Naturally occurring variation in gene-associated transposable elements impacts gene expression and phenotypic diversity in woodland strawberry

Ileana Tossolini[1,2,5,7], Rocio Tolley[1,6,7], Julia Llinares-Gómez[3], Camila Zlauvinen[3], Tuomas Toivainen[4], Timo Hytönen[4], David Posé[3], Carmen Martín-Pizarro[3*], Pablo A. Manavella[3*]

AFFILIATIONS

[1]Facultad de Ingeniería, Universidad Nacional de Entre Ríos, 3100 Oro Verde, Argentina.
[2]Instituto de Agrobiotecnología del Litoral (CONICET-UNL), 3000 Santa Fe, Argentina.
[3]Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" (IHSM "La Mayora"), Universidad de Málaga-Consejo Superior de Investigaciones Científicas (UMA-CSIC), 29010 Málaga, Spain.
[4]Department of Agricultural Sciences, Viikki Plant Science Centre, University of Helsinki, Finland.
[5]Current address: Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática (IBB, CONICET-UNER), 3100 Oro Verde, Argentina.
[6]Current address: Instituto Andino Patagónico de Tecnologias Biológicas y Geoambientales (IPATEC, CONICET-UNComa), 8400 Rio Negro, Argentina.
[7]These authors contributed equally to this work.

*To whom correspondence should be addressed. E-mail: pablomanavella@ihsm.uma-csic.es; cmmartin@uma.es

ABSTRACT

Transposable elements (TEs) constitute a major portion of plant genomes and play key roles in shaping genome architecture, regulating gene expression, and driving genome evolution. In this study, we generated a comprehensive and curated TE library for the woodland strawberry (*Fragaria vesca*) by integrating two bioinformatic pipelines (EDTA and DeepTE). Our annotation revealed that TEs account for approximately 37% of the *F. vesca* genome. Analysis of TE-derived inverted repeats (IRs) and miniature inverted-repeat transposable elements (MITEs) demonstrated their association with 24-nt small interfering RNA (siRNA) production and differential DNA methylation patterns across tissues, suggesting a role in the epigenetic regulation of gene expression, particularly during fruit ripening. This MITE-mediated epigenetic regulatory mechanism was confirmed by evaluating gene expression and chromatin organization at FvH4_7g18570, which encodes the alcohol acyl transferase (*FvAAT1*). Three MITEs located upstream or downstream of the *FvAAT1* coding sequence were shown to influence epigenetically this gene expression.

Furthermore, we analyzed 210 re-sequenced accessions from the *F. vesca* European germplasm collection to identify and annotate TE insertion and deletion polymorphisms. A principal component analysis (PCA) based on these polymorphisms revealed subpopulation structures that reflect geographic origins. A genome-wide association study (GWAS) uncovered significant associations between specific TE polymorphisms and economically important fruit traits, including aroma-related volatile compounds and fruit size. Among them, the insertion of a hAT MITE near FvH4_2g00610 correlated with increased levels of γ-decalactone, a desirable aroma compound in strawberries.

These findings underscore the functional significance of TE-derived elements as key contributors to phenotypic diversity through novel regulatory functions. By integrating TE polymorphisms into population-genomic and functional studies, this work provides valuable insights into strawberry fruit development and quality traits. It also highlights the potential of harnessing TE-mediated variation in breeding initiatives and genome editing strategies to improve fruit quality.

KEYWORDS
Transposons, Strawberries, DNA methylation, MITE, Inverted Repeats, Fragaria, Epigenetics

INTRODUCTION

Transposable elements (TEs) are repetitive, mobile DNA sequences widely dispersed across eukaryotic genomes. They constitute the most abundant genomic component in many species, accounting for nearly 50% of the human (*Homo sapiens*) genome [1] and up to 85% of the wheat (*Triticum aestivum*) genome [2]. Their extensive genomic presence directly impacts genome architecture, evolution, and gene regulation, particularly in response to environmental stressors [3-7].

TEs are classified into two major classes based on their transposition mechanisms [8]. Class I elements, or retrotransposons, transpose via an RNA intermediate through a “copy-and-paste” mechanism, whereas Class II elements, or DNA transposons, move via a “cut-and-paste” mechanism [9, 10]. TEs are further categorized into autonomous elements, which encode the proteins required for their own transposition, and non-autonomous elements, which rely on autonomous TEs for mobility [8].

Within these classes, TEs are subdivided into orders, superfamilies, and families based on structural and functional characteristics, such as replication strategies, sequence conservation, and enzymatic activity [8]. In plants, Class I Long terminal repeat (LTR) retrotransposons (e.g., Gypsy and Copia) are particularly abundant. Non-LTR retrotransposons (e.g., LINEs and SINEs), though more extensively studied in animals, also contribute significantly to plant genome evolution. Among DNA transposons, miniature inverted-repeat transposable elements (MITEs) and Helitrons are especially noteworthy. MITEs are short (<800 bp), non-autonomous Class II TEs that are highly abundant in eukaryotic genomes [7, 11, 12]. While many TEs accumulate in pericentromeric regions, MITEs and other short TE-derived inverted repeats (IRs) are frequently located near genes, where they can function as regulatory elements [13, 14]. A key feature that makes MITEs potent regulatory elements is their ability to form hairpin-shaped double-stranded RNAs (dsRNAs) upon transcription. These dsRNAs are processed by DCL3 into 24-nt small interfering RNAs (siRNAs) that trigger DNA methylation through a variant of the RNA-directed DNA methylation (RdDM) pathway. This non-canonical form of RdDM plays a crucial role in silencing TEs to maintain genome stability [15-17]. Different from the canonical RdDM-initiated silencing of TEs through heterochromatin formation, RdDM-mediated DNA methylation of MITEs does not commonly act as a silencing mark but instead could impact the expression, both positively and negatively, of nearby genes [7].

TE research has expanded significantly with the increasing availability of genome sequences, enabling improved annotation, classification, and functional analysis of these elements. However, challenges remain, particularly in non-model species with incomplete TE libraries [18]. Various computational approaches have been developed for TE

annotation [19], including *de novo* prediction [20-22], motif-based searches [23, 24], and similarity-based methods [25, 26]. Combining multiple strategies has proven most effective in generating accurate and comprehensive TE annotations [18, 22, 27, 28].

One of the most compelling open questions in TE research is understanding how the large natural variation in TE content within a species contributes to phenotypic diversity. Thanks to advancements in long-read high-throughput sequencing technologies, the study of TE polymorphisms has gained attention due to their role in complex traits and their potential as genetic markers for plant research and breeding [29, 30]. Integrating TE polymorphisms with Genome-Wide Association Studies (GWAS) has provided valuable insights into the functional implications of TEs in various biological processes [31-34]. For instance, transposon insertion polymorphisms (TIPs) have been shown to improve predictions of agronomic traits beyond what is possible with SNP-based approaches alone [31, 35-37].

The woodland strawberry (*Fragaria vesca*) is a wild diploid species from the Rosaceae family that serves as a model for strawberry research. It is one of the four ancestral species contributing to the polyploidization of the cultivated octoploid *F.* × *ananassa*, the most widely consumed strawberry species. Among the four subgenomes of *F.* × *ananassa*, the *F. vesca* subgenome exhibits the lowest density of TEs near genes and encodes significantly more dominantly expressed homeologs than the other three subgenomes combined [38]. While previous studies have examined TE diversity in the *F. vesca* genome [39-41], their complete annotation, functional characterization, and linkage with phenotypic variation remain poorly explored.

In addition to its genomic significance, *F. vesca* fruits are distinguished by their unique volatilome, which contributes to their intense floral and fruity aroma [9, 42]. The species exhibits substantial phenotypic diversity influenced by both genetic variation and environmental factors. The contribution of TEs to this phenotypic diversity remains largely unexplored. Given its importance as a genetic resource, understanding TE-driven variation in *F. vesca* could provide valuable insights into fruit development and quality traits, with potential applications for breeding.

In this study, we systematically annotated TEs in the *F. vesca* reference genome, generating a curated TE library. We analyzed TE polymorphisms in 210 re-sequenced *F. vesca* European accessions [43], uncovering their contribution to genetic diversity and subpopulation structure. Furthermore, using GWAS, we identified significant associations between TE polymorphisms and economically important fruit traits, including volatile compound profiles and fruit size. As a proof of concept, we validated the influence of a MITE located near FvH4_7g18570, which encodes the alcohol acyl transferase *FvAAT1*, a putative ortholog of *SAAT* involved in volatile ester biosynthesis in *F.* × *ananassa* [44]. A combination of gene expression analysis and profiling of chromatin topology at the *FvAAT1* locus in fruits with induced or depleted MITE activity confirmed the influence of this TE on the regulatory dynamics at this locus. Specifically, our results showed that DNA methylation triggered by MITE-derived siRNAs modifies chromatin organization, which in turn defines *FvAAT1* expression profile during each fruit ripening stage.

By integrating TE-driven variation into population-genomic and functional studies, this research provides new insights into the regulatory roles of TEs in strawberry biology and their potential applications in crop improvement. The identification of TE-derived regulatory elements and their functional impact on key agronomic traits highlights the potential for leveraging TE variation in breeding strategies aimed at improving strawberry fruit quality.

RESULTS

Comprehensive identification and annotation of transposable elements in *Fragaria vesca*

Transposable elements (TEs) do not typically experience selective pressures that preserve their sequences over time. Consequently, they tend to become fragmented due to the accumulation of random mutations, which drives divergence among different TEs and complicates their identification [22, 45, 46]. To address these challenges, we combined two recent bioinformatic pipelines —EDTA [47] and DeepTE [48] (Figure S1)— to generate a semi-automated *de novo* TE library for *F. vesca*. The combination of both approaches allowed us to comprehensively identify and classify TEs in the *F. vesca* v4.0 reference genome [49, 50] into classes, orders, and superfamilies (Table 1, Figure 1B).

The complete set of identified TEs accounted for approximately 37% of the *F. vesca* genome. In plants, TE content is typically proportional to genome size [51, 52]. For example, the maize genome (~2,500 Mb) contains over 80% TEs and repetitive elements, whereas the smaller *Arabidopsis thaliana* genome (~135 Mb) contains only 21% TEs [53-55]. The proportion of TEs identified for *F. vesca* (~37%), which has a genome size of ~220 Mb [56], aligns with these observations (Figure 1A).

Unlike in most plant species, where Class I retrotransposons are dominant, Class II DNA transposons constitute the largest component of the *F. vesca* genome, accounting for ~20% of the total genomic DNA (Figure 1B). In contrast, LTR retrotransposons, which are typically the most abundant TE class in plants, comprise ~16% of the *F. vesca* genome, with Gypsy being the most abundant LTR superfamily (Figure 1B). Non-LTR retrotransposons, including SINEs and LINEs, contribute the smallest proportion of Class I elements (~0.6% of the genome). This unusually low proportion of Class I elements may be associated with the relatively small genome size of *F. vesca* and the commonly poor quality of centromeric genome information.

Interestingly, *Fragaria* species exhibit higher diversity among DNA transposons, compared to other plant genomes, highlighting the unique composition of their genomes [53]. Within Class II, DeepTE allows for the subclassification of MITEs and non-MITEs (DNA TEs). Approximately 18.6% of identified TEs belong to subclass 1, including both MITEs and other DNA transposons from the CACTA, Harbinger, TcMar, hAT, and Mutator superfamilies. The remaining 1.4% corresponds to subclass 2, specifically Helitrons (Figure 1B). Despite their known difficulty in identification, the proportion of Helitrons identified in the *F. vesca* genome is consistent with reports in other plant genomes, where they typically constitute 1–4% of the genome [39, 47, 57].

We then analyzed the size distribution of TEs for each class based on the annotation generated for *F. vesca*, including both fragmented and intact TEs (Figure 1C, Figure S2, Table S1). Overall, no significant differences were observed in size distributions among the results obtained from EDTA, DeepTE, and their combined pipeline (Figure 1C). The average length of each intact TE type is detailed in Figure S2 and Table S1.

A high proportion of fragmented TEs was identified, particularly among Class I and Class II transposons (Figure S2). This explains the relatively low mean length values observed for the complete TE library (Figure S2), which is consistent with the accumulation of mutations and deletions in ancient TEs. In contrast, MITEs, which are known for being relatively short

183 [58, 59], showed a higher proportion of intact elements. Within Class I, most TEs have lengths
184 around 9 kb (Figure 1D), with the longest elements belonging to the LTR order [60]. Class II
185 TEs, however, display a broader range, which is strongly linked to their respective superfamily
186 classification (Figure 1D, Figure S2). The distinct size patterns observed among TE classes
187 suggest differences in replication strategies and selective pressures acting on different TE
188 families.
189 Finally, we made ready-to-use annotation files available in GFF3 format in a public
190 repository (https://github.com/Ro-Toll/TE-Annotation-and-Polymorphisms_Auxiliary-Scripts).
191 These files include all annotated TEs, both fragmented and intact, with detailed descriptions
192 including TE ID and name, classification, annotation method, and TSD/TIR sequences when
193 applicable.
194
195 **Genome-wide distribution and abundance of Transposable Elements**

196 The distribution, abundance, and genome coverage of TEs vary substantially across
197 species and are strongly influenced by subfamily classification. While some TEs tend to
198 accumulate near protein-coding genes or toward the centromeres, others are more evenly
199 distributed across chromosomes. To assess the distribution of TE types in *F. vesca*, circos
200 ideograms were generated for each TE class, with gene density included as a reference
201 (Figure 2). Each concentric circle represents the coverage of a specific TE type based on
202 density across 500 kb windows.
203 As shown in Figure 2A, regions with higher Class I TE content tend to have lower gene
204 density, a pattern particularly evident for chromosomes Fvb1, Fvb3, and Fvb4. This TE
205 distribution is predominantly driven by LTR transposons, which are the most abundant. A
206 significant proportion of LTR elements is unlikely associated to genes as they appeared
207 located more than 3,000 bp from genes (Figure S3). This distribution pattern of Class I
208 elements has been previously reported in several plant species, including tomato (*Solanum*
209 *lycopersicum*) and rice (*Oryza sativa*) [54, 55, 61].
210 In contrast, Class II TEs are more concentrated near pericentromeric regions compared to
211 chromosome arms except for Helitrons, which exhibit a more widespread distribution across
212 the genome (Figure 2B). Helitrons positioning varies considerably between species. For
213 example, in potato and rice Helitrons show a more erratic pattern of distribution [55, 62]
214 compared to Arabidopsis, where they are enriched in gene-poor pericentromeric regions, or
215 in maize (*Zea mays*), where they are more abundant in gene-rich regions [63].
216 MITEs exhibit a distinct distribution pattern characterized by a notable concentration in
217 gene-rich regions, particularly in the immediate vicinity of, or within, coding sequences (Figure
218 S3). This pattern aligns with previous reports describing the regulatory role of MITEs in gene
219 expression [54, 55, 64]. Notably, 37% of *F. vesca* genes harbor a MITE within 3,000 bp of their
220 coding sequence, and 5% of these genes contain MITEs within the coding regions. This
221 proximity to genes reinforces the potential role of MITEs in influencing gene expression
222 through epigenetic and chromatin-level mechanisms encouraging us focus our study on these
223 elements.
224
225 **Characterization of MITEs and short TE-derived IRs in *F. vesca***

226 MITEs located near genes have been shown to influence gene expression by modifying
227 local chromatin topology through the production of small RNAs that trigger localized DNA

methylation [65-68]. A similar effect can be triggered by gene-associated TE-derived short, inverted repeats (IRs) elements, which often originate from MITEs that have lost distinctive structural elements. The presence of IRs reflects the progressive fragmentation and degeneration of TEs over evolutionary time [7]. In any case, and for the purpose of the regulatory effects studied here, IRs and MITEs operate similarly. The matching distribution patterns of MITEs and IRs in the *F. vesca* genome motivated us to broaden our analysis to include TE-derived IRs. Like MITEs, short IRs are frequently located near genes, where they influence gene expression profiles and act as evolutionary agents contributing to adaptation and developmental processes [67, 68]. The expression of these elements has been linked to the production of 24-nt small RNAs through a non-canonical RdDM pathway, triggering DNA methylation and inducing changes in local chromatin organization. These chromatin changes can result in the formation of short-range chromatin interactions that affect the transcriptional activity of neighboring genes [65-67].

To identify all IRs neighboring protein-coding genes in the *F.vesca* v4.0.a2 reference genome, we used *einverted* from the European Molecular Biology Open Software Suite (EMBOSS) [69]. This analysis identified a total of 9,472 IRs, representing more than ten times the number of IRs characterized in the *Arabidopsis thaliana* genome [67]. Most of the identified IRs (7,476, ~80%) were found near or within annotated protein-coding regions (Figure 3A). Of these, 1,855 IRs are situated 500 bp from a gene, 1,217 between 501 bp and 1,000 bp, and 2,508 between 1,001 bp and 3,000 bp (Figure 3A). In total, 23% of *F. vesca* genes harbor an IR within these ranges. Similarly, of the 15,040 MITEs annotated (including fragmented elements), 12,246 MITEs were located near or within annotated protein-coding regions. Specifically, 3,702 MITEs were located within 500 bp of a gene, 2,166 between 501 bp and 1,000 bp, and 3,933 between 1,001 and 3,000 bp (Figure 3A). In total, 44% of *F. vesca* genes harbor at least one MITE or IR within 3,000 bp of their coding regions. Notably, 1,896 IRs (~20%) and 2,445 MITEs (~16%) are fully located within protein-coding genes, primarily in introns (Figure 3A-B), representing approximately 5% of all coding genes in the *F. vesca genome*. TE insertions in introns can have a significant regulatory effect. For example, in maize, insertion of a CACTA TE derived element into the *Par1* gene creates an epiallele whose methylation status directly affects gene expression and plant architecture [70].

