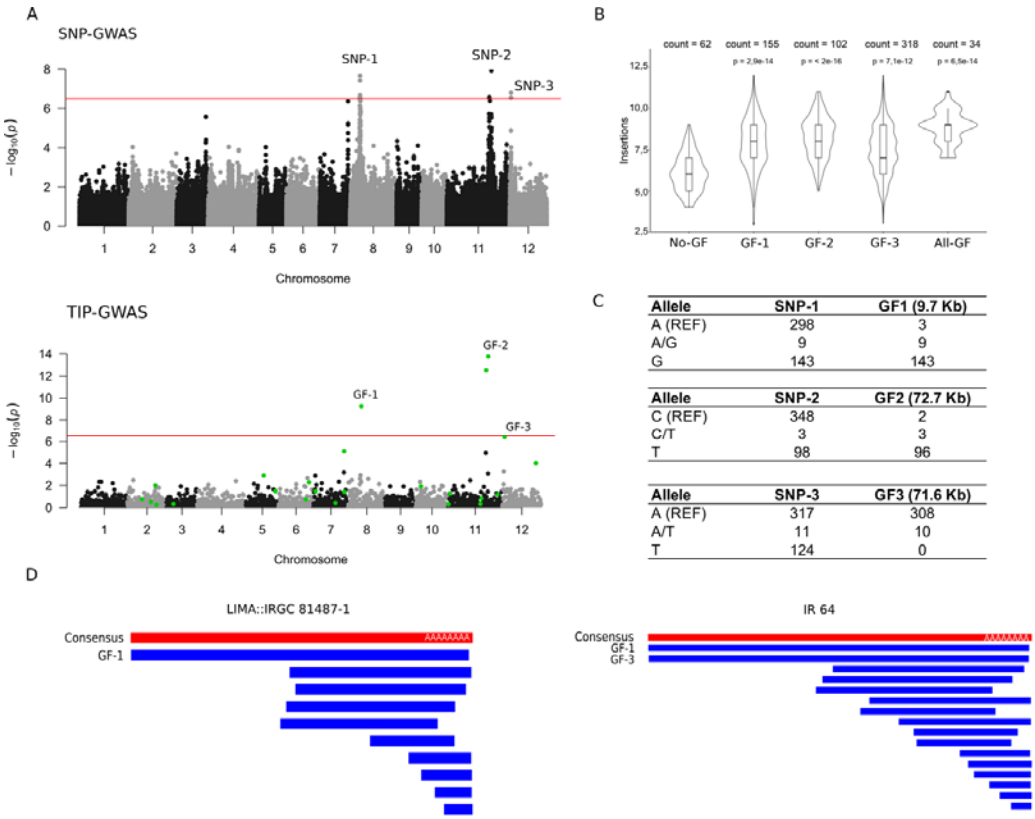
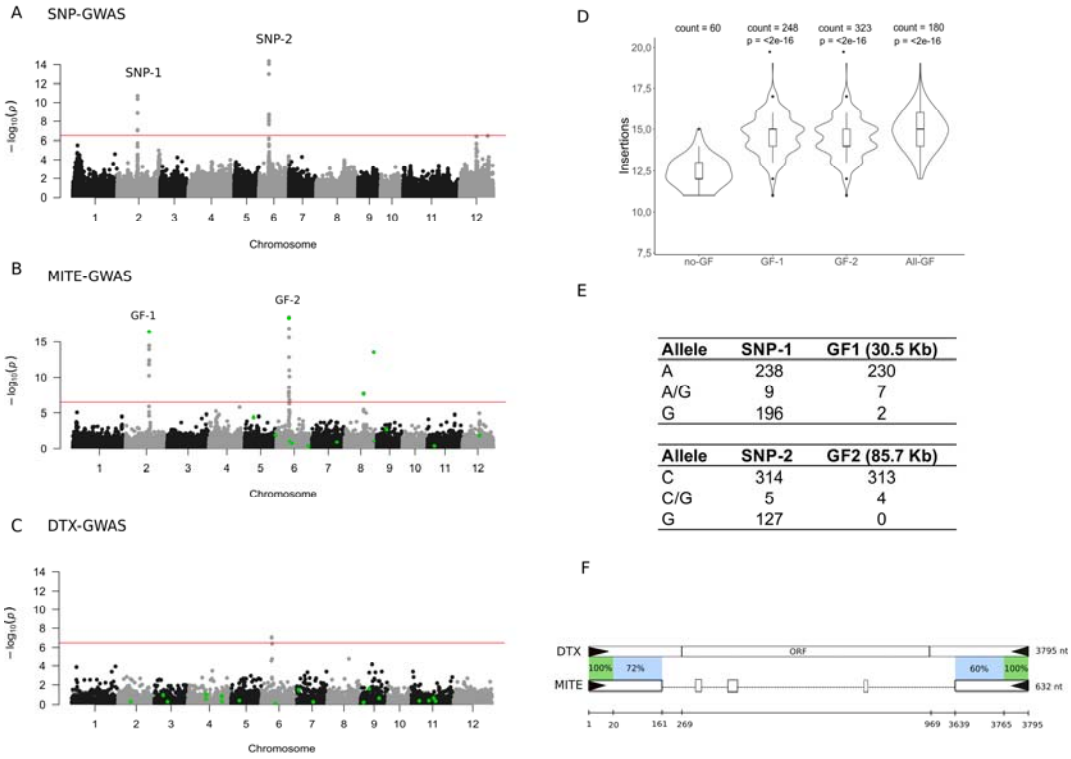



