The past decades have revealed an unexpected yet prominent role of so-called ‘junk DNA’ in the regulation of gene expression, thereby challenging our view of the mechanisms underlying phenotypic evolution. In particular, several mechanisms through which transposable elements (TEs) participate in functional genome diversity have been depicted, bringing to light the ‘TEs bright side’. However, the relative contribution of those mechanisms and, more generally, the importance of TE-based polymorphisms on past and present phenotypic variation in crops species remain poorly understood. Here, we review current knowledge on both issues, and discuss how analyses of massively parallel sequencing data combined with statistical methodologies and functional validations will help unraveling the impact of TEs on crop evolution in a near future.

Keywords: crop genomes; next generation sequencing; structural variation; functional validation; phenotypic variation; transposon insertion

INTRODUCTION

Transposable elements (TEs) were discovered 65 years ago in maize [1]. This has placed plants at the frontline for investigating the impact of TEs on genome structure and expression. Discovery of TEs and their genomic impact has also been leveraged by the sequencing efforts that have been particularly abundant for economically important plant species: first through cloning and sequencing of large genomic regions containing genes of agronomic interest [2–5], and more recently through whole genome sequencing (reviewed in [6]). While only a few genomes were sequenced in the early 2000s, development of so-called ‘next generation sequencing’ (NGS) and accompanying assembly methodologies has led to an explosion of the number of crop genomes sequenced, with 16 new crop genomes sequenced since 2012 [6].

While sequenced crop genomes cover distinct phyla, they exhibit common features: all plant genomes are polyploids or display traces of ancient polyploidy [7] and their largest genomic fraction is composed of TEs [8]. Among TEs, long terminal repeat (LTR) retrotransposons are often the most prevalent (Figure 1) because they are both more numerous and larger than DNA transposons. DNA TEs are also more difficult to annotate than LTR retrotransposons: DNA transposon annotation relies on sequence similarity to known transposase sequences, while LTR retrotransposons can also be identified based on detection of structural features (e.g. pair of LTRs). DNA transposons are therefore likely more abundant than depicted by current genome sequence analyses.

Crop genomes also differ in terms of genetic diversity [9], genome size [10], TE content (Figure 1)
and evolutionary potential [11]. Inspection of TE content of crop genomes in a phylogenetic context reveals no congruence in TE content even in closely related species—except perhaps in the genus *Solanum* (tomato and potato)—and highlights differences in success of TEs types (Figure 1). For instance, *Gypsy* LTR retrotransposons have been very successful at invading the maize and grapevine genomes, while non-LTR retrotransposons are prevalent in sorghum, barley, potato and tomato (Figure 1). Of course, these patterns are highly dependent on the quality of genome assemblies and annotation, the latter depending both on the age of the TEs found in a given genome (recent TEs are more easily detected than older ones) and the methods used for detection. In the next few years, genome assemblies will be improved. Use of a common method to annotate TEs from all crop species will then provide an unbiased picture of the trends observed. It is interesting to note that because TEs differ in their insertional properties (e.g. some are more often closely associated to genes than others [12]), the composition of TEs in a
genome may ultimately affect the potential for functional variation to be impacted by TEs.

When sequencing efforts include more than one genotype, an unsuspected level of structural variation is often found. For instance, comparison of a 1.2 Mb contiguous region of the indica and japonica subspecies of rice revealed that 13% of the sequence was not shared, the vast majority of the unshared regions being explained by differential TE insertions [32]. But this is particularly striking for maize, where as much as 76% of the genomic sequence between two inbred lines was unshared, mainly caused by TE presence/absence polymorphisms [33]. These variations likely predate maize domestication, but occurred within the past 4 million years. From both inter- and intra-specific analyses, it is now clear that TE differential insertions have shaped crop genomes. But whether this variation is a major contributor of phenotypic changes and, as such, has or could contribute to crop adaptation to human needs is still an open question. In this article, we review current knowledge about beneficial TE insertions in crop species and discuss how combining NGS, statistics and functional validation may be used to paint a more thorough picture of the impact of TE variation on crop phenotypic evolution.

**BENEFICIAL TE INSERTIONS IN CROP GENOMES**

TEs have long been recognized for their mutating potential, which generates genetic diversity upon which selection can act. Studies from the 1980s highlighted the adaptive potential of TEs in bacteria [34] and later in Drosophila in response to selection for bristle number [35]. In crops, active elements have been reported in maize and rice (reviewed in [36]) and in tobacco following stress activation (reviewed in [37]). While estimating transposition rate is often challenging [36], it is likely orders of magnitude higher than the nucleotide-base mutation rate, thus making TEs powerful agents of evolutionary changes [38].

Here, we list a total of 51 examples of TE-induced phenotypic changes associated with domestication or diversification of cultivated plants (Table 1). While in most cases TE causality has not been experimentally proven, these examples have provided first clues about the underlying mechanisms. We classified them according to the now well-established nomenclature of major TE-triggered mechanisms responsible for phenotypic variation [39]. We have identified five categories of mechanisms across the selected examples: (i) gene inactivation (or severe dysfunction) caused by TE insertions in coding regions or introns, (ii) differential gene expression caused by TE insertions in or in the vicinity of regulatory regions, (iii) TE-mediated genomic rearrangements resulting in gene insertion, deletion or duplication, (iv) transduction-based regulation of gene expression and (v) a few cases of transposase exaptation responsible for modification of transcription factor abilities. The earliest documented cases date back to the 1980s, but most studies were published between 2005 and 2012, revealing that the effect of TE insertions on plant regulation and evolution is being discovered at a growing pace (Figure 2).

These five mechanisms involve both class I and class II TEs. Nevertheless, the two first categories (i and ii), which relate to insertions of TEs in genes or in their vicinity, are mainly caused by class II transposons (61% of the examples of Table 1). This is in accordance with the frequent association of DNA transposons [40, 41] and notably miniature inverted repeats TEs (MITEs) [42–46] with coding sequences or regulatory regions of plant genes. Interestingly, more than 31% of the reviewed cases correspond to gene disruption, 56% of which occur in coding sequences. Such prevalence of coding regions may reflect a detection bias—it seems easier to identify TE insertions that cause a loss of gene function than to detect other types of TE-caused phenotypic variation.