The detected IRs ranged in size from 100 bp to 700 bp (the length limit of our analysis), with an average length of 427 bp. The annotated MITEs exhibited a broader size distribution, ranging from 51 bp to 2,120 bp, with an average length of 285 bp. An overrepresentation of elements between 100 bp and 500 bp (85%) was observed, with only three elements exceeding 1,500 bp (Figure 1D). These values align with the typical size range of MITEs in plants [71, 72]. However, the inclusion of fragmented transposable elements during the TE annotation process may have influenced the observed maximum size and average length.

We then classified the identified IRs and compared them with the superfamilies previously assigned to MITEs (Figure 3C). IRs can originate from both MITE-type transposons and DNA transposons, as well as from Class I elements or even non-TE sequences [67]. In line with this, 16% of the IRs were classified as elements distinct from MITEs ("Other" category in Figure 3C), predominantly belonging to DNA Mutator/CACTA, LTR Gypsy/Copia, and unknown origin (Figure S4). Conversely, 84% of the identified were classified as MITE-derived elements. The overall superfamily distribution of IRs was consistent with that of the annotated MITEs in our TE library (Figure 3C). However, the proportion of IRs from the MITE CACTA/hAT superfamilies was slightly lower than that of the annotated MITEs. This suggests that certain

IRs may have originated from more ancient MITE elements that have since diverged from their original sequence patterns.

Most MITEs and IRs in *F. vesca* are epigenetically active

As mentioned before, upon transcription, MITEs and IRs can be processed into siRNA that trigger non-canonical RdDM, leading to the regulation of neighboring genes and turning these TEs into autonomous regulatory elements. To identify potentially active regulatory MITEs and IRs, we used sRNA and bisulfite (BS) sequencing to determine whether the identified MITEs and IRs produced 24-nt siRNAs and triggered DNA methylation—clear indicators of active and potentially regulatory TEs. We analyzed three development stages: leaves, immature fruit receptacles (white stage) and ripe fruit receptacles (red stage) (Figure 3D, Figure S5). Our analysis revealed that 90.3% of MITEs and 91.6% of IRs produced siRNAs in at least one of the tested tissues (Table S2). Notably, approximately 80% of these elements are located within 3 kb of a gene, which is consistent with previous reports showing that gene-hosted MITEs and IRs are commonly transcribed by RNA polymerase II (Pol II) and are epigenetically active [67]. Metagene analysis of 24-nt siRNAs mapping to the MITEs and IRs showed a characteristic two-peak profile, indicating that the siRNAs are produced from the terminal inverted repeats (TIR) through DLC3 processing (Figure 3D). This behavior was particularly pronounced for IRs located within 1,000 bp of genes. In contrast, for MITEs, this pattern persisted even for elements located farther away from genes, although their expression levels decreased with increasing distance (Figure S5A).

BS-sequencing metagene analysis revealed that annotated TEs, MITEs, and IRs are methylated under all tested conditions (Figure 3E, Figure S5B). Gene-associated MITEs and IRs, particularly those closest to genes, showed a signature two-peak CHH methylation pattern, consistent with the siRNA production profile and RdDM-dependent methylation. Moreover, siRNA levels and CHH methylation decreased as the distance between MITEs or IRs and the associated gene increased (Figure S5B). This distance-dependent epigenetic activity of MITEs and IRs is consistent with previous reports showing greater transcriptional and regulatory activity of elements closer to a gene [67]. This observation supports the idea that 24-nt siRNAs derived from MITEs and IRs near protein-coding genes initiate a non-canonical RdDM pathway likely dependent on Pol II transcription.

In contrast, methylation in the CpG and CHG contexts remained relatively stable regardless of the distance between TEs, MITEs, or IRs and nearby genes (Figure S5B). When these elements were located farther from genes, the surrounding genomic context showed higher CpG and CHG methylation levels, likely reflecting the heterochromatic nature of these regions. Furthermore, CpG and CHG methylation levels were higher in leaves compared to ripe fruits, contrasting with the siRNA production profile and consistent with their independence from siRNA activity (Figure 3D-E, Figure S5B). This contrasts with immature and ripe fruits, which showed higher CHH methylation, suggesting that RdDM activity is required for dynamic regulation of TEs during fruit development and maturation. These differences underscore the distinct epigenetic regulation of transposons across tissues, with stable maintenance of CpG and CHG methylation in somatic tissues and active RdDM-driven CHH methylation in reproductive tissues. This tissue-specific regulatory pattern also reflects the dynamic role of TEs in gene regulation and genome plasticity during fruit ripening and development.

MITEs and short IRs regulate chromatin organization and expression of adjacent genes

To investigate the potential regulatory roles of gene-hosted MITEs and IRs in gene expression, we performed a differential accumulation analysis of 24-nt siRNAs derived from these elements across different tissues of *F. vesca* as a proxy for their differential expression and processing. We compared siRNA profiles between leaves and fruits at two ripening stages (immature and ripe). The analysis revealed a distinct production of 24-nt siRNAs derived from MITEs and IRs across these tissues, considering thresholds of |log2FC| ≥1 and adjusted p-value ≤0.05 (Figure 4A). For MITEs, we identified 19 upregulated and 13 downregulated elements in ripe fruits compared to immature fruits, 276 upregulated and 64 downregulated in ripe fruits compared to leaves, and 77 upregulated and 268 downregulated in leaves compared to immature fruits. Similarly, for IRs, 18 elements were upregulated and 10 downregulated in ripe fruits compared to immature fruits, 297 upregulated and 78 downregulated in ripe fruits compared to leaves, and 59 upregulated and 244 downregulated in leaves compared to immature fruits. This approach allowed us to identify MITEs and IRs actively producing siRNAs in a specific sample, which may serve as potential regulators of gene expression in a tissue-specific manner.

We then identified genes located within 3,000 bp of the differentially expressed (DE) MITEs and IRs and analyzed their differential expression in the same tissues to evaluate a potential regulatory effect of these TEs (Table S3, Table S4). More than 30% of DE MITEs or IRs were located near genes that also exhibited significantly altered expression under the same conditions (Figure S6A), with this proportion increasing to over 50% for MITEs during the ripening stage. MITEs and IRs can act as enhancers or silencers of gene expression by generating cis-regulatory elements [73, 74] or, by modifying the local chromatin organization, altering the gene product without changing overall expression levels (e.g., by triggering alternative splicing or termination) [65]. Their activity depends on their position relative to the gene, their interaction with regulatory elements such as transcription factors or epigenetic readers, and the way they reshape chromatin architecture, all of which can vary between conditions.

Given the established role of 24-nt siRNAs in mediating RdDM and chromatin modifications, this combined analysis of siRNA production, methylation profiles, and gene expression provides a comprehensive view of how TE-derived elements contribute to tissue-specific gene regulation and developmental processes in *F. vesca*. To test this idea, we focused on fruit ripening as a proof-of-concept scenario and analyzed publicly available mRNA-seq data from strawberries treated with 5-azacytidine (5-AZA) (Figure S6B, Table S5). 5-AZA inhibits DNA methylation allowing us to distinguish between genes regulated by MITEs or IRs and cases where differential siRNA accumulation and gene expression are merely correlations. Our analysis revealed that over 40% of differentially expressed MITEs or IRs are near genes that also exhibit significant expression changes following 5-AZA treatment (Figure S6B). This proportion increases to over 50% during ripening. These findings highlight the potential of inhibiting DNA methylation to study key regulatory interactions between transposons and genes during strawberry development.

Among the 32 up- or downregulated MITEs and 28 up- or downregulated IRs identified during fruit ripening (Figure 4A, Figure S6C), we identified a 416 bp IR element denoted as *IR_9195*, which corresponds to a MITE Mutator. At the ripe fruit stage, this element showed a 12-fold higher siRNA production than at the immature stage (Table S4) (Figure 4B-C). This MITE is located 1,915 bp upstream of the gene *FvAAT1* (FvH4_7g18570), which encodes a putative ortholog to alcohol acyl transferase SAAT—a key enzyme responsible for the

366 biosynthesis of aroma volatile esters in *F. × ananassa* [44]. Consistent with its functional role,
367 *FvAAT1* displayed substantial upregulation during fruit ripening; at the ripe stage, gene
368 expression was more than 1700 times higher than at the immature stage (Table S5), aligning
369 with previous findings that underscore its importance in producing characteristic ester-based
370 aroma compounds during ripening [75, 76]. This gene is also upregulated during the turning
371 stage in *F. × ananassa* compared to *F. vesca*, emphasizing its importance in commercial fruit
372 quality traits [77]. By contrast, treatment with 5-AZA resulted in a fivefold reduction in *FvAAT1*
373 expression (Table S5), suggesting that the identified epigenetically active MITE can influence
374 its expression.
375 In the same genomic region, we identified another element (*IR_9194*), which also
376 correspond to a Mutator MITE, located 4,761 bp upstream of *FvAAT1* and 3 bp downstream
377 of FvH4_7g18550, a gene encoding a papain family cysteine protease implicated in protein
378 degradation and defense responses during fruit ripening (Figure 4B). This gene showed a
379 marked increase in expression in ripe fruit compared to immature fruit, with levels
380 approximately 105 times higher (Table S5), mirroring patterns observed in other plant species
381 in which cysteine proteases have been proposed to regulate ripening processes [78-80]. A
382 third element (*IR_9196*), also a MITE, was found located 1,064 bp downstream of *FvAAT1*
383 (Figure 4B).
384 We then evaluated siRNA production and CHH methylation patterns around these
385 elements to gain further insights into potential epigenetic regulation (Figure 4C). We observed
386 that *IR_9194* and especially *IR_9195*—but not *IR_9196*—produce 24-nt siRNA with higher
387 levels in ripening fruits, consistent with active IRs (Figure 4C). This increase in siRNA
388 production in mature fruits coincides with an increase in CHH methylation (Figure 4C).
389 Interestingly, although we did not observe a significant accumulation of siRNA at *IR_9196,*
390 this region showed the highest methylation levels (Figure 4C). This may indicate that this IR
391 is permanently silenced, in contrast to *IR_9194 and IR_9195,* which remain epigenetically
392 active. Although *IR_9196* appears to be transcriptionally inactive, the increase in DNA
393 methylation at this element may serve as an "anchor point" in conjunction with the more
394 dynamic *IR_9195* element to trigger chromatin looping as observed in other systems [66]. To
395 confirm this hypothesis, we infiltrated *F. vesca* fruits with a construct either overexpressing the
396 viral silencing suppressor P19, which reduces siRNA activity and thus RdDM, or an additional
397 copy of *IR_9195* to increase the production of specific siRNAs in trans and thus induce
398 methylation at this MITE. We then extracted RNA from the resulting plants and measured the
399 expression of *FvAAT1* in these fruits and mock infiltrated controls. Confirming a regulatory
400 effect of the MITE over *FvAAT1,* we observed opposite expression of the gene when RdDM
401 at the MITE was induced or repressed (Figure 4D). Such induction or repression of the MITE
402 activity was achieved as previously reported [66, 68]. Induction was accomplished by
403 transiently overexpressing a copy of *IR_9195* in *F. vesca* receptacles, which increase siRNA
404 production and consequently enhance the MITE methylation. Repression was achieved by
405 transiently overexpressing either a fragment of *IR_9195* fused to GFP, which sequesters the
406 MITE-derived siRNAs, or the viral silencing suppressor P19, which nonspecifically sequester
407 siRNAs. Furthermore, using a chromatin conformation capture (3C) approach, we observed a
408 ripening stage dependent change in chromatin organization, with *IR_9195* as the most likely
409 anchor point (Figure 4E). In this sense, a chromatin loop encompassing the entire loci was
410 increasingly detected with the maturation of the fruits. This also suggest that this chromatin

411 loop has a positive effect on gene transcription, as previously reported for example during Pol
412 II reclining happening on gene looping events [65].
413 Given the apparent regulatory importance of *IR_9195*, we assessed its conservation in the
414 orthologous genomic regions of the commercial strawberry species *F. chiloensis* and *F.* ×
415 *ananassa*. In *F. chiloensis*, a copy of *IR_9195,* situated 1,674 bp upstream of the *FvAAT1*
416 homologu*e* was detected displaying 89.4% sequence identity (Figure S7A). A similar pattern
417 was observed in *F.* × *ananassa* (Camarosa genome assembly), where a copy of *IR_9195*
418 associated with *FaAAT1* exhibits 95.7% sequence identity (Figure S7A). Analysis of the
419 predicted secondary structures using RNAfold for the three elements revealed the formation
420 of a long dsRNA hairpin structure, suggesting that they are likely DCL3 substrates and will
421 function similarly to those observed in *F. vesca* (Figure S7B). These findings underscore the
422 evolutionary conservation of this MITE near key ripening-related genes across different
423 *Fragaria* species, potentially reflecting a conserved regulatory function in fruit quality traits.
424 We identified additional IR/MITE elements adjacent to genes with key roles in fruit
425 development and metabolism, suggesting potential regulatory contributions through small
426 RNA production, DNA methylation, and gene expression changes (Table S4 and S5). For
427 example, a 385 bp MITE Mutator element (*TE_00001870: TE_homo_120946*) exhibiting
428 approximately a 4.5-fold increase in expression in red fruit compared to leaf tissue (Table S4)
429 and pronounced CHH methylation in fruit was identified 1,972 bp upstream of the
430 SEPALLATA-like MADS-box gene FvH4_6g46420 (Figure S8A). This MITE overlaps with
431 *IR_8440* and is adjacent to a MITE Harbinger. Another IR (*IR_8439*) located downstream of
432 FvH4_6g46420 displays relatively constant expression across tissues, suggesting it could
433 serve as a potential “anchoring” region for 3D chromatin configurations. The FvH4_6g46420
434 gene itself is significantly upregulated in white fruit vs. leaf tissue, consistent with prior studies
435 underscoring the role of its ortholog in *F.* × *ananassa*, *FaMADS9*, which is involved in
436 strawberry fruit development and ripening [81]. Silencing of *FaMADS9* in *F.* × *ananassa* delays
437 or inhibits normal ripening and alters both auxin and ABA metabolism, implicating this MADS-
438 box transcription factor in pleiotropic regulatory networks [82, 83]. Furthermore, *FaMADS9* is
439 downregulated under AZA treatment, highlighting the broader interplay between epigenetic
440 modifications and fruit ripening pathways [84].
441 Additional cases reinforce the potential role of IRs and MITEs in regulating nearby gene
442 expression. *IR_2155* (377 bp) is located near FvH4_2g30230, annotated as molybdate
443 transporter 1 (Figure S8B). Molybdenum is an essential trace element in plant nutrition and
444 physiology, and its effects on strawberry flavor and fruit quality are well documented [85]. The
445 upregulation of this gene in ripe fruit and its downregulation in 5-AZA-treated strawberries,
446 along with the distinct expression profiles of the IR in leaf, white, and red tissues, together with
447 localized methylation and siRNA mapping, further support a role for IR-associated epigenetic
448 control. Moreover, *IR_6046* (448 bp) is located upstream of FvH4_5g23180 (Figure S8C),
449 encoding a probable xyloglucan endotransglucosylase/hydrolase (XTH30), which is critical for
450 cell wall remodeling during fruit softening [86, 87]. Differential expression of both the element
451 and XTH30, alongside distinct siRNA clusters in leaf versus fruit tissues, aligns with a
452 mechanism involving dynamic chromatin organization. Fruit softening is a critical factor
453 affecting the marketability and economic value of fruits; therefore, targeting cell wall-related
454 genes, including XTHs, may offer an effective strategy to delay this process and extend
455 postharvest shelf life.

Taken together, these observations underscore the potential of IR and MITE elements to modulate gene activity through epigenetic pathways, thereby influencing fruit maturation and quality traits. Further research, including targeted methylation profiling and chromatin conformation assays, is needed to clarify whether the mechanisms underlying TE-mediated gene regulation in strawberry fruit development mimics those described in Arabidopsis. These insights also hold promise for breeding strategies aimed at improving key quality traits, such as flavor and texture by targeting non-coding elements, like MITEs and IRs.

Discovery of TE Polymorphisms in resequenced *F. vesca* genomes

Due to their inherent mobility, TEs frequently vary between natural populations in copy number and genomic position, producing insertion or deletion polymorphisms [16]. A TE insertion polymorphism (TIP) occurs when a TE is present in an accession but absent from the reference genome, whereas a TE absence polymorphism (TAP) represents a TE found in the reference but missing in other accessions. Because many TEs, particularly MITEs, harbor their own regulatory sequences, they can strongly influence the expression of nearby genes, often more than other types of structural variation [36, 88]. Historically, genetic analyses have primarily focused on single nucleotide polymorphisms (SNPs); however, TE insertion/deletion polymorphisms remain an underexplored source of genomic variation. As such, they hold the potential to affect phenotypes and key biotechnological traits in nutritionally and commercially important plant species.