**Figure 6. Identification of full-length copies as main genetic factors responsible for LINE amplification in indica rice.**

A) SNP-GWAS and TIP-GWAS (451 indica varieties) analysis of LINE family RIX_comp_MH63_B_R3107_MAP6 using insertion numbers as phenotype. SNPs (upper manhattan plot) or LINE TIPs (lower manhattan plot) are used as genotype. Red line represents the Bonferroni-adjusted significant threshold (3.2E-07). In green, LINE TIPs of the target family (RIX_comp_MH63_B_R3107_MAP6). GF = Genetic factor. B) Mean insertion numbers of varieties containing or lacking the putative genetic factors. The number of varieties present in each category are indicated upstream each violin. P-values correspond to Wilcoxon tests comparing each category *vs* the no-GF category. C) Genotypes of the leading SNPs in all indica varieties and in the subsets of varieties carrying the three different genetic factors. Distance between each GF and the leading SNP is reported in Kb. D) Identification of Genetic Factors in the assembled genomes of LIMA::IRGC 81487-1 and IR64. In red, family consensus sequence. In blue, schematic representation of all genomic BLAST hits obtained using the consensus sequence as query.




| Allele | SNP-1 | GF1 (30.5 Kb) |
|---|---|---|
| A | 238 | 230 |
| A/G | 9 | 7 |
| G | 196 | 2 |

| Allele | SNP-2 | GF2 (85.7 Kb) |
|---|---|---|
| C | 314 | 313 |
| C/G | 5 | 4 |
| G | 127 | 0 |

**Figure 7. Identification of MITE copies as main genetic factors responsible for their amplification in indica rice.**

A) SNP-GWAS analysis of MITE family SE260300235fam318_632 using insertion numbers as phenotype and SNPs as genotype (451 indica varieties). B) MITE TIP-GWAS analysis using insertion numbers as phenotype and all MITE TIPs as genotype. C) TIP-GWAS analysis using insertion numbers as phenotype and all DTX TIPs as genotype. In green, TIPs from the family SE260300235fam318_632 (B) and from its distantly related transposase DTX_comp_IRGSP_B_R1932 (C). Red line represents the Bonferroni-adjusted significant threshold (3.2E-07). D) Mean MITE insertion numbers of varieties containing or not the putative genetic factors (GF). The number of varieties present in each category are indicated upstream each violin. P-values correspond to Wilcoxon tests comparing each category *vs* the no-GF category E) Genotypes of the leading SNPs in all indica varieties and in the subsets of varieties carrying the three different genetic factors. Distances between each GF and the leading SNP is reported in Kb. F) Schematic representation of the nucleotide conservation between the consensus sequences of the full DTX element DTX_comp_IRGSP_B_R1932 and its related MITE SE260300235fam318_632.