![Figure 2: Yearly distribution of 48 published studies depicting TE-triggered mechanisms responsible for phenotypic variation in crops (also listed in Table 1). GD, gene disruption (insertion in exon or intron); IRR, insertion in or close to regulatory regions; GR, genomic rearrangements; TR, transduction; TEX, transposase exaptation.](image-url)
<table>
<thead>
<tr>
<th>Mutation type</th>
<th>Gene</th>
<th>TE class, order, superfamily</th>
<th>Phenotype</th>
<th>Validation</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertions/deletions in coding regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zea mays ssp. mays</td>
<td>Insertion in exon</td>
<td>Alcohol dehydrogenase 1 (Adh1)</td>
<td>Class I, LTR retrotransposon, copia-like, Rsf</td>
<td>Allyl alcohol-resistant pollen grain</td>
<td>Molecular and sequence-based characterization of insertion within exon 9 of Adh1 gene</td>
</tr>
<tr>
<td>Z. mays ssp. mays</td>
<td>Insertion in exon</td>
<td>Granule-bound glycosyl transferase, waxy locus (Wx)</td>
<td>Class II, subclass I TIR, HAT-like, Ds</td>
<td>Waxy kernels</td>
<td>Molecular characterization of insertion within cDNA, comparison of expression between alleles with/without insertion</td>
</tr>
<tr>
<td>T. aestivum</td>
<td>Insertion in coding sequence</td>
<td>High-molecular-weight glutenin peptin (Glu-I)</td>
<td>Class I, LTR retrotransposon copia-like</td>
<td>Glutenin seed storage</td>
<td>Molecular and sequence-based characterization of the insertion, restriction mapping</td>
</tr>
<tr>
<td>Pisum sativum</td>
<td>Insertion in a exon near the 3' end of the gene</td>
<td>Isoform of starch-branching enzyme (SBEI) (turgosus locus)</td>
<td>Class II, subclass I TIR, Ac/Ds type</td>
<td>Wrinkled-seed—due to metabolic consequences on starch, lipid and protein biosynthesis in the seed</td>
<td>Molecular characterization of the insertion within cDNAs, linkage analysis of RFLPs/wrinkled phenotype</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>Insertion in exon</td>
<td>Floral organ number1 (FON1)</td>
<td>Class II, subclass I TIR, HAT-like, Ds-like, dBrk0</td>
<td>Alteration of floral number and architecture of primary rachis branches</td>
<td>Genotyping, TD and AFLP analysis</td>
</tr>
<tr>
<td>O. sativa</td>
<td>Insertion in exon</td>
<td>Thumbelina (Thl)</td>
<td>Class II, subclass I TIR, HAT-like, nDart1</td>
<td>Mutatable dwarfism</td>
<td>TD</td>
</tr>
<tr>
<td>O. sativa</td>
<td>Insertion in exon</td>
<td>Granule-bound starch synthase (GBSS), Wx</td>
<td>Class I, LTR retrotransposon, Dasheng-like</td>
<td>Glutinous rice</td>
<td>Molecular and sequence-based characterization of the insertion, expression analysis using RT-PCR and Northern blot</td>
</tr>
<tr>
<td>Glycine max</td>
<td>Insertion in exon</td>
<td>GlyPhyA2 (one of paralogs encoding phytochrome A)</td>
<td>Class I, LTR retrotransposon Ty1/copia-like, SORL1</td>
<td>Photoperiod insensitivity</td>
<td>Molecular and sequence-based characterization of insertion within exon1 (genomic DNA), comparison of expression between alleles with/without insertion, comparison of NILs phenotype</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>Excision from exon</td>
<td>Phavonoid 3', 5’-hydroxylase</td>
<td>Class II, MITE Stowaway-like</td>
<td>Skin color change obtained from somaclonal variant (from leaf protoplasts culture)</td>
<td>Anthocyanin content, sequencing, Southern blot (F)</td>
</tr>
<tr>
<td>Insertions in introns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zea mays ssp. mays</td>
<td>Insertion in intron</td>
<td>Knotted 1</td>
<td>Class II, subclass I TIR, Mutator</td>
<td>Leaf development</td>
<td>Molecular and sequence-based analysis of the insertions, correlation of phenotypic changes with structural variation, analysis of methylation</td>
</tr>
<tr>
<td>S. bicolor</td>
<td>Insertion in the second intron</td>
<td>Y (member of the MYB-transcription factor family)</td>
<td>Class II, subclass I TIR, CACTA-like CANDYSTRIPED</td>
<td>Variegated pericarp color</td>
<td>Molecular characterization of insertion, RFLP- and PCR-based detection of excision in red revertant</td>
</tr>
<tr>
<td>M. domestica (Rae lme, Seedless and Wellington Bloomless)</td>
<td>Insertion in intron 4 (Rae lme) or intron 6 (Spencer Seedless and Wellington Bloomless)</td>
<td>Apple homolog of Arabidopsis patellata (MdPi)</td>
<td>Class I, LTR retrotransposon, dem1</td>
<td>Parthenocarpic fruit production</td>
<td>Molecular and sequence-based analysis of the insertion, expression analysis using Northern blot</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Mutation type</th>
<th>Gene</th>
<th>TE class, order, superfamily</th>
<th>Phenotype</th>
<th>Validation</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. max</td>
<td>Insertion in intron</td>
<td>Flavanone 3-hydroxylase gene 1 (Wp locus)</td>
<td>Class II, subclass I TIR, CACTA-like, Tgm-Express1</td>
<td>Coloration flower and protein content of seed</td>
<td>Molecular and sequence-based analysis of the insertion, association of insertion with expression and phenotype</td>
</tr>
<tr>
<td>O. sativa</td>
<td>Insertion in untranslated upstream sequence</td>
<td>Nuclear-coded essential chloroplast protease (OsOmpS)</td>
<td>Class II, subclass I TIR, MAT-like, nDart1</td>
<td>Leaf variegation</td>
<td>Structural and sequence-based characterization of germinal revertants</td>
</tr>
<tr>
<td>G. max</td>
<td>Insertion in intron</td>
<td>Dihydroflavonol-4-reductase 2 (Dfr2)</td>
<td>Class II, subclass I TIR, CACTA-like, Tgm9</td>
<td>Variegated flower color</td>
<td>Variegated phenotype, molecular and sequence-based analysis of insertion, analysis of anthocyanins and flavonol content, expression analysis</td>
</tr>
<tr>
<td>Z. mays ssp. mays</td>
<td>Insertion in two introns and one in intron/exon junctions</td>
<td>Granule-bound starch synthase Wx</td>
<td>Class I, LTR retrotransposon</td>
<td>Waxy kernels</td>
<td>Molecular characterization of insertion, analysis of gene expression and alternative splicing</td>
</tr>
<tr>
<td>Insertions (more rarely deletions) in gene regulatory regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z. mays ssp. mays</td>
<td>Insertion 63 bp upstream of TATA box, affecting the steady-state level of mRNA and protein</td>
<td>Bronze 1, involved in production of anthocyanin pigments</td>
<td>Class II, subclass I TIR, Ds</td>
<td>Aleurone pigmentation</td>
<td>Molecular and sequence-based characterization of insertion, Northern blot, and enzyme assays compared between mutants and revertants</td>
</tr>
<tr>
<td>Z. mays ssp. mays</td>
<td>Excision of element from untranslated leader</td>
<td>Alcohol dehydrogenase 1 (ADH)</td>
<td>Class II, subclass I TIR, Ds1</td>
<td>Differential expression in pollen</td>
<td>Sequence analysis, phenotyping characterization of revertants, quantitative RTase protection assays</td>
</tr>
<tr>
<td>Z. mays ssp. mays</td>
<td>Insertion into the 5' regulatory region</td>
<td>Y1 phytoene synthase</td>
<td>Class II, MITE, Tourist-like, Ins2</td>
<td>Endosperm color</td>
<td>Analysis of selection traces, association with phenotype (but not the only polymorphism associated)</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>Insertion in the 5' flanking region</td>
<td>Vvmyba1, regulator of anthocyanin biosynthesis</td>
<td>Class I, LTR retrotransposon gypsy-like, Gtr1</td>
<td>Loss of pigmentation (anthocyanins)</td>
<td>Molecular and sequence-based characterization of insertion, association of TE presence/absence with expression and phenotype</td>
</tr>
<tr>
<td>T. aestivum</td>
<td>Insertion in promoter</td>
<td>TaFT/VRN3</td>
<td>Class I, LTR retrotransposon</td>
<td>Flowering time</td>
<td>Molecular and sequence-based characterization of insertion, allele frequency in germplasm, association of TE presence/absence with gene expression and phenotype, expression and phenotype analysis of plant transformed with construct carrying insertion (F)</td>
</tr>
<tr>
<td>S. bicolor</td>
<td>Insertion in upstream region</td>
<td>Aluminium tolerance locus AtSB</td>
<td>Class II, MITE Tourist-like</td>
<td>Aluminium resistance</td>
<td>Characterization of the insertion, expression and phenotypic analysis of plant transformed with construct carrying insertion (F)</td>
</tr>
<tr>
<td>Z. mays ssp. mays</td>
<td>Insertion ~70 kb upstream a gene (cis-acting regulatory change)</td>
<td>ZmRap2.7, an Ap2-like transcription factor</td>
<td>Class II, MITE Tourist</td>
<td>Flowering time</td>
<td>Positional cloning and association mapping</td>
</tr>
<tr>
<td>V. vinifera</td>
<td>Insertion in a promoter</td>
<td>Putative regulator of anthocyanin synthesis (W MMB1)</td>
<td>Class I, LTR retrotransposon gypsy-like</td>
<td>Fruit color</td>
<td>Characterization of the insertion, comparison of sequence, transient assays</td>
</tr>
<tr>
<td>Mutation type</td>
<td>Gene</td>