We analyzed 210 resequenced *F. vesca* genomes from a European germplasm collection, representing the species' continental diversity ([43] PRJNA1018297), to study natural variation in TEs, MITEs, and IRs using TEPID [89] (Figure S9). Table 2 shows the number of polymorphisms identified for each element annotation. A relatively small number of polymorphisms was detected in IR elements, particularly for insertions, consistent with the absence of autonomous mobilization mechanism for these elements [64]. Over 80% of all insertion polymorphisms exhibited a minor allele frequency (MAF) below 3%, suggesting that these variants may be recent events not yet widely propagated or may have been subjected to negative selection. All resequenced accessions were treated as homozygous for identified polymorphisms, and those detected in six or fewer accessions (out of 210) were defined as rare (MAF < 3%).

Among the 458 identified TE insertions, 34% belonged to Class I and 66% to Class II, closely matching the general TE distribution in the reference genome (37% vs 67%). In contrast, the 1,706 deletion polymorphisms showed a more balanced Class I/Class II ratio (854 vs. 852). Insertion polymorphisms were particularly enriched for LTR Copia, MITE TcMar, and MITE hAT elements, suggesting recent bursts of activity that have not yet spread across the population [90] (Figure 5A). LTR Copia and MITE TcMar insertions were especially frequent among low-frequency variants (MAF < 3%). While most TE polymorphisms occur at low frequency [90], deletion polymorphisms were more common (60.8%) at higher frequencies (MAF ≥ 3%), especially in gene-poor regions, suggesting a neutral effect that may facilitate their persistence in the population. A principal component analysis (PCA) based on the identified TE polymorphisms revealed a clear subpopulation structure (Figure 5B). Accessions from the Baltic Sea region formed a well-defined cluster, separated from those of Great Britain, Iceland, and the Iberian Peninsula. A similar separation between Icelandic and Baltic clusters was previously reported in a volatilome-based PCA analysis of a European *F. vesca* collection by Urrutia et al [91]. Furthermore, while the volatilome study showed a distinct separation

502 between the Iberian and Scandinavian peninsulas, our analysis placed most Scandinavian
503 accessions as genetically distinct from Iberian ones but with some degree of overlap. This can
504 be explained by recent findings that Scandinavian peninsula was colonized by both western
505 and eastern *F. vesca* ancestry after the last glacial maximum, which resulted in admixed and
506 genetically isolated subpopulations in different parts of Scandinavia [43].
507 The distribution of polymorphisms for all annotated elements showed a preference for gene
508 rich regions, where chromatin is typically less condensed, likely facilitating access for the TE
509 mobilization machinery (Figure 5C). Similar patterns have been observed in other species,
510 where TE insertions tend to occur in gene-rich regions, although variations can occur at both
511 the TE family and species levels [36, 90, 92]. Deletions of Class II TEs, particularly MITEs,
512 were more frequently observed among rare polymorphisms, possibly owing to their "cut-and-
513 paste" transposition mechanism and selective pressures in gene-rich regions. Class II
514 elements showed a preferential insertion into promoter regions, resulting in a higher proportion
515 of insertions within 1,000 bp upstream of genes, particularly for MITEs and IRs. Insertions also
516 frequently occurred within introns (Figure 5D), which can affect gene expression, as reported
517 in *Brassica rapa* and *Solanum lycopersicum* [36, 37]. In total, less than 20% of the TEs
518 annotated and about 5% of MITEs and IRs showed polymorphisms, with only 0.7% of IRs
519 exhibiting insertion polymorphisms. This limited fraction may reflect the small number of
520 elements retaining mobilization capacity, as well as potential methodological limitations
521 caused by the low sequencing coverage (<15×) in some accessions.
522
523 **Linking TE polymorphisms to phenotypic traits in woodland strawberry fruits**
524 Several phenotypic parameters, including fruit dimensions (length, width, volume) and the
525 concentrations of 99 secondary volatile metabolites (volatilome), were previously scored for a
526 collection of 130 wild *F. vesca* accessions across two independent harvest seasons (H16 and
527 H17) [91]. Since the genomes of these 130 accessions were included in our polymorphism
528 analysis, we performed a genome-wide association study (GWAS) to investigate the
529 relationship between TE, MITE, and IR polymorphisms and different strawberry fruit traits.
530 A significant association was detected between the insertion of a MITE and the relative
531 amount of γ-decalactone (Figure 6A). The insertion, located on chromosome Fvb2 and
532 classified as a MITE hAT element ('*TE_00001326_67_i1*'), was significantly associated with
533 higher γ-decalactone levels in both harvests (p-value < 0.05). γ-decalactone was one of the
534 most variable compounds in the *F. vesca* volatilome analysis by Urrutia et al. [91]. In *F.* ×
535 *ananassa*, γ-decalactone is the most abundant lactone [93], reaching maximum levels in fully
536 ripe fruits [93] and is positively correlated with consumer perception of favorable organoleptic
537 traits [91, 94]. A total of 11 accessions carried the insertion (of which nine have available
538 volatilome data [91]). The TIP is located 2,220 and 7,463 bp upstream of two genes,
539 FvH4_2g00610 and FvH4_2g00600, both annotated as endoglucanase 9-like. The proximity
540 of the MITE hAT insertion to the endoglucanase 9-like genes suggests a regulatory influence
541 on their expression, which could in turn affect γ-decalactone production. Endoglucanases are
542 involved in cell wall modification and have been linked to the metabolism and release of volatile
543 compounds in fruit [95, 96]. Changes in the structural integrity or cell wall remodeling
544 processes may facilitate or modulate the biosynthetic pathways responsible for lactone
545 production and aroma development.
546 Another significant association was identified between a MITE deletion polymorphism,
547 annotated as '*MITE_8692_d1*' and eight volatile compounds. Accessions carrying this deletion

548 (seven from Norway and one from Lithuania) showed significantly lower levels of all eight
549 compounds (p-value < 0.05; Figure 6B and Figure S10). The polymorphism is located on
550 chromosome Fvb7, approximately 2,412 and 2,297 bp upstream of the FvH4_7g10910 and
551 FvH4_7g10920 genes, respectively. The first gene is annotated as prenylated RAB acceptor
552 1.B5 (PRA1.B5), which encodes a protein involved in vesicular transport [97]. The second
553 gene encodes a member of the 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase
554 superfamily, known for its versatile functions in enzyme activation, metabolite biosynthesis,
555 and epigenetic regulation [98]. Enzymes of the 2OG-dependent dioxygenase family participate
556 in the synthesis of secondary metabolites and phytohormones, as well as in growth,
557 development, and stress responses [99]. The eight compounds found at lower levels in
558 carriers of this MITE deletion are derived from three major pathways: fatty acid degradation,
559 cuticular wax biosynthesis, and volatile ester biosynthesis—pathways that are among the most
560 important contributors to the woodland strawberry volatilome [91]. Because these volatiles
561 span multiple metabolic routes, the polymorphism may have downstream effects that influence
562 multiple pathways. For example, changes in FvH4_7g10910 expression could affect vesicular
563 transport of enzymes to their respective compartments, while changes in FvH4_7g10920
564 expression might modulate phytohormone and secondary metabolite levels. Such pleiotropic
565 effects on fruit metabolism could explain why compounds from different biosynthetic routes
566 are significantly altered in deletion-carrying accessions.
567 A significant association was also identified between the TE deletion polymorphism
568 '*TE_00001968_9_d1*' and reduced levels of (Z)-3-hexenyl acetate (Figure S11A), a compound
569 linked to negative organoleptic properties in strawberry [91, 94]. This association appeared
570 only in harvest H16 data but involved a large number of carrier accessions distributed across
571 diverse geographic regions (142 total, of which 99 were included in the Urrutia et al. volatilome
572 study [91]). The polymorphism, classified as an LTR Gypsy element, is located near three
573 genes (FvH4_3g14362, FvH4_3g14363, and FvH4_3g14364), none of which contain a
574 functional annotation. The broad geographic distribution of this deletion suggests that it is not
575 deleterious and may even confer a selective advantage.
576 Another association involved fruit volume and a deletion polymorphism of a CACTA-type
577 DNA TE located on chromosome Fvb2 (Figure S11B). The eight genomes carrying this
578 deletion (spanning a relatively wide geographic range) exhibited a significantly larger average
579 fruit volume (p-value < 0.05) compared to those without the deletion. The TE overlaps intron
580 1 and exon 2 of FvH4_2g05211 (without functional annotation) and is situated 2,970 bp
581 upstream of FvH4_2g05210, which is annotated as a MYB44-like transcription factor.
582 Members of the MYB44 family are involved in growth regulation and hormone signaling (e.g.,
583 ABA, JA, and auxin). In *F.* × *ananassa*, FaMYB44.2 regulates sucrose accumulation in fruits
584 [100]. Therefore, although FvH4_2g05210 has not been functionally studied, the presence of
585 this TE deletion could influence the expression or function of this gene, potentially contributing
586 to the increased fruit volume observed in the carrier accessions.
587 Collectively, these findings underscore the importance of having a fully annotated genome,
588 as comprehensive functional information is essential for elucidating how such polymorphisms
589 influence fruit quality and other agronomically valuable traits. It also positions these genes and
590 TEs as important targets for future studies. The observed associations also highlight the
591 potential of TE polymorphisms, often overlooked in population-genomic studies of crops, to
592 act as significant drivers of phenotypic diversity. While these findings suggest potential links
593 between TE variation and agronomically relevant traits, improvements in genome annotation

594 and further in-depth studies—potentially involving CRISPR/Cas9-mediated editing of these
595 TEs—are needed to validate these associations and fully harness TE polymorphisms in future
596 breeding and crop improvement strategies.
597
598 **DISCUSSION**

599 Our study provides a comprehensive analysis of TEs in *F. vesca*, highlighting their role in
600 genome composition, organization, and functional impact. By systematically annotating TEs
601 and analyzing their polymorphisms across 210 *F. vesca* accessions, we have gained valuable
602 insights into their evolutionary significance and their potential contributions to gene regulation
603 and phenotypic diversity.
604 Our TE annotation revealed that TEs occupy ~37% of the *F. vesca* genome, a proportion
605 consistent with expectations for plant species of similar genome size. However, unlike most
606 plant genomes where LTR retrotransposons are dominant [51] *F. vesca* exhibits a strikingly
607 high proportion of Class II DNA transposons, comprising ~20% of the genome, a feature also
608 observed in the pan-genome TE content analysis across the *Fragaria* genus [39]. The
609 predominance of DNA transposons, particularly MITEs, suggests that *F. vesca* has undergone
610 a distinct evolutionary trajectory in TE composition. The enrichment of LTR retrotransposons
611 in pericentromeric regions of plant genomes is well documented; thus, the relatively low
612 abundance of LTR retrotransposons in *F. vesca* could reflect incomplete assembly of
613 centromeric regions. However, because Class I elements are typically linked to genome
614 expansion due to their replicative nature, the relatively low abundance of LTR
615 retrotransposons in *F. vesca* and its small genome size may indicate the presence of effective
616 silencing mechanisms limiting their proliferation. Like other plant genomes, the distribution
617 patterns of TEs in *F. vesca*, particularly the enrichment of MITEs and IR elements near genes,
618 reinforce their potential role as regulatory sequences influencing gene expression.
619 One of the most significant findings of our study is the widespread production of small RNAs
620 (siRNAs) from MITEs and IRs in *F. vesca*, supporting their potential regulatory role in gene
621 expression. Over 90% of MITEs and IRs produced 24-nt siRNAs in at least one tissue type
622 studied, with a characteristic two-peak siRNA profile indicative of processing through the
623 RdDM pathway. The strong correlation observed between siRNA production and CHH
624 methylation, particularly in reproductive tissues, underscores the functional relevance of these
625 elements in tissue-specific gene regulation. In contrast, CpG and CHG methylation remained
626 relatively stable, supporting the role of siRNA-driven CHH methylation as a key mechanism
627 controlling TE activity and gene expression in a dynamic and reversible manner. These
628 findings highlight the functional significance of MITEs and IRs as autonomous regulatory
629 elements that modulate gene expression through small RNA-mediated epigenetic
630 modifications. The presence of these regulatory TEs near genes involved in key biological
631 processes, including fruit development and metabolism, suggests that they function as
632 epigenetic modulators of gene expression.
633 Beyond their role in siRNA production, MITEs and IRs near genes in *F. vesca* appear to
634 influence gene expression through chromatin reorganization. Our analysis of differentially
635 expressed TEs and their associated genes revealed a correlation between TE and gene
636 expression levels, suggesting a more complex regulatory mechanism. Indeed, our analysis of
637 the *FvAAT1* locus indicates that MITEs and IRs act as chromatin organizers, facilitating or
638 repressing gene transcription through changes in chromatin topology.

A particularly compelling example is the regulatory role of a MITE located upstream of FvH4_7g18570, encoding the alcohol acyltransferase *FvAAT1*, a key enzyme responsible for volatile ester biosynthesis in strawberry [44, 101, 102]. The upregulation of *FvAAT1* during ripening, coupled with siRNA production from the adjacent MITE, supports a model in which this TE influences gene expression through methylation-dependent chromatin remodeling. Further validation using 5-AZA demethylation assays confirmed that *FaAAT1* expression is methylation-dependent, highlighting the direct regulatory effect of MITEs in *F. vesca* fruit ripening. The conservation of this particular MITE in *F. chiloensis* and *F.* × *ananassa*, with high sequence identity and structural similarities, suggests that it plays a conserved role in fruit aroma regulation across strawberry species. This discovery adds to the growing evidence that TEs contribute to the evolution of crop traits by introducing regulatory variation that affects key metabolic pathways [7, 36, 103, 104]. It also positions TEs, especially gene-associated MITEs and short IRs, as promising targets for genome editing. The CRISPR/Cas9-mediated removal of a MITE could significantly alter the expression of adjacent genes and the associated phenotypic traits. Genome editing of MITEs also offers the advantage that these elements are non-coding and naturally variable in populations, potentially simplifying regulatory approval.

In addition to their regulatory potential, TE polymorphisms provide an important source of genetic diversity that can influence trait variation [37]. Our GWAS linked TE polymorphisms to multiple fruit traits, including volatile composition and fruit size, demonstrating the phenotypic relevance of TEs in *F. vesca*. One of the strongest associations was found between a MITE hAT insertion on chromosome Fvb2 and increased levels of γ-decalactone, a key volatile compound in strawberry aroma [93, 105]. This TE is located near two *endoglucanase 9-like* genes, which are involved in cell wall remodeling and potentially linked to volatile biosynthesis. The presence of this MITE may enhance gene expression through its regulatory sequences or by altering chromatin accessibility, thereby contributing to increased γ-decalactone production in carrying accessions.

Similarly, a MITE deletion polymorphism on chromosome Fvb7 was associated with decreased levels of eight volatile compounds derived from fatty acid degradation, cuticular wax biosynthesis, and volatile ester pathways. This deletion occurs near genes encoding a prenylated RAB acceptor protein and a 2OG-dependent oxygenase, both of which play critical roles in metabolite biosynthesis and cellular transport [97-99]. The proximity of the deletion to these genes suggests that it may affect the expression or activity of both genes, leading to a coordinated reduction in volatile production. We also identified a significant association between a CACTA-type DNA TE deletion and increased fruit volume. This TE is located near a MYB44-like transcription factor, which may be involved in hormone signaling and sucrose accumulation [100]. Its removal could alter gene expression in a way that promotes increased fruit growth. Given the importance of fruit size in commercial strawberry production, understanding how TE polymorphisms modulate these traits could provide new targets for genetic improvement.

Our findings demonstrate that TE-driven variation significantly contributes to the natural diversity of fruit metabolic profiles, with potential implications for breeding strategies aimed at improving fruit aroma and flavor. Furthermore, understanding how TEs modulate gene expression and metabolite production at the molecular level provides valuable insights into the complex regulatory networks underlying crop traits. The high degree of TE polymorphism observed in *F. vesca*, combined with the functional relevance of associated traits, suggests that TEs represent an untapped source of genetic variation for crop improvement.

CONCLUSIONS

Our findings demonstrate that, as in other plants, TEs in strawberries are not merely passive genomic elements but rather key contributors to genome function, gene regulation, and phenotypic variation. The identification of TE-derived regulatory elements and polymorphisms associated with fruit traits provides a valuable resource for future breeding efforts aimed at improving strawberry quality.

In particular, the regulatory roles of MITEs and IRs in fruit ripening suggest that targeted epigenetic modifications could be leveraged to optimize fruit quality. The use of genome-editing tools such as CRISPR/Cas9 to precisely manipulate TE-derived regulatory regions may enable the fine-tuning of gene expression without introducing foreign DNA, making it a promising strategy for crop improvement.

Additionally, the conservation of key regulatory MITEs in commercial strawberry species suggests that similar TE-driven mechanisms may underlie important agronomic traits in cultivated varieties. Further comparative genomic studies across *Fragaria* species could help elucidate the broader evolutionary impact of TEs and their role in shaping fruit diversity.

METHODS

Transposable element identification and classification

The genome sequence of *Fragaria vesca* and annotation files (v4.0.a2) were downloaded from the Phytozome web portal at https://phytozome-next.jgi.doe.gov/ [106]. TEs were detected and preliminarily classified using the EDTA v2.1.0 [47], executed as follows: `EDTA.pl --genome $genome --exclude $annotation_bed --anno 1`. where `$genome` is the *F. vesca* reference genome, and `$annotation_bed` is the assembly annotation converted to BED format from the original GFF3 file using the *gff2bed* tool from BEDOPS v2.4.41 [107]. EDTA produces a filtered non-redundant TE library for annotation of structurally intact and fragmented elements. TEs are clasified into several non-overlapping categories, including LTR (referring to LTR retrotransposons), non-LTR (including SINEs and LINEs), TIR (referring to DNA transposons with TIRs, including MITEs), Helitron, and non-TE repeat sequences.

TEs identified by EDTA were reclassified with DeepTE [48] that uses convolutional neural networks and detectes domains inside TEs to correct false classification. DeepTE was run with the following command: `DeepTE.py -d /path/intermed_files -o /path/results -i $TE_lib -sp P -m_dir /path/Plants_model`. Where `$TE_lib` is the FASTA file with the TE sequences detected by EDTA, the parameter `-sp P` specifies a plant organism, and the `-m_dir` flag indicates the path to the training model. A custom bash script was then used to merge the GFF3 EDTA output files with the DeppTE classification (TE order and superfamily). For TEs categorized as “unknown” by DeepTE, the original EDTA classification was preserved, translating it into the classification used in this work (Figure 1B). Two GFF3 files were generated with the genomic coordinates and assigned TE class/superfamily. One includes all TEs (including fragments), while the other contains only structurally intact TEs, that is, those containing all domains corresponding to their superfamily.

TE-derived IRs were classified by grouping sequences based on their sequence identity, using the CD-HIT-EST module of CD-HIT v4.8.1 [108], as previously reported for *Arabidopsis thaliana* [64], running the following command: `cd-hit-est -i $fa -o $output -c 0.8`

730 `-A 0.8 -T 2 -n 5`. A minimum sequence identity threshold of 80% (`-c`), a minimum
731 alignment coverage of 80% (`-A`), and a word size of 5 (`-n`) were set. Next, DeepTE was run
732 as previously described for TE annotation, using the representative sequences from each
733 cluster to classify the IRs. When IRs overlapped with annotated TEs but were not initially
734 classified as Class II, they were reclassified based on the overlapping element's category
735 using a custom bash script.
736
737 **IR analysis**
738 Inverted repeats were identified in the *F. vesca* v4 genome using *einverted* from the EMBOSS
739 program suite v6.6.0.0 [69]. The following parameters were used: maximum repeat of 700 bp,
740 a gap penalty of 8, a minimum score threshold of 150, a match score of 3, and a mismatch
741 score of −4. A custom Perl script was used to parse the output and generate a GFF3 file with
742 the IR annotation.
743 To assess the conservation of *IR_9195*, the Tripal MegaSearch tool from the Genome
744 Database for Rosaceae (GDR) website (https://www.rosaceae.org/) was first used to identify
745 orthologous genes in the strawberry species *Fragaria chiloensis* (KIB CAS genome assembly)
746 and *Fragaria* × *ananassa* (Camarosa genome assembly). A 3000 bp sequence upstream of
747 each gene was retrieved from JBrowse (GDR) for the respective genomes and locally aligned
748 with the *F. vesca IR_9195* sequence using EMBOSS Water
749 (https://www.ebi.ac.uk/jdispatcher/psa/emboss_water), which employs the Smith-Waterman
750 algorithm for pairwise alignment.
751 The predicted secondary structures with minimum free energy for each IR were generated
752 using the RNAfold web server (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi)
753 with default parameters.
754
755 **Plant material and growth conditions**
756 Plants from the *F. vesca* accession 'Reine des Vallées' used in this study were grown and
757 maintained in a greenhouse under natural sunlight conditions (IHSM "La Mayora", Málaga,
758 Spain), using a mixture of universal substrate and river sand (3:1 ratio by volume). Fruits at
759 the immature (white) and ripe (red) stages were collected for small RNA-seq (sRNA-seq) and
760 RNA-seq transcriptome analyses.
761
762 **sRNA-seq**
763 RNA was extracted as previously described [109] from fruit receptacles at two developmental
764 stages, immature (white) and ripe (red). In detail, for each stage, two biological replicates were
765 prepared; each biological replicate was composed of 30 receptacles after the removal of the
766 achenes. RNA quality of the samples was measured on a Bioanalyzer 2100 (Agilent
767 Technologies, Santa Clara, CA). RNA purity was checked using the NanoPhotometer®
768 spectrophotometer (IMPLEN, CA, USA). RNA integrity was analyzed by lab-on-a-chip
769 electrophoresis using the RNA Nano 6000 Assay Kit with the Agilent Bioanalyzer 2100 system
770 (Agilent Technologies, CA, USA). Quantification of samples was performed with Qubit® RNA
771 Assay Kit in the Qubit® 3.0 Flurometer (Life Technologies, CA, USA). Subsequently, 1 μg of
772 total RNA per sample was used to generate small RNA libraries using the Truseq® Small RNA
773 Library Prep Kit according to manufacturer's instructions (Illumina Inc., San Diego, CA, USA).
774 Library pool was sequenced on the Illumina Nextseq550 instrument at the University of Málaga

775 (Spain), with 1 × 75 cycles according to the manufacturer's protocol. Publicly available sRNA-
776 Seq reads from *F. vesca* leaves were retrieved from GEO under accession number
777 GSM1513956 and analyzed together with immature and ripe reads.
778 The small RNA reads were first cut to remove 3' adapters using cutadapt v4.9 and their quality
779 checked using FastQC v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
780 and MultiQC v1.25.1 [110]. They were then mapped with ShortStack v4.1.1 [111] to *F. vesca*
781 v4 genome using the following parameters: `--locifile` (with MITE or IR annotation files), –
782 `-nohp` (prevents search for microRNAs) and `--make_bigwigs`. The DESeq2 Bioconductor
783 package ([112], R Core Team 2022) was used to normalize the total raw read counts from
784 ShortStack for differential expression analysis. MITEs/IRs with 10 or more 24-nucleotide reads
785 across their entire region in all samples were considered to have potential siRNA production.
786 After count normalization, the expression of IR/MITE-derived siRNA was compared between
787 leaf tissue vs. immature stage, ripe vs. immature stages, and ripe stage vs. leaf tissue, using
788 the negative binomial test. Differentially expressed MITEs/IRs were determined by adjusting
789 p-values with the Benjamini–Hochberg false discovery rate, and those with an absolute log2
790 fold change ≥ 1 and an adjusted p-value ≤ 0.05 were defined as differentially expressed
791 elements.
792
793 **RNA-seq**
794 Total RNA was isolated as described previously [102] using the same protocol applied to the
795 fruit receptacles used for sRNA-seq analysis. For the RNA-seq experiment, however, three
796 independent biological replicates were prepared per developmental stage. After validating the
797 quality and quantity of the total RNA using Bioanalyzer 2100 (Agilent Technologies Santa
798 Clara, CA, USA) and Qubit 2.0, libraries were prepared using Illumina's TruSeq RNA Sample
799 Prep Kit v2. Pools of libraries were sequenced by paired-end sequencing (100x2) in an
800 Illumina HiSeq 2500 sequencer by Sistemas Genómicos (Valencia, Spain). Also,
801 transcriptome data for leaf tissue (GEO: GSM3243286 and GSM3243287), fruits treated with
802 5-AZA and its control (GEO: GSE151534) produced by previous studies were used [84, 113].
803 The analysis started by quality trimming and filtering the raw reads with Trimmomatic v0.39
804 [114], the reads were then aligned to the *F. vesca* v4 reference genome using STAR v2.7.11b
805 [115] which was guided by the gene and exon annotation from *F. vesca* v4.0.a2. *Sort* and
806 *bamCoverage* tools from Samtools v1.21 [116] were then used to sort BAM files by genomic
807 coordinates and generate normalized coverage tracks for visualization and further analysis,
808 respectively. Read quality before and after trimming was analyzed with FastQC v0.12.1
809 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and, together with mapping
810 efficiency, they were summarized with MultiQC v1.25.1 [110]. Read counts on each gene were
811 then calculated with featureCounts v2.0.8 [117]. This pipeline was run with the aid of the
812 Snakemake workflow engine [118]. Differential expression analysis was performed by the
813 DESeq2 package [112], filtering out genes with counts below 10 in all samples. Genes were
814 then annotated using the information (BLAST, InterPro, GO term, GO accession) provided by
815 the GDR database (https://www.rosaceae.org/).
816
817 **BS-seq**
818 Publicly available Bisulfite-Seq data [113] from *F. vesca* leaf tissue, immature receptacle and
819 ripe receptacle were retrieved from GEO under accession numbers GSM3243379,
820 GSM3243380 and GSM3243381, respectively. The EpiDiverse WGBS v1.0 pipeline for

821 bisulfite read trimming, mapping and methylation calling was implemented
822 (https://github.com/EpiDiverse/wgbs), which was specifically designed for nonmodel plant
823 species [119]. For each sample, bedGraph files corresponding to each sequence context
824 (CpG, CHG, CHH) were obtained. Then *bedGraphToBigWig* v.2.10 from ENCONDE
825 (https://www.encodeproject.org/software/bedgraphtobigwig/) was used to generate coverage
826 tracks for visualization and plotting.
827
828 **3C assay and RT–qPCR**
829 The 3C assay was conducted following a published procedure [66]. Approximately 2 g of tissue
830 was collected per sample and crosslinked with 1% formaldehyde. Crosslinking was stopped
831 by adding glycine to a final concentration of 0.125 M, after which nuclei were isolated. To
832 detect FvAAT1 loops DNA was digested with XhoI (Thermo Fisher Scientific). DNA fragments
833 were ligated with 25 U of high concentrated T4 DNA ligase (Thermo Fisher Scientific) at 22
834 °C for 5 h in a 2-ml volume. Subsequent steps included reverse crosslinking, proteinase K
835 treatment (Qiagen) and DNA purification with the phenol–chloroform. To measure interaction
836 frequency, qPCR was performed, and results were analyzed with the 2−ΔΔCt formula using
837 ACT2 as a control. The final qPCR products were subjected to agarose gel electrophoresis
838 (2%) and bands corresponding to the expected size were purified and sequenced using the
839 Sanger method to confirm ligation products in the case of 3C experiments. For RT–qPCR, 500
840 ng of total RNA underwent RT using the RevertAid RT kit (Thermo Fisher Scientific) after
841 treatment with DNase (Thermo Fisher Scientific). Subsequently, qPCR was conducted using
842 SYBR green (Thermo Fisher Scientific Maxima SYBR Green qPCR Master Mix (2×)). Relative
843 expression levels were determined using the 2−ΔΔCt method with *FvCHP1* as the
844 housekeeping gene control.
845
846 **Identification of TE Polymorphisms**
847 A set of 210 resequenced woodland strawberry genomes from a European germplasm
848 collection (PRJNA1018297), representing the species' continental diversity, was used for the
849 polymorphism study [43]. Detailed information on each accession's geographic origin, read
850 depth, and chromosomal coverage can be found in Table S6. Since the study of TE
851 polymorphisms requires a coverage over 15X and no less than 10X [89], resequencing data
852 was analyzed using SAMtools *coverage* v1.17 [120]. The results were processed using
853 custom bash and R scripts. Next, the accessions were filtered based on a mean per-
854 chromosome coverage exceeding 10X. Consequently, seven of the 217 available accessions
855 (Alta, FIN55, NOR2, NOR-P1-10, RUS3, UK3, and UK9) were excluded from subsequent
856 analyses.
857 TEPID v0.10 [89] was employed for the detection of polymorphisms, using the non-redundant
858 TE library previously generated by EDTA and reclassified by DeepTE, and the IR annotation
859 generated by *einverted*. The three stages of the TEPID algorithm were executed using the
860 commands listed below:
861 Initial Mapping Stage:
862 `tepid-map -x $bowtie_index -y $yaha_index -p $threads -s $insert_size`
863 `-n $reseqid -1 $r1 -2 $r2 -z`
864 Where `$bowtie_index` and `$yaha_index` indicate the paths to the reference indices
865 generated by each program; `$insert_size` refers to the average insert size of the
866 sequencing library, set to 250 bp based on data obtained using the *CollectInsertSizeMetrics*

module from Picard v3.1.1 (https://broadinstitute.github.io/picard/); `$reseqid` represents the genome under analysis; `$r1` and `$r2` denote the forward and reverse reads, respectively; and the `-z` flag is used to handle compressed FASTQ files.

Discovery Stage:

```
tepid-discover -p $threads --name $reseqid --conc $conc_BAM --split
$split_BAM --te $TE_bed
```

Where `$reseqid` is the genome identifier; `$conc_BAM` and `$split_BAM` are the Bowtie2 [121] and YAHA [122] alignment files resulting from the mapping stage, respectively; and `$TE_bed` is the BED file with TE annotations in the format specified by TEPID.

Refinement Stage

```
tepid-refine -p $threads --insertions $insertions --deletions
$deletions --name $reseqid --conc $conc_BAM --split $split_BAM -a
$reseq_list --te $TE_bed
```

Where `$insertions` and `$deletions` contain the identified insertion and deletion polymorphisms for all samples in the population; `$reseqid` is the genome identifier to run the *refine* command on; `$conc_BAM` and `$split_BAM` are the Bowtie2 and YAHA alignment files, respectively; `$reseq_list` is the list with all the genome identifiers in the population; and `$TE_bed` is the TE annotation BED file.

The resulting BED files were processed using custom bash scripts and command-line tools from bedtools v2.26.0 [57].

Principal Component Analysis

PCA was performed using the R packages FactoMineR [123] and factoextra (https://cran.r-project.org/web/packages/factoextra/index.html). The input VCF files were generated by transforming the output files from TEPID using custom bash scripts.

Genome-Wide Association Study

GWAS was conducted on the polymorphisms present in 130 of the accessions and their associated fruit traits and volatile compound profiles, as reported in [91]. TASSEL v5.2.89 [124] was employed using the generalized linear model (GLM) with 100 permutations. Analyses were conducted separately for each polymorphism type (insertion, deletion) and for each element annotation (TE, MITE, IR). Polymorphisms were filtered to include only those with a minor allele frequency (MAF) of ≥3%, considering only genomes with available phenotypic data. These analyses were performed using custom bash scripts.

Data processing and visualization

To calculate the number of genes near each type of TE and differentially expressed MITEs/IRs, the tool *slop* and *intersect* from bedtools v2.26.0 [57] were used for distances of 500, 1000, and 3000 bp both upstream and downstream of each element. Additionally, *intersect* was used to count elements located within coding regions, utilizing different annotation files for each feature.

In order to evaluate the global methylation and siRNA level, metagene analysis was performed using deepTools v3.5.5 [125]. First, *computeMatrix* was executed to generate read abundance from all samples over gene, TE, MITE and IR regions. This matrix was then used to create, using *plotProfile*, a metagene profile from 2 kb upstream of the Transcription Start Site (TSS) to 2 kb downstream of the Transcription End Site (TES).

Moreover, circos ideograms were created using Circos software vv0.69-6 [126], displaying only the assembled chromosomes due to their higher density of gene and transposon annotations. Also, JBrowse2 desktop v2.18.0 [127] was used to visualize gene and MITE/IR annotations, along with RNA-seq and BS-seq data, at specific loci in the *F. vesca* v4 genome. Data obtained from the different analyses of TE annotation and from the sequencing experiments were further processed, statistically analyzed, and plotted in R using a diversity of packages, including dplyr (https://github.com/tidyverse/dplyr), tidyr (https://github.com/tidyverse/tidyr), scales (https://github.com/r-lib/scales), forcats (https://github.com/tidyverse/forcats), ggplot2, RColorBrewer (https://CRAN.R-project.org/package=RColorBrewer), rtracklayer [128], and ggman (https://github.com/drveera/ggman).
Statistical significance in the GWAS analysis was assessed using either a T-test or a Wilcoxon test, depending on the data distribution, as specified in each figure legend.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The datasets generated during the current study are available from European Nucleotide Archive (ENA) under the accession number PRJEB86847. Scripts and annotation files generated for the analysis of *F. vesca* transposable elements are available at: https://github.com/Ro-Toll/TE-Annotation-and-Polymorphisms_Auxiliary-Scripts.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by grants from the Spanish Ministry of Science and Innovation (PID2022-137037NB-I00 and CNS2023-145312 to P.A.M., and PID2021-123677OB-I00 to D.P.), from the European Research Council (ERC-2014-Stg 638134 to D.P), and from the Junta de Andalucía (“Proyecto QUAL21 012 IHSM, Consejería de Universidad, Investigación e Innovación, Junta de Andalucía”. UMA20-FEDERJA-093 and Postdoctoral program, POSTDOC_21_00893 to C.M.-P.).