<td>TE class, order, superfamily</td>
<td>Phenotype</td>
<td>Validation</td>
<td>References</td>
</tr>
<tr>
<td>---------------</td>
<td>------</td>
<td>------------------------------</td>
<td>-----------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>O. sativa</strong></td>
<td>Insertion upstream gene</td>
<td>Resistance gene Pit</td>
<td>Class I, LTR retrotransposon copia-like, Renovator</td>
<td>Disease resistance</td>
<td>Characterization of the insertion, construction of chimeric plants, analysis of pathogenicity</td>
</tr>
<tr>
<td><strong>C. melo</strong></td>
<td>Insertion near the promoter provokes its methylation</td>
<td>C2H2 zinc-finger domain transcription factor of the WIP protein subfamily</td>
<td>Class II, subclass I TIR, HAT-like, Gyno-HAT</td>
<td>Sex determination: carpel development, stamen repression</td>
<td>Molecular and sequence-based characterization of insertion, DNA methylation profile, TILLING, association of TE presence/absence with gene expression and phenotype</td>
</tr>
<tr>
<td><strong>Brassica oleracea</strong></td>
<td>Insertion in the upstream regulatory region</td>
<td>Purple (Pr)</td>
<td>Class II, subclass I TIR, Harbinger</td>
<td>Purple coloration</td>
<td>Association mapping (S), plant transformation (F)</td>
</tr>
<tr>
<td><strong>V. vinifera</strong></td>
<td>Insertion in the promoter</td>
<td>VvTFL1 A</td>
<td>Class II, subclass I hAT-like, Hatvine1-rrm</td>
<td>Inflorescence structure</td>
<td>Molecular characterization of the insertion, comparison of expression profile and presence of the insertions</td>
</tr>
<tr>
<td><strong>G. max</strong></td>
<td>Insertion in the promoter</td>
<td>Dihydroflavonol-4-reductase 2 (DFR2)</td>
<td>Class II, subclass I TIR, CACTA-like, Tgm9</td>
<td>Diluted (pale) flower color</td>
<td>Molecular and sequence-based characterization of insertion, analysis of anthocyanins and flavonols content, expression analysis, sequencing</td>
</tr>
<tr>
<td><strong>Triticum turgidum</strong></td>
<td>Insertion in 5' UTR, near regulatory sites</td>
<td>Vrn-B1, one of the genes controlling vernalization</td>
<td>Class I, LTR retrotransposon, gypsy-like</td>
<td>Expression of the gene without vernalization</td>
<td>QTL analysis, cloning and sequencing, association of TE presence/absence with gene expression</td>
</tr>
<tr>
<td><strong>Z. mays</strong> ssp. <em>mays</em></td>
<td>Insertion in regulatory region localized about 60 kb of the ORF</td>
<td>TCP family transcriptional regulator B3 domain, hopscotch (tb1)</td>
<td>Class I, LTR retrotransposon apio-like, hopscotch</td>
<td>Increase of apical dominance</td>
<td>Test of selection, haplotype group analysis, phenotypic fine-mapping, transient assays of reporter gene under control of regulatory sequence with and without the insertion</td>
</tr>
<tr>
<td><strong>C. sinensis</strong></td>
<td>Insertion in upstream regulatory region leading to TE-derived new transcription start site</td>
<td>Ruby, a MYB transcriptional activator of anthocyanin production</td>
<td>Class I, LTR retrotransposon apio-like, Tcs1</td>
<td>Cold dependent induction of anthocyanin biosynthesis in fruit</td>
<td>Molecular and sequence-based characterization of insertion, association of insertion with expression and phenotype, RACE-PCR based characterization of new transcript start site</td>
</tr>
<tr>
<td><strong>C. sinensis</strong></td>
<td>Insertion in upstream regulatory region</td>
<td>Ruby, a MYB transcriptional activator of anthocyanin production</td>
<td>Class I, LTR retrotransposons copia-like, Tcs2</td>
<td>Cold dependent induction of anthocyanin biosynthesis in fruit</td>
<td>Molecular and sequence-based characterization of insertion</td>
</tr>
<tr>
<td><strong>Z. mays</strong> ssp. <em>mays</em></td>
<td>Insertion near a regulatory sequence</td>
<td>Auxin-binding protein 1</td>
<td>Class II, MITE, Torist; Class II, subclass I, Ac-like, Algrim; Class I, solo-LTR, Botuta</td>
<td>Unknown</td>
<td>Expression of reporter gene (luciferase) under control of regulatory sequence with and without the insertion</td>
</tr>
<tr>
<td><strong>O. sativa</strong></td>
<td>Insertion into the 5' UTR of the gene</td>
<td>Chalcone isomerase (OsCHI)</td>
<td>Class I, LTR retrotransposon, Dasheng</td>
<td>Golden hull and internode coloration</td>
<td>Complementation test (F)</td>
</tr>
<tr>
<td><strong>Brassica napus</strong></td>
<td>Insertion in the ORF</td>
<td>TRANSPARENT TEST8 locus (BTT8)</td>
<td>Class II, subclass II Heliton</td>
<td>Yellow seed coloration</td>
<td>Construction of linkage map, functional complementation (F)</td>
</tr>
<tr>
<td><strong>Z. mays</strong> ssp. <em>mays</em></td>
<td>Tandem insertion at 289 nt of the gene</td>
<td>Cytochrome P450-like</td>
<td>Class II, MITE, Tourist-like, ZEmVl-9</td>
<td>Flowering time</td>
<td>Association mapping (S)</td>
</tr>
<tr>
<td><strong>Z. mays</strong> ssp. <em>mays</em></td>
<td>Insertion in 3' untranslated region</td>
<td>ZmGE2 (gene encoding a cytochrome P450 protein)</td>
<td>Class II, subclass I TIR, Mutator</td>
<td>Kernel oil content</td>
<td>Characterization of the insertion, association analysis of expression profile and presence of the insertion</td>
</tr>
<tr>
<td>Mutation type</td>
<td>Gene</td>
<td>TE class, order, superfamily</td>
<td>Phenotype</td>
<td>Validation</td>
<td>References</td>
</tr>
<tr>
<td>---------------</td>
<td>------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>Z. mays ssp. mays</strong></td>
<td>Insertion in promoter</td>
<td><em>ZmCCT</em> (encoding a CTT domain protein)</td>
<td>Class II, subclass I TIR, CACTA-like</td>
<td>Flowering time</td>
<td>Genome-wide association study, transformation</td>
</tr>
<tr>
<td><strong>Rearrangements</strong></td>
<td>TE-mediated chromosomal rearrangement</td>
<td>R-v complex transcription factor regulating the anthocyanin pigment pathway</td>
<td>Class II, subclass I TIR, CACTA-like, dppio</td>
<td>Modification of tissue-specificity of the duplicated genes</td>
<td>Molecular and sequence-based characterization of insertion and rearrangement breakpoints, expression and genetic analysis of aleurone mRNA populations in vivo</td>
</tr>
<tr>
<td><strong>Triticum and Aegilops species</strong></td>
<td>TE-mediated insertion of a 24.7 kb rearranged fragment inducing gene duplication and expression modification from new cis-acting elements</td>
<td>Hardness (Ho) locus</td>
<td>Class I retrotransposons (different families)</td>
<td>Variation in grain hardness</td>
<td>Sequence-based characterization of the corresponding locus</td>
</tr>
<tr>
<td><strong>S. lycopersicum</strong></td>
<td>TE-mediated insertion of a 24.7 kb rearranged fragment inducing gene duplication and expression modification from new cis-acting elements</td>
<td>SUN locus (IQD2 gene)</td>
<td>Class I LTR retrotransposon copia-like, Rider</td>
<td>Bongsted fruit shape</td>
<td>Molecular and sequence-based characterization of insertion and rearrangement breakpoints, analysis of fruit shape index and gene expression in transformants (i) carrying deleted versions of the rearrangement, (ii) overexpressing the gene and (iii) underexpressing the gene through RNAi</td>
</tr>
<tr>
<td><strong>Transduction</strong></td>
<td>Plasma membrane proton ATPase gene, Mha1</td>
<td>Class I, LTR retrotransposon, BsaI</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Molecular and sequence-based comparison of TE and complete gene</td>
</tr>
<tr>
<td><strong>Z. mays ssp. mays</strong></td>
<td>Fragments of plasma membrane protein ATPase gene, xylan endohydrolase gene, beta-1,3-glucanase gene</td>
<td>Class I, LTR retrotransposon, BsaI</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Spatio-temporal expression (transcription and translation) of hybrid ORF combining the retrotransposon gag domain and the three gene fragments, reconstruction of sequence evolution, test of selection</td>
</tr>
<tr>
<td><strong>Z. mays ssp. mays</strong></td>
<td>Fusion of the coding and regulatory sequences of two linked paralogous genes</td>
<td>Myb-homologous transcriptional regulator (p1)</td>
<td>Class II, subclass I TIR, HAT-like, Ac</td>
<td>Kernel color and cob glume</td>
<td>Molecular characterization of the transposition, comparison of expression analysis and phenotype</td>
</tr>
<tr>
<td><strong>Transposase exaptation</strong></td>
<td>Gary hAT-transposase like</td>
<td>Class II, subclass I TIR, HAT-like, gary</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Comparison of gary genes and transposase at the structural and sequence levels, phylogeny</td>
</tr>
<tr>
<td><strong>Z. mays ssp. mays</strong></td>
<td>CYP72A27 -Zm gene duplicated from the monoxygenase (P450) gene</td>
<td>Class II, subclass II Helitron</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Characterization of the insertion at the sequence level, reconstruction of sequence evolution</td>
</tr>
</tbody>
</table>