Authors' contributions
I.T. and R.T. performed all the bioinformatics analysis. J.L.G. and C.Z performed all the validations of the MITE-triggered effects over *FvAAT1*. C.M.P. collected the samples and performed sRNA-sequencing and RNA-seq experiments. T.H., D.P.P. and T.T. provided genomic data of *F. vesca* accessions (bam files). I.T., D.P.P., C.M.P., and P.A.M. conceived the study. C.M.P. and P.A.M. supervised the work and secured project funding. I.T. and P.A.M. wrote the manuscript with the help of all authors.

Acknowledgements

We would like to acknowledge the Sequencing Unit from the University of Malaga for the support they have provided for the small RNA-seq analysis.

REFERENCES

1. Mills RE, Bennett EA, Iskow RC, Devine SE: **Which transposable elements are active in the human genome?** *Trends Genet* 2007, **23:**183-191.
2. International Wheat Genome Sequencing C: **Shifting the limits in wheat research and breeding using a fully annotated reference genome.** *Science* 2018, **361**.
3. Liu P, Cuerda-Gil D, Shahid S, Slotkin RK: **The Epigenetic Control of the Transposable Element Life Cycle in Plant Genomes and Beyond.** *Annu Rev Genet* 2022, **56:**63-87.
4. Hosaka A, Kakutani T: **Transposable elements, genome evolution and transgenerational epigenetic variation.** *Curr Opin Genet Dev* 2018, **49:**43-48.
5. Judd J, Feschotte C: **Transposons take remote control.** *Elife* 2018, **7**.
6. Hirsch CD, Springer NM: **Transposable element influences on gene expression in plants.** *Biochim Biophys Acta Gene Regul Mech* 2017, **1860:**157-165.
7. Tossolini I, Mencia R, Arce AL, Manavella PA: **The genome awakens: transposon-mediated gene regulation.** *Trends in Plant Science* 2025.
8. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al: **A unified classification system for eukaryotic transposable elements.** *Nat Rev Genet* 2007, **8:**973-982.
9. Ulrich D, Komes D, Olbricht K, Hoberg E: **Diversity of aroma patterns in wild and cultivated Fragaria accessions.** *Genetic Resources and Crop Evolution* 2007, **54:**1185-1196.
10. Kumar A, Bennetzen JL: **Plant retrotransposons.** *Annu Rev Genet* 1999, **33:**479-532.
11. Hu J, Zheng Y, Shang X: **MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes.** *BMC Med Genomics* 2018, **11:**101.
12. Nouroz F, Noreen S, Heslop-Harrison JS: **Evolutionary genomics of miniature inverted-repeat transposable elements (MITEs) in Brassica.** *Mol Genet Genomics* 2015, **290:**2297-2312.
13. Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P, Yang G: **Miniature inverted-repeat transposable elements: discovery, distribution, and activity.** *Genome* 2013, **56:**475-486.
14. Kornienko AE, Nizhynska V, Molla Morales A, Pisupati R, Nordborg M: **Population-level annotation of lncRNAs in Arabidopsis reveals extensive expression variation associated with transposable element-like silencing.** *Plant Cell* 2023, **36:**85-111.
15. Cuerda-Gil D, Slotkin RK: **Non-canonical RNA-directed DNA methylation.** *Nat Plants* 2016, **2:**16163.
16. Yang DL, Huang K, Deng D, Zeng Y, Wang Z, Zhang Y: **DNA-dependent RNA polymerases in plants.** *Plant Cell* 2023, **35:**3641-3661.
17. Vaucheret H, Voinnet O: **The plant siRNA landscape.** *Plant Cell* 2024, **36:**246-275.
18. Bell EA, Butler CL, Oliveira C, Marburger S, Yant L, Taylor MI: **Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines.** *Mol Ecol Resour* 2022, **22:**823-833.
19. Loreto ELS, Melo ES, Wallau GL, Gomes T: **The good, the bad and the ugly of transposable elements annotation tools.** *Genet Mol Biol* 2024, **46:**e20230138.
20. Shi J, Liang C: **Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection.** *Plant Physiol* 2019, **180:**1803-1815.
21. Guo R, Li YR, He S, Ou-Yang L, Sun Y, Zhu Z: **RepLong: de novo repeat identification using long read sequencing data.** *Bioinformatics* 2018, **34:**1099-1107.

1014 22. Storer JM, Hubley R, Rosen J, Smit AFA: **Methodologies for the De novo Discovery**
1015 **of Transposable Element Families.** *Genes (Basel)* 2022, **13**.
1016 23. Mao H, Wang H: **SINE_scan: an efficient tool to discover short interspersed**
1017 **nuclear elements (SINEs) in large-scale genomic datasets.** *Bioinformatics* 2017,
1018 **33:**743-745.
1019 24. Ou S, Jiang N: **LTR_retriever: A Highly Accurate and Sensitive Program for**
1020 **Identification of Long Terminal Repeat Retrotransposons.** *Plant Physiol* 2018,
1021 **176:**1410-1422.
1022 25. Kennedy RC, Unger MF, Christley S, Collins FH, Madey GR: **An automated**
1023 **homology-based approach for identifying transposable elements.** *BMC*
1024 *Bioinformatics* 2011, **12:**130.
1025 26. Huda A, Jordan IK: **Analysis of transposable element sequences using CENSOR**
1026 **and RepeatMasker.** *Methods Mol Biol* 2009, **537:**323-336.
1027 27. Baril T, Galbraith J, Hayward A: **Earl Grey: A Fully Automated User-Friendly**
1028 **Transposable Element Annotation and Analysis Pipeline.** *Mol Biol Evol* 2024, **41**.
1029 28. Riehl K, Riccio C, Miska EA, Hemberg M: **TransposonUltimate: software for**
1030 **transposon classification, annotation and detection.** *Nucleic Acids Res* 2022,
1031 **50:**e64.
1032 29. Castanera R, Morales-Díaz N, Gupta S, Purugganan M, Casacuberta JM:
1033 **Transposons are important contributors to gene expression variability under**
1034 **selection in rice populations.** *eLife* 2023, **12:**RP86324.
1035 30. Rohilla M, Mazumder A, Saha D, Pal T, Begam S, Mondal TK: **Genome-wide**
1036 **identification and development of miniature inverted-repeat transposable**
1037 **elements and intron length polymorphic markers in tea plant (Camellia sinensis).**
1038 *Sci Rep* 2022, **12:**16233.
1039 31. Yan H, Haak DC, Li S, Huang L, Bombarely A: **Exploring transposable element-**
1040 **based markers to identify allelic variations underlying agronomic traits in rice.**
1041 *Plant Commun* 2022, **3:**100270.
1042 32. Liang Y, Liu HJ, Yan J, Tian F: **Natural Variation in Crops: Realized Understanding,**
1043 **Continuing Promise.** *Annu Rev Plant Biol* 2021, **72:**357-385.
1044 33. Yin Z, Yang Q, Shen D, Liu J, Huang W, Dou D: **Online data resource for exploring**
1045 **transposon insertion polymorphisms in public soybean germplasm accessions.**
1046 *Plant Physiol* 2023, **193:**1036-1044.
1047 34. Akakpo R, Carpentier MC, Ie Hsing Y, Panaud O: **The impact of transposable**
1048 **elements on the structure, evolution and function of the rice genome.** *New Phytol*
1049 2020, **226:**44-49.
1050 35. Vourlaki IT, Castanera R, Ramos-Onsins SE, Casacuberta JM, Perez-Enciso M:
1051 **Transposable element polymorphisms improve prediction of complex**
1052 **agronomic traits in rice.** *Theor Appl Genet* 2022, **135:**3211-3222.
1053 36. Cai X, Lin R, Liang J, King GJ, Wu J, Wang X: **Transposable element insertion: a**
1054 **hidden major source of domesticated phenotypic variation in Brassica rapa.**
1055 *Plant Biotechnol J* 2022, **20:**1298-1310.
1056 37. Dominguez M, Dugas E, Benchouaia M, Leduque B, Jimenez-Gomez JM, Colot V,
1057 Quadrana L: **The impact of transposable elements on tomato diversity.** *Nat*
1058 *Commun* 2020, **11:**4058.
1059 38. Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD,
1060 Teresi SJ, Nelson ADL, Wai CM, et al: **Origin and evolution of the octoploid**
1061 **strawberry genome.** *Nat Genet* 2019, **51:**541-547.
1062 39. Lyu K, Xiao J, Lyu S, Liu R: **Comparative Analysis of Transposable Elements in**
1063 **Strawberry Genomes of Different Ploidy Levels.** *Int J Mol Sci* 2023, **24**.
1064 40. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal
1065 P, Mockaitis K, Liston A, Mane SP, et al: **The genome of woodland strawberry**
1066 **(Fragaria vesca).** *Nat Genet* 2011, **43:**109-116.

41. Yu Z, Li J, Wang H, Ping B, Li X, Liu Z, Guo B, Yu Q, Zou Y, Sun Y, et al: **Transposable elements in Rosaceae: insights into genome evolution, expression dynamics, and syntenic gene regulation.** *Hortic Res* 2024, **11:**uhae118.
42. Ulrich D, Olbricht K: **Diversity of metabolite patterns and sensory characters in wild and cultivated strawberries 1.** *Journal of Berry Research* 2014, **4:**11-17.
43. Toivainen T, Salonen JS, Kirshner J, Lembinen S, Kort HD, Lyyski A, Edger PP, Hilmarsson HS, Hallsson JH, Sargent DJ, et al: **Late Quaternary climatic impact on the woodland strawberry genome: a perennial herb's tale.** *bioRxiv* 2024**:**2024.2010.2009.617376.
44. Aharoni A, Keizer LC, Bouwmeester HJ, Sun Z, Alvarez-Huerta M, Verhoeven HA, Blaas J, van Houwelingen AM, De Vos RC, van der Voet H, et al: **Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays.** *Plant Cell* 2000, **12:**647-662.
45. Mhiri C, Borges F, Grandbastien MA: **Specificities and Dynamics of Transposable Elements in Land Plants.** *Biology (Basel)* 2022, **11**.
46. Goerner-Potvin P, Bourque G: **Computational tools to unmask transposable elements.** *Nat Rev Genet* 2018, **19:**688-704.
47. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al: **Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline.** *Genome Biol* 2019, **20:**275.
48. Yan H, Bombarely A, Li S: **DeepTE: a computational method for de novo classification of transposons with convolutional neural network.** *Bioinformatics* 2020, **36:**4269-4275.
49. Edger PP, VanBuren R, Colle M, Poorten TJ, Wai CM, Niederhuth CE, Alger EI, Ou S, Acharya CB, Wang J, et al: **Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity.** *Gigascience* 2018, **7:**1-7.
50. Li Y, Pi M, Gao Q, Liu Z, Kang C: **Updated annotation of the wild strawberry Fragaria vesca V4 genome.** *Hortic Res* 2019, **6:**61.
51. Wang D, Zheng Z, Li Y, Hu H, Wang Z, Du X, Zhang S, Zhu M, Dong L, Ren G, Yang Y: **Which factors contribute most to genome size variation within angiosperms?** *Ecol Evol* 2021, **11:**2660-2668.
52. Beric A, Mabry ME, Harkess AE, Brose J, Schranz ME, Conant GC, Edger PP, Meyers BC, Pires JC: **Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales).** *G3 (Bethesda)* 2021, **11**.
53. Buisine N, Quesneville H, Colot V: **Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets.** *Genomics* 2008, **91:**467-475.
54. Noshay JM, Marand AP, Anderson SN, Zhou P, Mejia Guerra MK, Lu Z, O'Connor CH, Crisp PA, Hirsch CN, Schmitz RJ, Springer NM: **Assessing the regulatory potential of transposable elements using chromatin accessibility profiles of maize transposons.** *Genetics* 2021, **217:**1-13.
55. Zavallo D, Crescente JM, Gantuz M, Leone M, Vanzetti LS, Masuelli RW, Asurmendi S: **Genomic re-assessment of the transposable element landscape of the potato genome.** *Plant Cell Rep* 2020, **39:**1161-1174.
56. Zhou Y, Xiong J, Shu Z, Dong C, Gu T, Sun P, He S, Jiang M, Xia Z, Xue J, et al: **The telomere-to-telomere genome of Fragaria vesca reveals the genomic evolution of Fragaria and the origin of cultivated octoploid strawberry.** *Hortic Res* 2023, **10:**uhad027.
57. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

1118 58. Macko-Podgorni A, Stelmach K, Kwolek K, Grzebelus D: **Stowaway miniature**
1119 **inverted repeat transposable elements are important agents driving recent**
1120 **genomic diversity in wild and cultivated carrot.** *Mob DNA* 2019, **10:**47.
1121 59. Huang T, Li Y, Wang W, Xu L, Li J, Qi Y: **Evolution of lmiRNAs and their targets**
1122 **from MITEs for rice adaptation.** *J Integr Plant Biol* 2022, **64:**2411-2424.
1123 60. Ramakrishnan M, Satish L, Sharma A, Kurungara Vinod K, Emamverdian A, Zhou M,
1124 Wei Q: **Transposable elements in plants: Recent advancements, tools and**
1125 **prospects.** *Plant Molecular Biology Reporter* 2022, **40:**628-645.
1126 61. Jouffroy O, Saha S, Mueller L, Quesneville H, Maumus F: **Comprehensive**
1127 **repeatome annotation reveals strong potential impact of repetitive elements on**
1128 **tomato ripening.** *BMC Genomics* 2016, **17:**624.
1129 62. Yang L, Bennetzen JL: **Structure-based discovery and description of plant and**
1130 **animal Helitrons.** *Proc Natl Acad Sci U S A* 2009, **106:**12832-12837.
1131 63. Yang L, Bennetzen JL: **Distribution, diversity, evolution, and survival of Helitrons**
1132 **in the maize genome.** *Proc Natl Acad Sci U S A* 2009, **106:**19922-19927.
1133 64. Guo C, Spinelli M, Ye C, Li QQ, Liang C: **Genome-Wide Comparative Analysis of**
1134 **Miniature Inverted Repeat Transposable Elements in 19 Arabidopsis thaliana**
1135 **Ecotype Accessions.** *Sci Rep* 2017, **7:**2634.
1136 65. Gagliardi D, Manavella PA: **Short-range regulatory chromatin loops in plants.** *New*
1137 *Phytol* 2020, **228:**466-471.
1138 66. Gagliardi D, Cambiagno DA, Arce AL, Tomassi AH, Giacomelli JI, Ariel FD, Manavella
1139 PA: **Dynamic regulation of chromatin topology and transcription by inverted**
1140 **repeat-derived small RNAs in sunflower.** *Proc Natl Acad Sci U S A* 2019,
1141 **116:**17578-17583.
1142 67. Arce AL, Mencia R, Cambiagno DA, Lang PL, Liu C, Burbano HA, Weigel D, Manavella
1143 PA: **Polymorphic inverted repeats near coding genes impact chromatin topology**
1144 **and phenotypic traits in Arabidopsis thaliana.** *Cell Rep* 2023, **42:**112029.
1145 68. Mencia R, Arce AL, Houriet C, Xian W, Contreras A, Shirsekar G, Weigel D, Manavella
1146 PA: **Transposon-triggered epigenetic chromatin dynamics modulate EFR-related**
1147 **pathogen response.** *Nat Struct Mol Biol* 2025, **32:**199-211.
1148 69. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open**
1149 **Software Suite.** *Trends Genet* 2000, **16:**276-277.
1150 70. Li R, Xu Y, Xu Q, Tang J, Chen W, Luo Z, Liu H, Li W, Yan J, Springer NM, et al: **An**
1151 **epiallele of a gene encoding a PfkB-type carbohydrate kinase affects plant**
1152 **architecture in maize.** *Plant Cell* 2024, **37**.
1153 71. Li R, Yao J, Cai S, Fu Y, Lai C, Zhu X, Cui L, Li Y: **Genome-wide characterization**
1154 **and evolution analysis of miniature inverted-repeat transposable elements in**
1155 **Barley (Hordeum vulgare).** *Front Plant Sci* 2024, **15:**1474846.
1156 72. Lu C, Chen J, Zhang Y, Hu Q, Su W, Kuang H: **Miniature inverted-repeat**
1157 **transposable elements (MITEs) have been accumulated through amplification**
1158 **bursts and play important roles in gene expression and species diversity in**
1159 **Oryza sativa.** *Mol Biol Evol* 2012, **29:**1005-1017.
1160 73. Yang X, Liu Q, Wang MM, Wang XY, Han MQ, Liu FP, Lu TF, Liu J, Wang YZ: **A single**
1161 **dominant GLOBOSA allele accounts for repeated origins of hose-in-hose**
1162 **flowers in Sinningia (Gesneriaceae).** *Plant Cell* 2024, **37**.
1163 74. Sun Q, He Z, Wei R, Ye J, Chai L, Cheng Y, Xu Q, Deng X: **Red peel regulator 1**
1164 **links ethylene response factor 25 and beta-citraurin biosynthetic genes to**
1165 **regulate ethylene-induced peel reddening in citrus.** *Plant Cell* 2024, **37**.
1166 75. Defilippi BG, Kader AA, Dandekar AM: **Apple aroma: alcohol acyltransferase, a rate**
1167 **limiting step for ester biosynthesis, is regulated by ethylene.** *Plant Science* 2005,
1168 **168:**1199-1210.

1169 76. Zhang Z, Yu S, Zhang Z, Zhang J, Li H: **Comparative Characterization of Fruit**
1170 **Volatiles and Volatile-Related Genes Expression of 'Benihoppe' Strawberry and**
1171 **Its Somaclonal Mutant.** *Plants (Basel)* 2023, **12**.
1172 77. Li Y-p, Liu T-j, Luo H-f, Liu S-c: **The transcriptional landscape of cultivated**
1173 **strawberry (Fragaria×ananassa) and its diploid ancestor (Fragaria vesca) during**
1174 **fruit development.** *Journal of Integrative Agriculture* 2021, **20:**1540-1553.
1175 78. Nieuwenhuizen NJ, Maddumage R, Tsang GK, Fraser LG, Cooney JM, De Silva HN,
1176 Green S, Richardson KA, Atkinson RG: **Mapping, complementation, and targets of**
1177 **the cysteine protease actinidin in kiwifruit.** *Plant Physiol* 2012, **158:**376-388.
1178 79. Zhai Y, Cui Y, Song M, Vainstein A, Chen S, Ma H: **Papain-Like Cysteine Protease**
1179 **Gene Family in Fig (Ficus carica L.): Genome-Wide Analysis and Expression**
1180 **Patterns.** *Front Plant Sci* 2021, **12:**681801.
1181 80. Wang W, Wang Y, Chen T, Qin G, Tian S: **Current insights into posttranscriptional**
1182 **regulation of fleshy fruit ripening.** *Plant Physiol* 2023, **192:**1785-1798.
1183 81. Sanchez-Gomez C, Pose D, Martin-Pizarro C: **Insights into transcription factors**
1184 **controlling strawberry fruit development and ripening.** *Front Plant Sci* 2022,
1185 **13:**1022369.
1186 82. Vallarino JG, Merchante C, Sanchez-Sevilla JF, de Luis Balaguer MA, Pott DM, Ariza
1187 MT, Casanal A, Pose D, Vioque A, Amaya I, et al: **Characterizing the involvement**
1188 **of FaMADS9 in the regulation of strawberry fruit receptacle development.** *Plant*
1189 *Biotechnol J* 2020, **18:**929-943.
1190 83. Seymour GB, Ryder CD, Cevik V, Hammond JP, Popovich A, King GJ, Vrebalov J,
1191 Giovannoni JJ, Manning K: **A SEPALLATA gene is involved in the development**
1192 **and ripening of strawberry (Fragaria x ananassa Duch.) fruit, a non-climacteric**
1193 **tissue.** *J Exp Bot* 2011, **62:**1179-1188.
1194 84. Martinez-Rivas FJ, Blanco-Portales R, Molina-Hidalgo FJ, Caballero JL, Perez de
1195 Souza L, Alseekh S, Fernie AR, Munoz-Blanco J, Rodriguez-Franco A: **Azacytidine**
1196 **arrests ripening in cultivated strawberry (Fragaria x ananassa) by repressing key**
1197 **genes and altering hormone contents.** *BMC Plant Biol* 2022, **22:**278.
1198 85. Liu L, Xiao W, Ji M-l, Yang C, Li L, Gao D-s, Fu X-l: **Effects of molybdenum on**
1199 **nutrition, quality, and flavour compounds of strawberry (Fragaria×ananassa**
1200 **Duch. cv. Akihime) fruit.** *Journal of Integrative Agriculture* 2017, **16:**1502-1512.
1201 86. Cosgrove DJ: **Growth of the plant cell wall.** *Nat Rev Mol Cell Biol* 2005, **6:**850-861.
1202 87. Del Olmo I, Romero I, Alvarez MD, Tarradas R, Sanchez-Ballesta MT, Escribano MI,
1203 Merodio C: **Transcriptomic analysis of CO(2)-treated strawberries (Fragaria**
1204 **vesca) with enhanced resistance to softening and oxidative stress at**
1205 **consumption.** *Front Plant Sci* 2022, **13:**983976.
1206 88. Ariel FD, Manavella PA: **When junk DNA turns functional: transposon-derived**
1207 **non-coding RNAs in plants.** *J Exp Bot* 2021, **72:**4132-4143.
1208 89. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R: **Population scale**
1209 **mapping of transposable element diversity reveals links to gene regulation and**
1210 **epigenomic variation.** *Elife* 2016, **5**.
1211 90. Catlin NS, Josephs EB: **The important contribution of transposable elements to**
1212 **phenotypic variation and evolution.** *Curr Opin Plant Biol* 2022, **65:**102140.
1213 91. Urrutia M, Meco V, Rambla JL, Martin-Pizarro C, Pillet J, Andres J, Sanchez-Sevilla
1214 JF, Granell A, Hytonen T, Pose D: **Diversity of the volatilome and the fruit size and**
1215 **shape in European woodland strawberry (Fragaria vesca).** *Plant J* 2023, **116:**1201-
1216 1217.
1217 92. Baduel P, Leduque B, Ignace A, Gy I, Gil J, Jr., Loudet O, Colot V, Quadrana L:
1218 **Genetic and environmental modulation of transposition shapes the evolutionary**
1219 **potential of Arabidopsis thaliana.** *Genome Biol* 2021, **22:**138.

1220 93. Sanchez-Sevilla JF, Cruz-Rus E, Valpuesta V, Botella MA, Amaya I: **Deciphering**
1221 **gamma-decalactone biosynthesis in strawberry fruit using a combination of**
1222 **genetic mapping, RNA-Seq and eQTL analyses.** *BMC Genomics* 2014, **15:**218.
1223 94. Schwieterman ML, Colquhoun TA, Jaworski EA, Bartoshuk LM, Gilbert JL, Tieman
1224 DM, Odabasi AZ, Moskowitz HR, Folta KM, Klee HJ, et al: **Strawberry flavor: diverse**
1225 **chemical compositions, a seasonal influence, and effects on sensory**
1226 **perception.** *PLoS One* 2014, **9:**e88446.
1227 95. Cano-Salazar J, López ML, Crisosto CH, Echeverría G: **Volatile compound**
1228 **emissions and sensory attributes of 'Big Top' nectarine and 'Early Rich' peach**
1229 **fruit in response to a pre-storage treatment before cold storage and subsequent**
1230 **shelf-life.** *Postharvest Biology and Technology* 2013, **76:**152-162.
1231 96. Zhang B, Shen JY, Wei WW, Xi WP, Xu CJ, Ferguson I, Chen K: **Expression of genes**
1232 **associated with aroma formation derived from the fatty acid pathway during**
1233 **peach fruit ripening.** *J Agric Food Chem* 2010, **58:**6157-6165.
1234 97. Alvim Kamei CL, Boruc J, Vandepoele K, Van den Daele H, Maes S, Russinova E,
1235 Inze D, De Veylder L: **The PRA1 gene family in Arabidopsis.** *Plant Physiol* 2008,
1236 **147:**1735-1749.
1237 98. Farrow SC, Facchini PJ: **Functional diversity of 2-oxoglutarate/Fe(II)-dependent**
1238 **dioxygenases in plant metabolism.** *Front Plant Sci* 2014, **5:**524.
1239 99. Wei S, Zhang W, Fu R, Zhang Y: **Genome-wide characterization of 2-oxoglutarate**
1240 **and Fe(II)-dependent dioxygenase family genes in tomato during growth cycle**
1241 **and their roles in metabolism.** *BMC Genomics* 2021, **22:**126.
1242 100. Wei L, Mao W, Jia M, Xing S, Ali U, Zhao Y, Chen Y, Cao M, Dai Z, Zhang K, et al:
1243 **FaMYB44.2, a transcriptional repressor, negatively regulates sucrose**
1244 **accumulation in strawberry receptacles through interplay with FaMYB10.** *J Exp*
1245 *Bot* 2018, **69:**4805-4820.
1246 101. Gonzalez M, Gaete-Eastman C, Valdenegro M, Figueroa CR, Fuentes L, Herrera R,
1247 Moya-Leon MA: **Aroma development during ripening of Fragaria chiloensis fruit**
1248 **and participation of an alcohol acyltransferase (FcAAT1) gene.** *J Agric Food*
1249 *Chem* 2009, **57:**9123-9132.
1250 102. Beekwilder J, Alvarez-Huerta M, Neef E, Verstappen FW, Bouwmeester HJ, Aharoni
1251 A: **Functional characterization of enzymes forming volatile esters from**
1252 **strawberry and banana.** *Plant Physiol* 2004, **135:**1865-1878.
1253 103. Studer A, Zhao Q, Ross-Ibarra J, Doebley J: **Identification of a functional**
1254 **transposon insertion in the maize domestication gene tb1.** *Nat Genet* 2011,
1255 **43:**1160-1163.
1256 104. Gill RA, Scossa F, King GJ, Golicz AA, Tong C, Snowdon RJ, Fernie AR, Liu S: **On**
1257 **the Role of Transposable Elements in the Regulation of Gene Expression and**
1258 **Subgenomic Interactions in Crop Genomes.** *Critical Reviews in Plant Sciences*
1259 2021, **40:**157-189.
1260 105. Chambers AH, Pillet J, Plotto A, Bai J, Whitaker VM, Folta KM: **Identification of a**
1261 **strawberry flavor gene candidate using an integrated genetic-genomic-**
1262 **analytical chemistry approach.** *BMC Genomics* 2014, **15:**217.
1263 106. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W,
1264 Hellsten U, Putnam N, Rokhsar DS: **Phytozome: a comparative platform for green**
1265 **plant genomics.** *Nucleic Acids Res* 2012, **40:**D1178-1186.
1266 107. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
1267 Maurano MT, Vierstra J, Thomas S, et al: **BEDOPS: high-performance genomic**
1268 **feature operations.** *Bioinformatics* 2012, **28:**1919-1920.
1269 108. Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next-**
1270 **generation sequencing data.** *Bioinformatics* 2012, **28:**3150-3152.

109. Gambino G, Perrone I, Gribaudo I: **A Rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants.** *Phytochem Anal* 2008, **19:**520-525.
110. Ewels P, Magnusson M, Lundin S, Kaller M: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics* 2016, **32:**3047-3048.
111. Axtell MJ: **ShortStack: comprehensive annotation and quantification of small RNA genes.** *RNA* 2013, **19:**740-751.
112. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.
113. Lu P, Yu S, Zhu N, Chen YR, Zhou B, Pan Y, Tzeng D, Fabi JP, Argyris J, Garcia-Mas J, et al: **Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening.** *Nat Plants* 2018, **4:**784-791.
114. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.
115. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.
116. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
117. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30:**923-930.
118. Koster J, Rahmann S: **Snakemake--a scalable bioinformatics workflow engine.** *Bioinformatics* 2012, **28:**2520-2522.
119. Nunn A, Can SN, Otto C, Fasold M, Diez Rodriguez B, Fernandez-Pozo N, Rensing SA, Stadler PF, Langenberger D: **EpiDiverse Toolkit: a pipeline suite for the analysis of bisulfite sequencing data in ecological plant epigenetics.** *NAR Genom Bioinform* 2021, **3:**lqab106.
120. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H: **Twelve years of SAMtools and BCFtools.** *Gigascience* 2021, **10**.
121. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
122. Faust GG, Hall IM: **YAHA: fast and flexible long-read alignment with optimal breakpoint detection.** *Bioinformatics* 2012, **28:**2417-2424.
123. Lê S, Josse J, Husson F: **FactoMineR: An R Package for Multivariate Analysis.** *Journal of Statistical Software* 2008, **25:**1 - 18.
124. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES: **TASSEL: software for association mapping of complex traits in diverse samples.** *Bioinformatics* 2007, **23:**2633-2635.
125. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T: **deepTools2: a next generation web server for deep-sequencing data analysis.** *Nucleic Acids Res* 2016, **44:**W160-165.
126. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19:**1639-1645.
127. Diesh C, Stevens GJ, Xie P, De Jesus Martinez T, Hershberg EA, Leung A, Guo E, Dider S, Zhang J, Bridge C, et al: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol* 2023, **24:**74.

128. Lawrence M, Gentleman R, Carey V: **rtracklayer: an R package for interfacing with genome browsers.** *Bioinformatics* 2009, **25:**1841-1842.

FIGURE LEGENDS

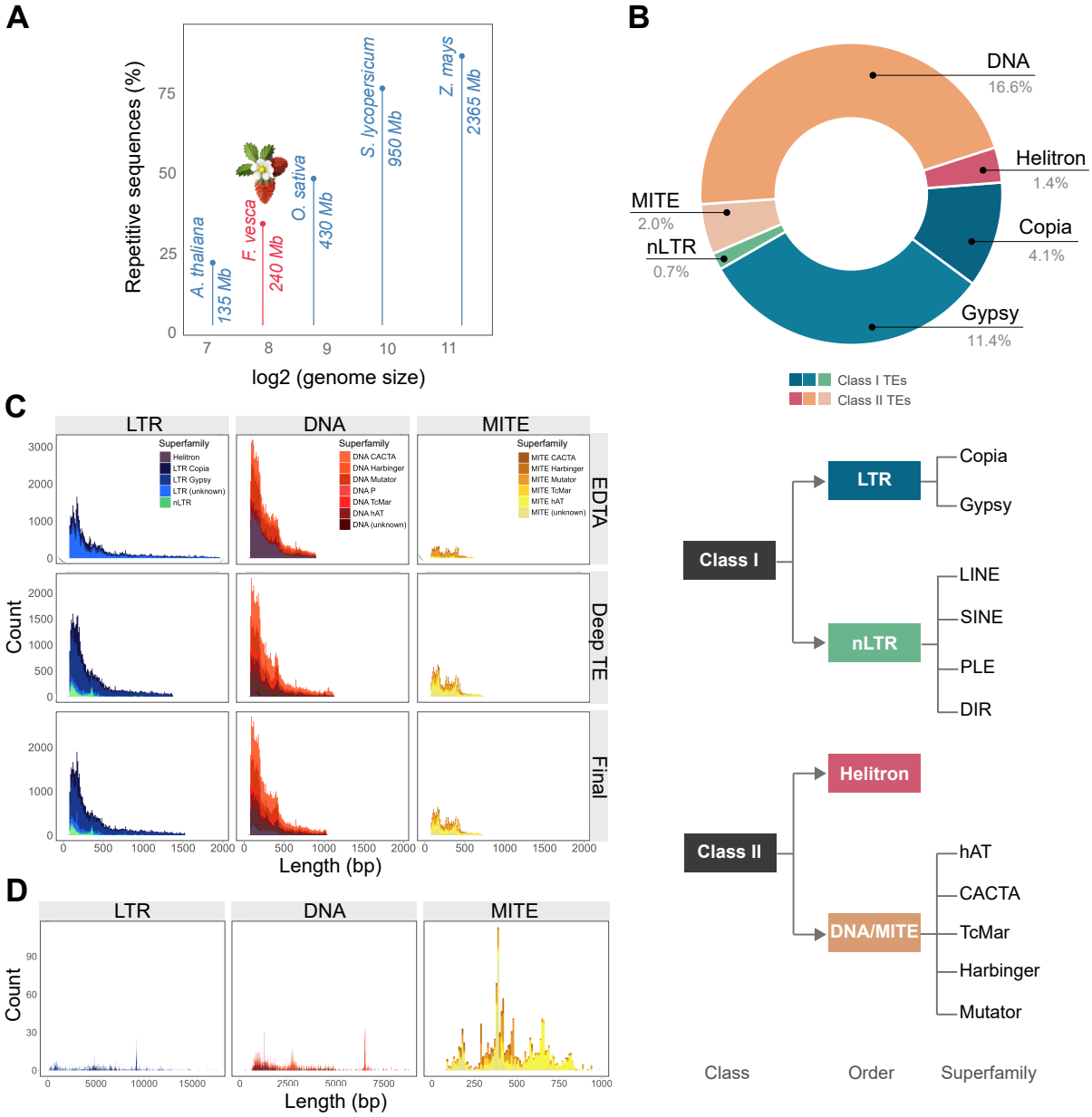

Figure 1. Transposable Element (TE) diversity in *Fragaria vesca*. (A) Repetitive sequence content as a proportion of total genome size in several plant species, listed in ascending order: *Arabidopsis thaliana*, woodland strawberry (*Fragaria vesca*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), and maize (*Zea mays*). **(B)** Total coverage of the *F. vesca* genome assembly by the main TE classes, shown as percentages relative to the entire genome. Representation of TE classification scheme generated in this study is displayed. **(C)** Length distribution of TEs (including fragmented TEs) in the *F. vesca* v4.0.a2 annotation. The

classifications are listed in descending order for EDTA, DeepTE, and the final annotation. **(D)** Length distribution of intact TEs in the final *F. vesca* v4 annotation.