(continued)
mutations such as TE-mediated local genome rearrangements requiring sequencing of large genomic regions and for which the demonstration of the causative effect remains challenging. This is particularly clear from the observation that older studies in Table 1 (eight papers published between 1985 and 1994) report the characterization of TE insertions resulting in direct gene disruption. Several of these studies revealed TE insertions in genes underlying agronomic traits such as seed content of amylose (maize, [47]), high-molecular weight glutenin (wheat, [48]) or starch (pea, [49]), which were cloned following fine mapping or genetic linkage approaches.

The predominant TE-mediated mechanism involved in crop phenotypic variation is related to modifications in the regulatory regions of plant genes (45%). This mechanism has been discovered more recently than gene disruption (most studies postdate 2006), which is likely due to parallel improvement of methodologies that allow positional cloning, association mapping in complex traits and the analysis of gene mutations via tilling or plant transformation. With this technological development, more complex traits such as flowering time have been analyzed (wheat, [50]; maize, [46, 51]) and highlighted the more subtle impact of TEs on quantitative variation via modulation of gene expression. Improvement of sequencing also provided the means to identify regulatory sites located tens of kilobases upstream to the target gene, as exemplified by the textbook example of ‘teosinte branched 1’ (tb1), where insertion of a LTR retrotransposon Hopscotch element ~60 kb upstream of the gene [52] induces gene overexpression, leading to apical dominance in maize. Such studies have also brought considerable insights into the epigenetic impact of TE insertions in crops. For instance, transition from male to female flowers in melon is caused by DNA methylation spreading of the class II Gyno-hAT transposon to the promoter of the CmWIPI transcription factor gene [53].

The three other categories of TE-mediated mechanisms (iii, iv and v) have been reported more rarely. The complexity of genome rearrangements mediated by TEs has nevertheless been well illustrated through comparative genomics. For instance, comparative analysis of the Hardness locus sequenced for seven Triticeum and Aegilops species demonstrated the role of various retroelements in gene loss and the further association of deleted haplotypes to trait quality in polyploid wheats [54].
With the recent boom of NGS-based technologies allowing the identification of TE polymorphisms (see below) together with the transcriptome and epigenome characterization, the number of TE-mediated phenotypic changes is expected to greatly increase in the coming years. Hence, NGS-based technologies will soon offer a less biased view of the relative contribution of the different types of mechanisms through which TEs have shaped crop evolution.

**USING NGS TO DETECT TE INSERTION POLYMORPHISMS**

Most of the examples presented in Table 1 were found through the analysis of specific traits and/or elements. Whole genome sequencing allows for detection of potentially all TEs of a given genome—but for most species, the assembled sequence does not fully access the heterochromatic part of the genome, and estimation of TE content is likely underestimated. Nevertheless, sequencing costs remain too high to apply de novo assembly to detect TE polymorphisms at the population/germplasm level. Moreover, complete genome sequences based on de novo assembly of massive parallel data alone are often not of enough high quality to accurately detect TE presence/absence polymorphisms. Alternatively, analysis of sequence reads from massive parallel sequencing data provides unprecedented opportunities to investigate without a priori assumptions the propensity to which TE insertions can underlie adaptive variation. Over the past decade, several methodologies have been developed to detect TE insertion polymorphisms in complex eukaryotic species. In this section, we give an overview of these methodologies, their use and limitations and how they could be improved.

**Sequencing transposon display**

Polymerase chain reaction (PCR)-based methods following genome digestion with restriction enzymes and amplification of TE insertion flanking regions are known as sequence-specific amplification polymorphism (S-SAP, [55]) or transposon display (TD, [56]). They have long been used to detect insertion polymorphisms in plants [57–60]. Several improvements of the original TD technique that allow distinguishing polymorphisms of the restriction site from TE-based polymorphisms include vectorette, splinkerette and digestion–ligation–amplification (DLA)-454. They provide fast, extensive and reliable parallel characterization of TE insertions in many individuals. For instance, mPing-specific vectorette on 24 rice siblings of a self-cross allowed the identification of ~1600 mPing insertion sites [61]. Similarly, DLA-454 applied on active Mu transposon maize lines led to discover 21 new Mu elements and to characterize the distribution of 40 000 non-redundant Mu insertion sites along maize chromosomes [62]. While parallel sequencing of differently tagged amplicons targeting various TEs is feasible, the obvious drawback of TD-derived techniques is that they specifically target the set of TEs chosen for PCR amplification (e.g. mPing, Mu). They therefore do not offer a full characterization of the TE insertion landscape. Availability of a reference genome sequence is useful to localize the TE insertions discovered. Quality of this reference sequence will not highly impact detection of the TE polymorphism itself. Rather, it will impact connection of the insertion with other genomic features, such as nearby genes. To some extent, direct comparison of TD derived sequence reads can also be achieved to detect presence/absence polymorphisms.

**Mapping of single reads on a reference genome**

One way to avoid targeting a set of chosen TE families is to directly analyze sequence reads from genomic DNA, therefore, bypassing the PCR amplification step. Over the past few years, the growing interest for structural variation has led to the development of methodologies allowing for detection of presence–absence and copy-number variants from high-throughput sequencing data (mostly Illumina). These methodologies provide a deep characterization of genome wide TE-based polymorphisms, but are highly dependent on the quality of the reference sequence.

The identification of TE insertion breakpoints with base-pair precision can be achieved by detecting a ‘split mapping’ signature of reads on the reference genome. One split may correspond to an insertion breakpoint, thus separating the sequencing read into one part that matches the genomic region flanking the TE insertion site, and another part that is anchored in the TE sequence. After mapping reads onto the reference genome, uniquely mapping reads are then analyzed. This allows finding reads matching low copy regions flanking TE insertions. A read spanning a region deleted in the sample genome will be mapped twice onto the reference genome (i.e. on
both sides of the insertion), but at each genomic location, only part of the read will be mapped (i.e., each read will be split) (Figure 3A). Long reads (such as 454/Roche or Pacific Bioscience—PacBio) are therefore more adapted, as they more reliably capture partial mappings. Alternatively, as proposed by [63], use of paired reads can provide additional anchoring of the split reads when these are short (Figure 3A). To partially control for false positive, paired-end mapping (PEM) strategies (see below) can be used in complement to detect pairs mapping with larger-than-expected insert size.