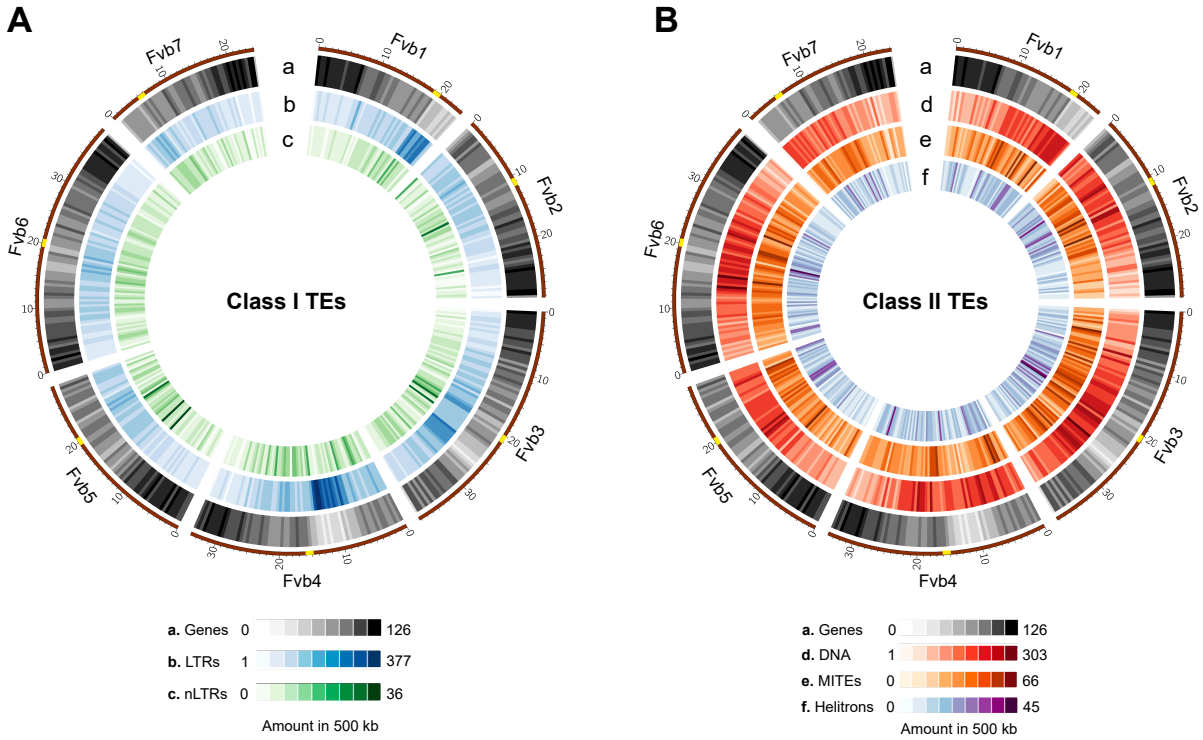

Figure 2. Distribution of TEs in the *F. vesca* genome, represented as density plots within 500 kb windows. (A) Class I TEs: LTR (Track b in blue), non-LTR (Track c in green). **(B)** Class II TEs: DNA (Track "d" red), MITE (Track "e" in orange), Helitron (Track "f" in purple). The color scale indicates the number of elements found within each 500 kb window. The distribution of genes is shown in grayscale (Track "a") in both graphs. Centromeres are marked as a yellow dot in the brown outer line.

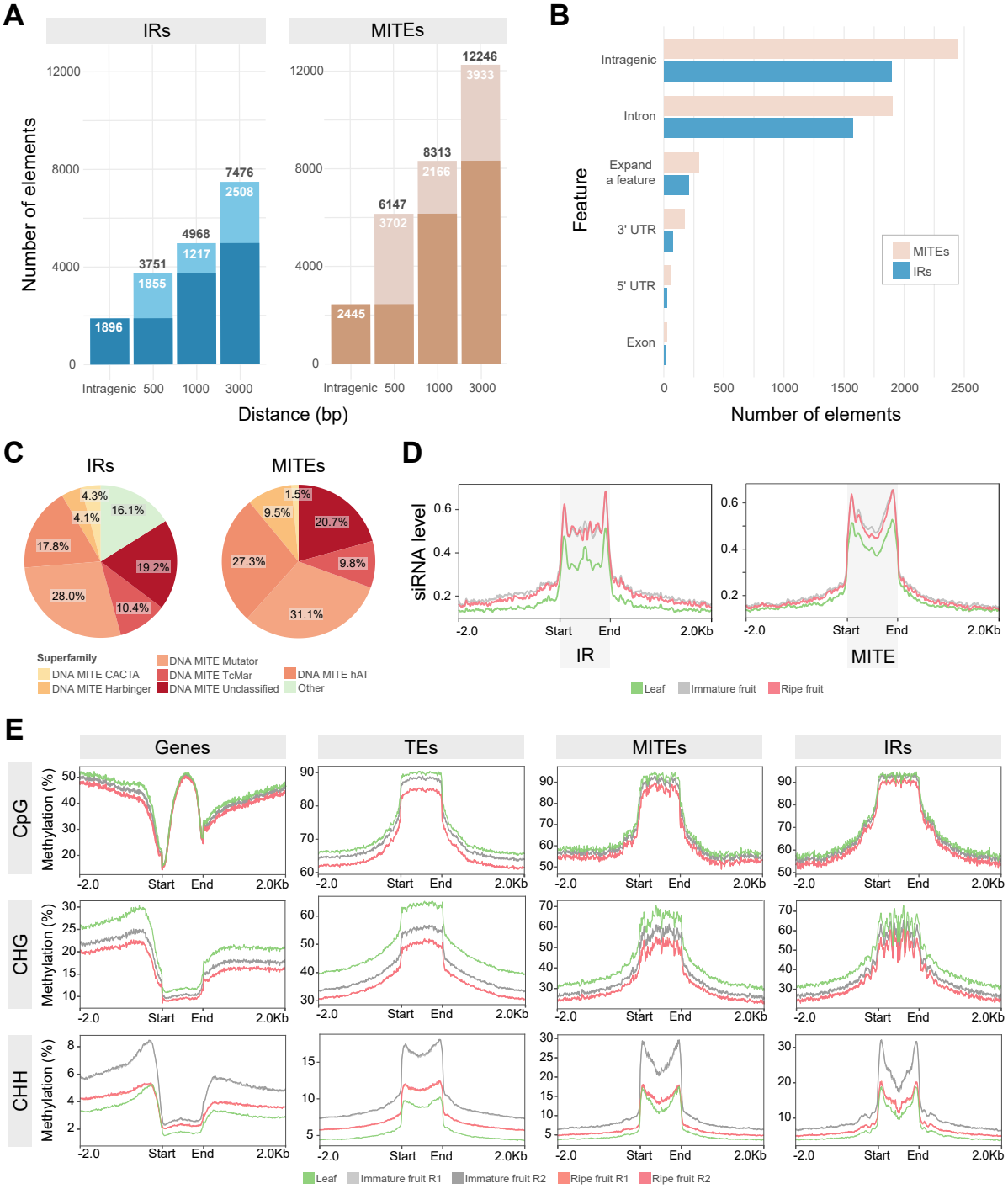

Figure 3. Characterization of TE-Derived IRs and MITEs in *F. vesca*.
(A) Number of IRs (left) and MITEs (right) detected in the *F. vesca* v4.0.a2 reference genome, counting elements located within gene bodies ('intragenic') or within the upstream or downstream regions defined as 0–500 bp ('500'), 501–1000 bp ('1000'), and 1001–3000 bp ('3000') of annotated genes (the counts for these flanking regions also include any intragenic elements). Black numbers indicate the total counts of IRs and MITEs, while white numbers show the non-cumulative counts. **(B)** Number of IRs and MITEs that are completely contained within a gene ('intragenic'). Among these, the number that are entirely within a single gene feature ('exon', 'intron', '3′-UTR', or '5′-UTR') or that span more than one gene feature as indicated. **(C)** Proportion of the identified IRs (left) and annotated MITEs (right) classified by TE superfamily in *F. vesca*. **(D)** Metagene profile of 24-nt siRNAs mapping to all IRs and MITEs

in leaf (green), and immature (gray) and ripe fruits (pink). Plots show MITEs/IRs scaled from the start to the end plus 2,000 bp on each side. sRNA-seq replicates are plotted together. **(E)** Metagene profile of CpG, CHG, and CHH DNA methylation of all genes, TEs, MITEs, and IRs in leaf (green), immature (gray), and ripe fruits (pink). Plots show Genes, TEs, MITEs, and IRs scaled from the start to the end plus 2,000 bp on each side. Individual BS-seq replicates are plotted.

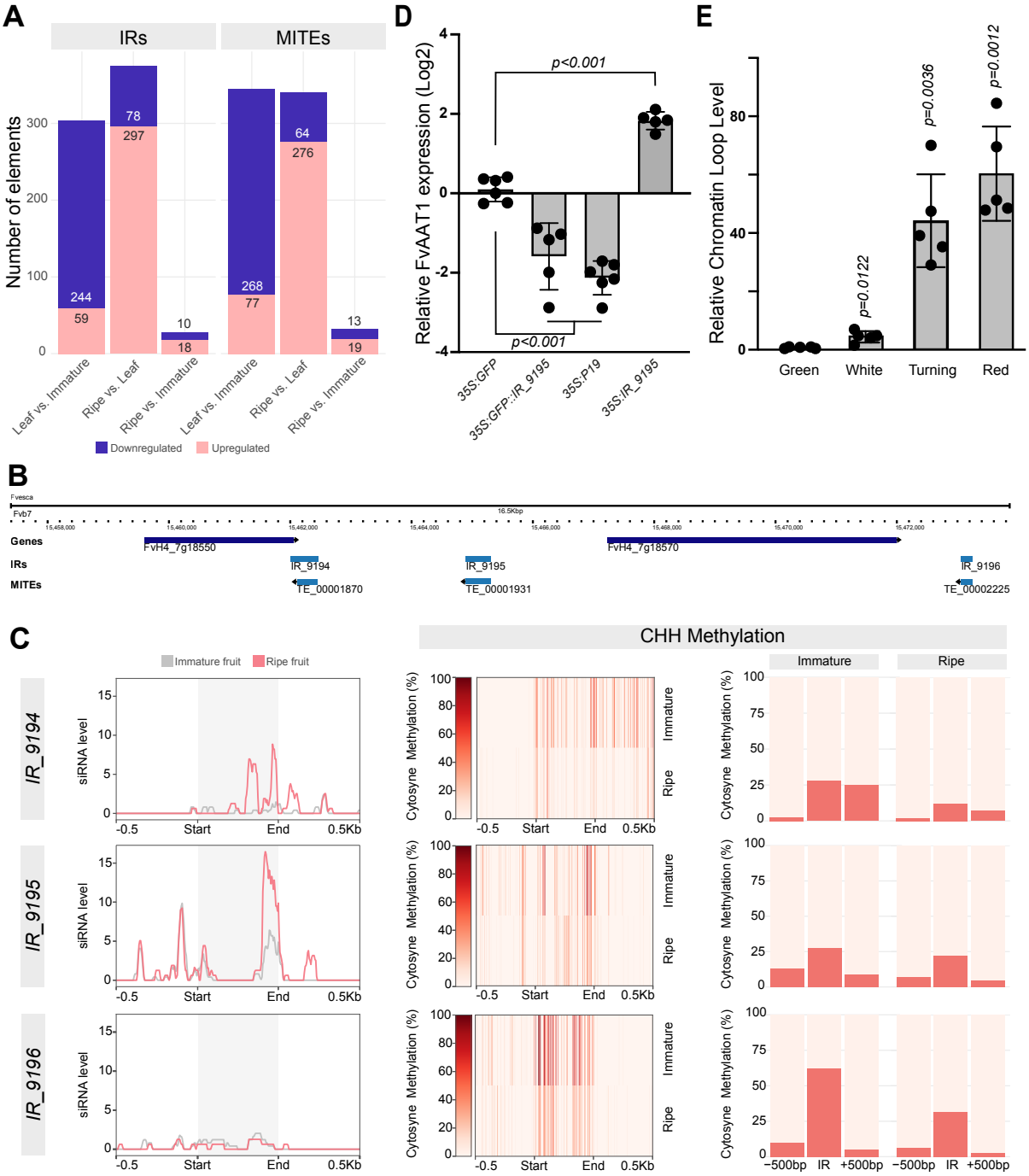

Figure 4. Regulatory role of MITEs and IRs in gene expression.
(A) Differential accumulation analysis of 24-nt siRNAs derived from IRs and MITEs across leaf tissue, immature and ripe fruits, using thresholds of |log2-fold change| ≥ 1 and an adjusted p-value (padj) ≤ 0.05 **(B)** Genomic region of the *F. vesca* v4 assembly containing the *IR_9195*

locus near FvH4_7g18550 and FvH4_7g18570 genes. IRs and MITEs identified in this region are also displayed. **(C)** Left: 24-nt siRNAs mapping to *IR_9194*, *IR_9195*, and *IR_9196* (highlighted by a gray box) in immature (gray) and ripe (pink) fruits. Each IR is scaled from start to end plus 500 bp on each side, and all sRNA-seq replicates are plotted together. Right: Cytosine methylation in the CHH context for *IR_9194*, *IR_9195*, and *IR_9196*, including 500-bp flanking regions, in immature and ripe fruits. The average of individual BS-seq replicates is plotted for each condition. Right: average percentage of CHH methylation in each IR and its 500-bp flanking regions. **(D)** *FvAAT1* expression in transiently transformed strawberry receptacles. RNA levels are expressed relative to the control samples (35S:GFP) plants. Values of n = 6 biologically independent samples. Data are presented as the mean values ± s.d. P values were calculated with a two-tailed unpaired t-test with Welch's correction. **(E)** Chromatin conformation capture (3C) analysis showing the interaction frequency of *IR_9195* with the FvAAT1 locus in immature and ripe fruits as measured by 3C–qPCR. Values of n = 4 biologically independent samples are expressed relative to green immature stage. Data are presented as the mean values ± s.d. P values were calculated with a two-tailed unpaired t-test with Welch's correction.

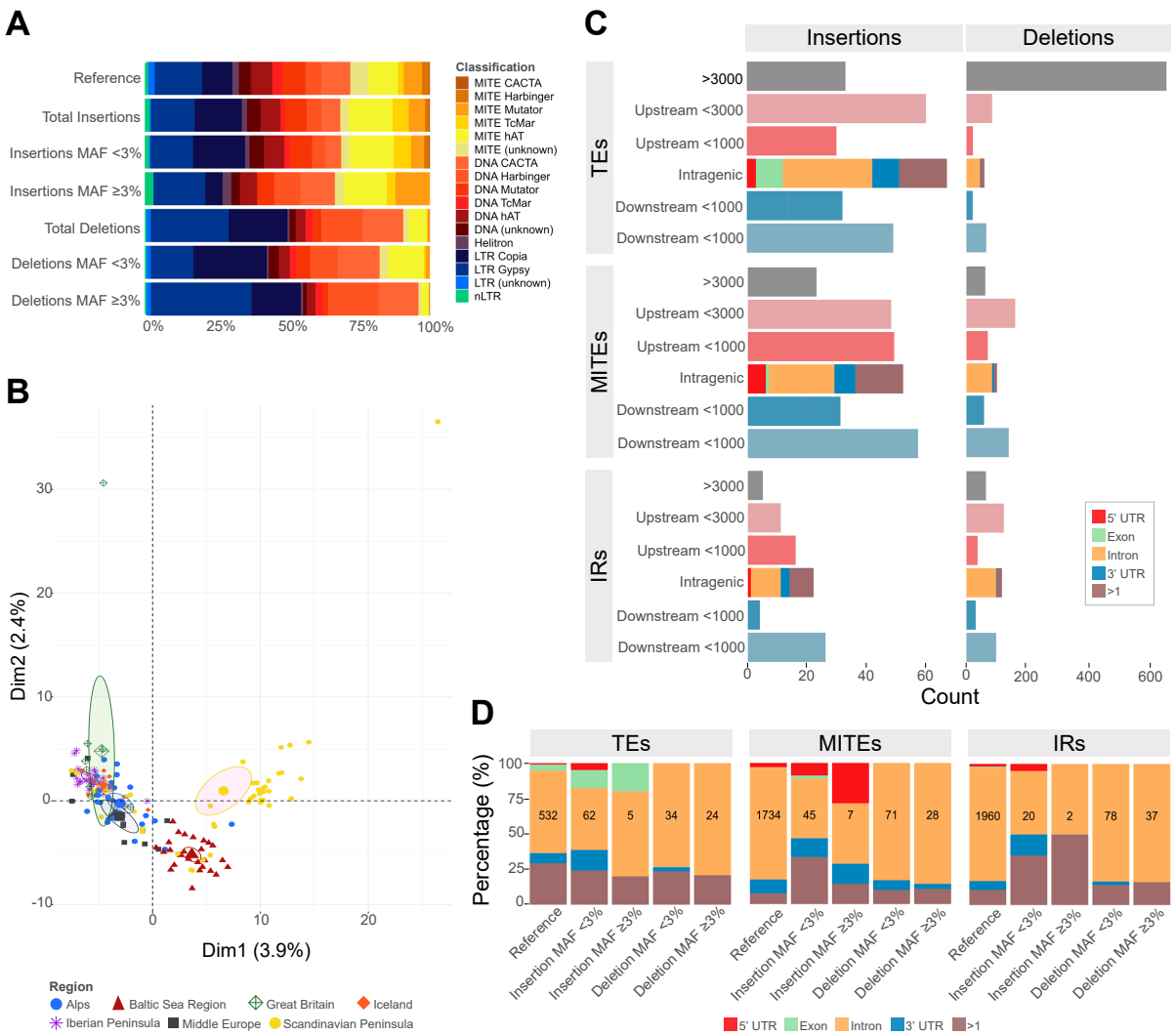

Figure 5. Distribution of insertion and deletion polymorphisms in the *F. vesca* v4.0.a2 annotation.