**Figure 3**: Split-read-based methodologies to detect TE presence/absence polymorphisms from NGS data. TE sequence is represented as a rectangle, and empty site of insertion is shown as a vertical bar. (A) Detection of a TE insertion that is absent from the sample genome using 'split read' signature. A read spanning a deletion breakpoint will map as a split read in two locations (corresponding to the two sides flanking the insertion) in the reference genome. Long reads help detection as they reduce spurious mapping of the split reads. Alternatively, as shown here, use of paired reads can increase proper anchoring of the split reads. Annotation as a TE sequence of the DNA fragment located between the two signatures of a pair points to a TE presence/absence polymorphism. 

(B) Detection of a new TE insertion in the sample genome using 'split-hanging reads'. Pairs where one read is uniquely mapped, and the other is spanning one of the insertion breakpoints reveal one boundary of a sequence absent from the reference genome. De novo assembly of these hanging reads with reads from pairs where one read is hanging and the other is multiply mapping can be used to walk inside the inserted fragment sequence (construction of a scaffold, depicted as discontinuous blocks). Annotation of the assembled fragment as a TE allows characterization of a new TE insertion in the sample genome. 

(C) Detection of a new TE insertion in the sample genome with the use of long reads. Reads for which the middle section does not align to the reference genome while both 5’ and 3’ ends do point to a sequence absent from the reference genome. Further annotation of the middle section may reveal a TE insertion.
Detection of split reads that map only once but with partial length mapping (or ‘hanging reads’) can be used to detect insertions in the sample genome, but additional clustering work is required to reconstruct the inserted segment (Figure 3B). As proposed by [64], de novo assembly of hanging reads detected as one cluster can be performed to reconstruct the insertion, but will help only if the inserted segment is smaller than the library insert size. For longer TEs, de novo assembly of unmapped reads is also required (Figure 3B).

Finally, if reads are long enough to cover the entire TE length, split read signature detection can be used to directly detect insertions in the sample genome. Because the size of the insertion needs to be smaller than read length, Roche/454 sequencing is suitable for detection of new MITE or SINE insertions, but discovery of insertions of larger elements such as autonomous DNA transposons or LTR retrotransposons requires the use of technologies providing longer reads, such as PacBio (Figure 3C).

**Mapping of paired reads on a reference genome**

PEM (from paired-end or mate pair libraries) can be used to detect structural variants. Paired ends are generated from the sample genome. During library preparation, the distance between the two paired ends is size fractionated, hence the expected distribution of distances between them is known. These paired reads are then mapped onto a reference genome. Pairs of reads for which the distance on the reference genome sequence substantially differs from the expected distribution indicate the presence of an insertion/deletion between the sample and the reference genomes. This can be used to detect TEs that are present in the reference and absent from the sample genome (Figure 4A). They can also be used to detect insertions in the sample genome, but in this case (i) the size of the insertion needs to be smaller than the library insert size and (ii) it does not provide any information on the inserted DNA fragment, therefore, hampering to detect whether it is a TE or not.

**Figure 4:** PEM-based methodologies to detect TE presence/absence polymorphisms from NGS data. TE sequence is represented as a rectangle, and empty site of insertion is shown as a vertical bar. (A) Detection of a TE insertion that is absent from the sample genome. Paired-end reads are generated at a distance that is approximately known and homogeneous (dotted lines). Paired reads are mapped onto a reference genome, and distance between reads is computed. Pairs of reads mapping at a distance that is substantially different from the average expected library length (long dotted lines) suggests the presence of a DNA segment in the reference genome that is absent in the sample genome. Annotation of this fragment as a TE sequence then points to a TE presence/absence polymorphism. (B) Detection of a new TE insertion in the sample genome. Pairs of reads in which one read maps on a unique sequence in the reference genome sequence and the other maps in a TE sequence are detected. Here, a reference TE database is used to search for TE-mapping reads, but search for reads mapping at multiple positions annotated as TEs in the reference genome is also possible.
Alternatively, PEM can be used to detect new TE insertions in the sample genome by searching for pairs in which one read maps unambiguously onto the reference sequence and the other maps to a TE sequence (Figure 4B). This TE sequence can be either multiple genomic regions annotated as TEs, or more simply, come from a TE sequence database that has been collected from genomic sequences of the species. In plants, suitability of the PEM to detect neo-insertions from active TEs was first demonstrated in *Arabidopsis thaliana* and was used to identify TEs activated through the alteration of silencing pathways [65]. It was then successfully applied to detect 34 TE insertions in a rice mutant line regenerated from callus culture [66].

Of course, the efficiency of TE polymorphism detection greatly depends on the quality of the reference genome sequence and that of the TE annotation. This is why its use in crops has been limited to rice. With a median contig length over 100 kb and an assembly covering more than 85% of the genome (reviewed in [6]), other crop species such as sorghum (*Sorghum bicolor*), cucumber (*Cucumis melo*) or peach (*Prunus persica*) are likely to soon provide the level of information required to pursue such approaches. Another parameter that should be considered is the degree to which the reference genome is representative for the whole species. While PEM may be fruitful for species with low structural diversity, such as grapevine [67], they are likely more challenging for outcrossing species with high and complex structural diversity such as maize [33, 68]. Note that PEM does not provide precise detection of the TE insertion site (variant breakpoint), and collection of such information therefore requires further investigation once the TE polymorphisms are detected.

**Depth of coverage**

Under the assumption that sequencing uniformly covers the sample genome, the number of reads mapping to a genomic location of the reference genome follows a Poisson distribution and is expected to be proportional to the number of times the region is present within the genome. Methods similar to those applied to comparative genome hybridization array data (segmentation, change-point analysis) can be used to find signal breakpoints [69, 70]. However, because strength of the depth of coverage (DOC) signature increases with variant size, it is mainly used to detect large variants. Note also that DOC is efficient for detecting duplicates or triplicates, but becomes more challenging when the variants are in high copy numbers in both the sample and the reference genomes, which is often the case for TEs. In addition, it provides a poor characterization of the breakpoints. For these multiple reasons, DOC is usually not the preferred approach to detect TE insertions but can be used in combination other approaches, such as PEM [71, 72].

**Current limitations and future directions**

Several of the methods presented above have been implemented into algorithms that are publicly available. Those include CNVnator, CNVseq, SegSeq, SVdetect, PEMer, VariationHunter, MoDIL, Pindel BreakDancer, HYDRA and SVDetect (reviewed in [73–75]). These softwares may be used *a priori* for analyzing one genotype in different conditions or population-based samples. However, our ability to detect TE polymorphism will ultimately depend on the level of structural diversity between the genotype(s) under survey and the species reference genome. Two recent tools have also been specifically designed to simultaneously discover and characterize TE polymorphisms in large populations: T-lex [76], which detects whether a TE present in the reference genome sequence is present or absent in other individuals, and RelocaTE [77], which has been developed to detect new insertion events from movement of active TEs.

Even with the development of these new tools, we are still facing several challenges inherent to the use of NGS. First, correct mapping on the reference genome sequence is affected by sequencing errors. Second, PEM relies entirely on insert size estimates, which are both sensitive to the presence of chimeric clones and accuracy of library construction. Indeed, while libraries are generally considered homogeneous in size, this is usually questionable. Third, DOC usage is affected by uneven sampling of the genome and may therefore not be robust enough for an accurate prediction of duplications. Because these different methods are sensitive to diverse biases, combining evidence from different types of signature to detect the same event allows for more sensitive detection, better characterization of breakpoints, together with the need of a lesser coverage.

One limitation of all these methodologies is that they are applicable only to uniquely mapped reads (reviewed in [74]) and hence, limited to the detection of TE polymorphisms in the non-repeated
genomic content of the reference and sample genomes. While this is problematic to understand TE-mediated species diversity, it likely provides most of the information important to discover insertions with adaptive value. The main limitation so far is thus the availability of a good genome sequence and a good TE annotation. While the dramatic increase in read size that will arise in coming years will bring solution to assembly issues, efforts need to be made to develop tools allowing for parallel high-level annotation of TEs in multiple genomes.

METHODS AND CHALLENGES FOR THE DETECTION OF SELECTED TE INSERTIONS

Genome-wide characterization of TE-based polymorphisms using NGS technologies does not predict their adaptive contribution. In fact, TE insertions that have been selected for (such as those presented in Table 1) likely represent a negligible fraction of genomic TE content. To capture this fraction, TE-based population genetic studies and association studies can be undertaken. On one hand, TE-based population genetic studies can help revealing past selection of adaptive insertions, the ones that have been selected for during the domestication and/or breeding history of crops (Table 1). On the other hand, TE-based association mapping can assist both characterization of additional adaptive variation in germplasm collections and elucidation of TE contribution to the genetic architecture of quantitative traits.

Current crop literature, however, lacks such studies. So far, only a single study (in maize) reports the use of polymorphic MITE insertions to screen a population, in this case a sample of 26 maize landraces, as well as an association panel of 367 maize inbred lines. Results of this analysis reveal a significant association between male flowering time and a tandem MITE insertion located in the 3' end of a cytochrome P450-like gene [46].

In principle, TE-based markers could be used as any other markers in population genetic studies to investigate the genetic bases of adaptation. However, as demonstrated in several model species both in animals and plants [78–82], TE variants are specific in at least one respect: they are often deleterious and therefore maintained at low population frequencies as a result of purifying selection. As shown in Arabidopsis species, the strength of selection varies among TE families [82] and strongly depends on parameters influencing the effective population size such as demography [83] as well as the mating system [78, 81, 82]. Because neutrality tests aiming at detecting positive selection are often based on frequency spectrum, deviation of TE variants toward low population frequencies deserves specific attention.

First, the strength of selection acting against TE insertions for a given TE family can be estimated [84]. The most recent methodological development [85] incorporates ascertainment bias and demographic inferences. It could be used to estimate the strength of purifying selection independently for multiple TE families in several crop genomes for which demographic scenarios for domestication and diversification have been established—see, for instance, maize [86], rice [87], sunflower [88] and apple [89]. Relying on the hypothesis that TE insertions are unlikely to drift to high frequencies when they belong to families subjected to strong purifying selection, it is then possible to identify putatively selected insertions. This procedure has been proposed and employed successfully in Drosophila melanogaster [90, 91]. But, in theory, estimates of selection coefficients could also serve in coalescent simulations that explicitly model background selection [92] to approximate null distributions of various summary statistics and even ultimately to determine the strength of selection acting on individual TE insertions (although such framework has yet to be developed).

Alternatively to these model-based approaches, TE-insertion genome scans based on empirical distribution of statistics summarizing population data can be utilized to identify insertions that have contributed to past adaptation, with the known pitfall that by definition empirical distributions always contain extreme values in their tails [93]. But copies of a given TE family are spread in the genome (and most likely not in linkage), and therefore statistics relying on an expected ‘spatial’ pattern of polymorphisms as a function of linkage to the selected allele (where multiple contiguous markers deviate from the neutral expectation) are not applicable. Perhaps, the only summary statistic that could be used because it does not necessitate information at multiple loci but rather information across populations is the population-average Fst. Fst estimates the sample variance in allele frequencies among populations and therefore their degree of differentiation and may be used to
provide evidence for selected insertions [94]. The idea behind the use of estimates of population differentiation is that recent adaptations in populations often reflect the peculiarities of local environments. But these \( F_{st} \) measures suffer from several known drawbacks: they are highly variable [95], the detection power depends on the polymorphism frequency at the onset of positive selection [96] and non-independence between populations increases the false positive rate [97].

Finally, correlation of population allele frequencies with environmental variables, e.g. along an environmental cline, offers a valuable tool to investigate how crops have adapted to particular selection pressures (e.g. biotic or abiotic). Correlation methods offer more power than differentiation-based methods but suffer from a higher rate of false positive [98]. Providing a detailed environmental characterization [99], genotype–environment correlation methods allow detection of candidate polymorphisms that have been selected for and contribute to variation in allele frequencies among populations. It also gives insights into the nature of the selective biotic or abiotic ‘agent’ involved. Along this line, Gonzalez et al. [100] have used a regression model to test the association between the frequency of TE candidate insertions, and geographic/climatic variables (latitude, maximum and minimum temperature, rainfall) in four Australian populations of \( D. \) melanogaster. Their results revealed a proportion of significant associations greater than that expected at 5% false discovery rate, suggesting the contribution of TE insertions to climate adaptation. However, treating neighboring populations as independent observations is problematic because populations may share a recent common population history. Coop et al. [101] have proposed a Bayesian approach to overcome these difficulties. They provided a null model in which derived population allele frequencies evolve from an ancestral population allele frequency, and estimated the covariance of allele frequencies across derived populations using a Monte Carlo Markov chain. This covariance matrix is used to correct for relatedness among populations. The null model is compared with the alternative model that encompasses a linear relationship between allele frequencies and an environmental variable. Loci in the tail of the resulting Bayes factors’ empirical distribution, i.e. outlier loci, are considered as candidates for adaptation. More recent developments include the estimation of genotype–environment correlations and random effects linked to population history simultaneously and hence without \textit{a priori} on a ‘neutral’ set of markers [102].

Finally, association mapping studies directly evaluate the link between genotypes and phenotypes measured in the field in large panels of distantly related individuals [103]. It therefore offers a valuable tool to characterize adaptive variation segregating within species—including polymorphisms generated by TEs. It has been applied to crops with great success in mapping genes involved in flowering time [51], photoperiod response [104] and aluminium tolerance in maize [105]; in flowering time, spike length and stem diameter in pearl millet [106]; in grain color, size and quality in rice [107]; in resistance to late blight in potato [108] and in sclerotinia head rot in sunflower [109]. Just as for \( F_{st} \) scans or associations with environmental variables, however, the consistency of a pattern/association among markers in tight linkage is usually taken as evidence for true positive as opposed to false positive and facilitates statistical detection. With the lack of linkage among TE insertions, such evidence will remain difficult to obtain except if information at other markers surrounding the insertion is available. In addition, the frequency of TE variants is expected to be lower than other types of variants, stressing the need for screening larger dataset [110] and for adjusting statistical association mapping methods [111].

**FUNCTIONAL VALIDATION**

Coupling NGS detection with the use of population genetic/statistical tools allows for finding candidate TE insertions whether they stand as outliers in accordance with a past selective history or whether they are significantly associated to phenotypic variation. Among these candidate insertions, however, some are false positive. Hence, while these strategies represent a major first step toward finding TE adaptive insertions, further functional validation is necessary.

As exemplified by cases presented in Table 1, several methods can be applied. In many cases, fine mapping may be used for cloning the candidate TE insertion. While this allows characterizing the type of TE involved and gives first clues about the possible mechanism underlying the phenotypic change, it does not provide any direct evidence of causality, except in particular cases where TE insertion occurs within an exon and thus disrupts or highly
In such cases, characterization of the complete cDNA and/or protein remains necessary, as it ensures that the TE sequence is not spliced out. An intermediate level of validation, which is often conducted, resides in comparing the phenotype (at the whole plant level or at the gene expression level) of individuals that do carry or do not the candidate TE insertion. This is usually efficient although individuals often harbor, in addition to the insertion itself, other genomic differences making it hard to rule out possible effect of other cis or trans factors on the phenotype observed. If however the effect of the insertion is consistent across different genetic backgrounds then the most plausible explanation would be the causality of the TE insertion. Methylation studies can also be performed and used as an additional link between TE presence and gene expression change.

When possible, the best validation resides in analyzing the phenotypic effect of a series of transgenic plants or plants originating from culture of protoplasts injected with constructs that differ only by the presence/absence of the candidate TE. However, in most cases, transformation analyses rule out long distance cis effects as well as trans effects, but does not exclude possible phenotypic effect due to SNPs or other small indels located in the close vicinity of the candidate TE.

CONCLUSION

While TEs have long been considered as junk DNA, numerous examples of TE insertions generating new phenotypes of interest for human needs can be uncovered in crops. Technological advances have driven the discovery of TE-triggered phenotypic changes and the underlying mechanisms. Among them, the most recent studies have revealed a prominent role of TE insertions in gene regulatory regions at the origin of expression changes. Full characterization of the functional impact of specific TE insertions is costly, long and therefore rarely undertaken. NGS technologies can nevertheless provide a number of correlates between TE presence/absence and molecular variation (e.g. DNA methylation, transcription or translation) that will contribute to elucidate mechanisms through which TEs can impact phenotypes. The growing number of crop complete genome sequences will be a gold mine in which to dig up TE-mediated polymorphisms. Statistical developments also offer new opportunities to test the association of these TE polymorphisms with phenotype and/or environmental variation and will ultimately contribute to unravel TE contribution to crop evolution.

Key points

- There are several mechanisms through which TEs create phenotypic changes: gene disruption, regulation of gene expression, rearrangements, transduction and transposase exaptation. While gene disruption has been historically the easiest to detect, the flow of omics data will provide unprecedented material to estimate their relative contribution to crop evolution.
- The use of NGS data to detect TE-based polymorphisms is a challenging task. Four methods are currently available: sequencing of family specific TD reactions, mapping of paired reads on a reference genome, detecting split read signatures and detecting variations in DOC. Recent advances suggest combining several of these methods promises to lead to more reliable detection.
- Population genetics offer valuable tools to identify TE insertions that have potentially contributed to crop evolution. In theory, both selection coefficients for a given TE family and demographic parameters for the considered crop could be incorporated in coalescent-based models. Such models would be useful to detect signatures of past selection.
- While association mapping provides first hints about the impact of transposon insertions on phenotypic variation, true functional validation is necessary to establish a causal relationship. This critical step has revealed several striking examples of TE-driven changes in crops.
- Up to now, relatively few cases of TE-induced DNA methylation-based expression changes have been characterized. Deciphering the intimate interplay between TEs, DNA methylation and nearby gene expression will provide new insights into how TEs can shape crop phenotypic diversity.

Acknowledgements

The authors are grateful to Sylvain Fustier for help in arranging Figure 1 photographs, Aleksandra Walczak, Johann Joets and Yves Vigouroux for valuable inputs on specific sections of the manuscript. They thank Josep Casabuberta and Brandon Gaut for precious comments that helped improving the manuscript as well as two anonymous reviewers.

FUNDING

The PhD fellowship of M.A.F. is supported by the Agence Nationale de la Recherche, project grant ANR-12-ADAP-002 to M.I.T.

References