(A) Proportion of identified polymorphisms by TE superfamily. Rare TE polymorphisms are indicated by minor allele frequency (MAF) < 3% and frequent TE polymorphisms by MAF ≥

3%. **(B)** Principal Component Analysis (PCA) based on TE insertion and deletion polymorphisms. The geographic regions of the analyzed *F. vesca* accessions are indicated according to their clustering, as reported by Urrutia et al [91]. Ellipses represent the 95% confidence intervals around the mean for each geographic region. **(C)** Number of polymorphic elements located within genes ('intragenic'), within 500 bp, 1000 bp, or 3000 bp upstream or downstream of genes, or farther than 3000 bp from genes (>3000). **(D)** Proportion of TEs, MITEs, and IRs that fully overlap each gene annotation feature (5′ UTR, intron, exon, 3′ UTR) or overlap more than one feature (>1). The total number of elements in each category is indicated.

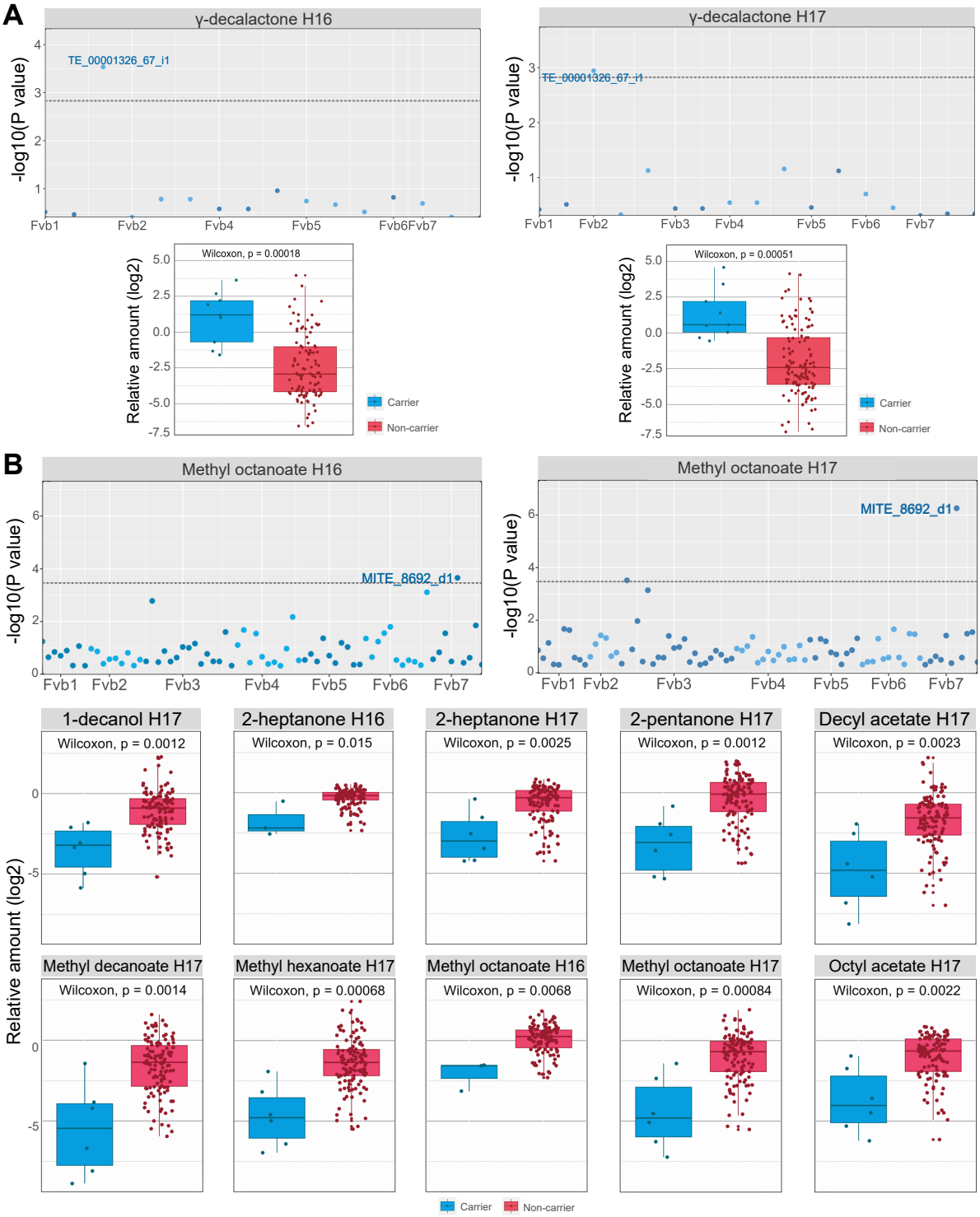

Figure 6. TE polymorphisms linked to volatile compound variation in a collection of *F. vesca* accessions.

(A) Association between the TE insertion polymorphism 'TE_00001326_67_i1' and the relative amount of γ-decalactone in the H16 and H17 harvests. Top: Manhattan plot illustrating the association between TE insertion polymorphisms and the relative amount of γ-decalactone. A dotted gray line marks the significance threshold determined by the Bonferroni test (threshold = 2.82), and the polymorphism with a significant association is highlighted. Bottom: Box plot showing the relative amount of γ-decalactone for accessions carrying the polymorphism (light blue) versus non-carriers (red). The p-value from the Wilcoxon test is

1419 indicated. **(B)** Association between a MITE deletion polymorphism and the relative amounts
1420 of eight volatile compounds in the H16 and H17 harvests. Top: Manhattan plot illustrating the
1421 association between the deletion polymorphism 'MITE_8692_d1' and the relative amount of
1422 methyl octanoate for the H16 and H17 harvests. A dotted gray line indicates the significance
1423 threshold determined via the Bonferroni test (threshold = 3.47), and the polymorphism with a
1424 significant association is labeled. Bottom: Box plots showing the relative amounts of the
1425 compounds that are significantly associated with the MITE deletion polymorphism
1426 'MITE_8692_d1' for accessions carrying the polymorphism (light blue) and non-carriers (red).
1427 The p-value obtained from the Wilcoxon test is indicated.
1428

1429 **TABLES**

1430 **Table 1.** Number of unique TEs, number of copies, and total TE coverage of the *F. vesca*
1431 genome assembly v4.0, according to their classification.

Classification	N° of TEs	N° of Copies	Genome Coverage (%)
Class I			
LTR			
Copia	635	12,903	4.09
Gypsy	632	30,584	11.38
Unclassified	136	4,786	0.96
Total LTRs	1,403	48,273	16.43
non-LTR			
LINE (L1)	34	760	0.18
SINE (tRNA)	3	1,649	0.12
PLE	19	1,273	0.27
DIRS	3	80	0.02
Unclassified	2	4	0.00
Total non-LTRs	61	3,766	0.65
Total Class I	1,464	52,039	17.08
Class II			
Subclass 1			
MITE			
CACTA	28	223	0.03
Harbinger	80	1,431	0.17
Mutator	218	4,683	0.60
TcMar	86	1,478	0.16
hAT	257	4,111	0.63
Unclassified	132	3,114	0.36
Total MITEs	801	15,040	1.95

DNA (non-MITE)			
CACTA	247	24,711	7.58
Harbinger	85	3,748	1.43
Mutator	353	14,507	2.53
P	1	6	0.00
TcMar	142	4,146	0.94
hAT	324	9,348	2.09
Unclassified	220	10,403	2.06
Total DNA	1,372	66,869	16.63
Subclass 2			
Helitron	55	5,987	1.35
Total Class II	2,228	87,896	19.93

Table 2. Polymorphisms detected by TEPID for each element annotation. MAF ≥ 3% refers to polymorphisms with a minor allele frequency above 3%. The percentage shown represents the proportion of each value relative to the total number of polymorphic elements detected for each element and each type (insertion/deletion).

		TEs		MITEs		IRs	
Polymorphisms		**Amount**	**%**	**Amount**	**%**	**Amount**	**%**
Insertions	Total	229	100.0	213	100.0	71	100.0
	MAF ≥ 3%	33	14.4	32	15.0	12	16.9
	MAF < 3%	196	85.6	181	85.0	59	83.1
Deletions	Total	853	100.0	459	100.0	388	100.0
	MAF ≥ 3%	519	60.8	145	31.6	123	31.7
	MAF < 3%	334	39.2	314	68.4	265	68.3

1459 **SUPPLEMENTARY MATERIAL**
1460 **Table S1.** Size distribution of Transposable Elements by Class annotated in *Fragaria vesca*,
1461 including both fragmented and intact TEs.
1462
1463 **Table S2.** ShortStack analysis of sRNA-seq data from *Fragaria vesca* leaves, immature (white
1464 stage), and ripe (red stage) fruits using annotated loci files for MITEs and IRs.
1465
1466 **Table S3.** Differential expression analysis of 24-nt siRNA-producing MITEs and adjacent
1467 genes (<3 kb) in *Fragaria vesca*. Genomic coordinates, functional annotation and expression
1468 changes between leaf tissue, and immature and ripe stages are shown.
1469
1470 **Table S4.** Differential expression analysis of 24-nt siRNA-producing IRs and adjacent genes
1471 (<3 kb) in *Fragaria vesca*. Genomic coordinates, functional annotation and expression
1472 changes between leaf tissue, and immature and ripe stages are shown.
1473
1474 **Table S5.** Differential gene expression analysis of all *F. vesca* genes across conditions: leaf
1475 tissue vs. immature stage, ripe stage vs. leaf tissue, ripe vs. immature stages, and AZA-treated
1476 vs. control strawberries.
1477
1478 **Table S6.** Resequenced genomes of *Fragaria vesca* analyzed in the present study, including
1479 accession identifiers, country and origin coordinates, read counts, and average chromosome
1480 coverage (excluding unassembled contigs). Accessions in italics were excluded from analysis
1481 due to low coverage; accessions in bold were included in the GWAS study.
1482
1483 **Figure S1. Annotation and classification of transposable elements (TEs) in the *Fragaria***
1484 ***vesca* v4.0.a2 reference genome. (A)** Representation of the pipeline used in this work for
1485 transposon annotation. Software employed are highlighted in blue boxes, while input and
1486 output files are shown in pink boxes, with their formats in parentheses. **(B)** Distribution of TE
1487 superfamilies as annotated by EDTA (left), by DeepTE (center), and by their combined pipeline
1488 (right) for the *F. vesca* v4.0 genome.
1489
1490 **Figure S2. The average length of intact TEs and all annotated TEs in *F. vesca* genome**
1491 **(including fragmented TEs).**
1492
1493 **Figure S3. Distribution of intact TE superfamilies in the *Fragaria vesca* v4.0.a2**
1494 **annotation based on their distance from a gene**. Categories include: within a gene
1495 ('intragenic'), within 1 kb ('1000'), between 1 and 3 kb ('3000'), and more than 3 kb ('>3000').
1496 Left: Total counts by superfamily. Right: Proportion by superfamily.
1497
1498 **Figure S4. Proportion of IRs classified as 'Other', excluding MITE-type elements.**
1499 Values are shown as percentages relative to the total number of IRs in the 'Other' category
1500 (16.1%).
1501
1502 **Figure S5. (A)** Metagene profile of 24-nt siRNAs mapping to MITEs and IRs located at
1503 different ranges of distance to the closest protein-coding gene, in leaf (green), and immature
1504 (gray) and ripe fruits (pink). Plots show MITEs/IRs scaled from the start to the end plus 2,000
1505 bp to each side. sRNA-seq replicates are plotted together. **(B)** Metagene profile of CpG, CHG,
1506 and CHH DNA methylation at MITEs/IRs located at different ranges of distance to the closest
1507 protein-coding gene, in leaf (green), and immature (gray) and ripe fruits (pink). Plots show

1508 MITEs/IRs scaled from the start to the end plus 2,000 bp to each side. Individual BS-seq
1509 replicates are plotted.
1510
1511 **Figure S6. (A)** Differential accumulation analysis of 24-nt siRNAs derived from IRs (left) and
1512 MITEs (right) elements across leaf tissue, immature and ripe fruits, using thresholds of log2-
1513 fold change (FC) ≥0.5 and adjusted p-value (padj) ≤0.05. The percentages of elements located
1514 within 3000 bp of either differentially expressed genes (DE) or non-DE genes are shown. **(B)**
1515 Volcano plots showing the log2FC of the differential expression analysis for IRs (left) and
1516 MITEs (right) between immature and ripe stages. Red points represent elements with an
1517 adjusted p-value ≤0.05. One element from each analysis, which had an extremely low adjusted
1518 p-value, was removed only for clarity.
1519
1520 **Figure S7. Conservation of *IR_9195* from *Fragaria vesca* in the orthologous genomic**
1521 **regions of commercial strawberry species *Fragaria chiloensis* and *Fragaria* × *ananassa*.**
1522 **(A)** Local alignment of *F. vesca IR_9195* with the IR located near the FvH4_7g18570 ortholog
1523 in *F. chiloensis* (left) and *F.* × *ananassa* (right). **(B)** Minimum free energy (MFE) RNA
1524 secondary structure predicted by RNAfold server of each IR. The color scale shows base-
1525 pairing probability 0 (purple) to 1 (red).
1526
1527 **Figure S8.** Region of the F. vesca v4.0.a2 reference genome containing the FvH4_6g46420
1528 **(A)**, FvH4_2g30230 **(B),** and FvH4_5g23180 **(C)** loci displaying the IR and MITE elements
1529 annotated in the region and expression and epigenetic profiles. (I) 24-nt siRNAs mapping to
1530 the genomic regions as determined by sRNA sequencing of leaf (green), immature (gray) and
1531 ripe (pink) tissues. Replicates are plotted together. (II) Expression of each gene in leaf (green),
1532 immature (gray) and ripe (pink) tissues as well as strawberries treated with 5-azacytidine (5-
1533 AZA) and its control, measured by RNA-seq. (III) Cytosine DNA methylation in CG, CHG, and
1534 CHH contexts in leaf (green), immature (gray) and ripe (pink) fruits. The average of individual
1535 BS-seq replicates is plotted for each condition.
1536
1537 **Figure S9. (A)** Representation of the pipeline used in this study. The main inputs are shown
1538 at the top, while the main outputs are placed at the bottom. File formats are indicated in
1539 parentheses. **(B)** Representation of the TEPID pipeline used for identifying transposable
1540 element insertion polymorphisms (TIPs) and absence polymorphisms (TAPs).
1541
1542 **Figure S10. Association between a MITE deletion polymorphism and the relative**
1543 **amounts of volatile compounds in the H16 and H17 harvests.** Manhattan plots illustrate
1544 the association between the deletion polymorphism 'MITE_8692_d1' and the relative amount
1545 of 1-decanol, 2-heptanone, 2-pentanone, decyl acetate, methyl decanoate, methyl hexanoate
1546 and octyl acetate, for the H16 and H17 harvests as indicated. A gray line indicates the
1547 significance threshold determined via the Bonferroni test (threshold = 3.47), and the
1548 polymorphism with a significant association is labeled.
1549
1550 **Figure S11. TE polymorphisms linked to phenotypic characteristics of woodland**
1551 **strawberry fruits. (A)** Association between the TE deletion polymorphism
1552 'TE_00001968_9_d1' and the relative quantity of (Z)-3-hexenyl acetate in the H16 harvest.
1553 Top: Manhattan plot showing the association between TE deletion polymorphisms and the
1554 relative quantity of (Z)-3-hexenyl acetate. A dotted gray line indicates the significance
1555 threshold calculated using the Bonferroni test (threshold = 4.05), and the polymorphism with
1556 a significant association is labeled. Bottom: Box plot of the relative quantity of (Z)-3-hexenyl
1557 acetate for accessions carrying the polymorphism (light blue) versus non-carriers (red). The
1558 p-value from t-test is shown. **(B)** Association between the TE deletion polymorphism
1559 'TE_00000043_1_d1' and average fruit volume in the H16 harvest. Top: Manhattan plot

1560 showing the association between TE deletion polymorphisms and average strawberry fruit
1561 volume. The significance threshold calculated via the Bonferroni test (threshold = 4.05) is
1562 indicated by a dotted gray line, and the polymorphism with significant association is labeled.
1563 Bottom: Box plot of average fruit volume for accessions carrying the polymorphism (light blue)
1564 and non-carriers (red). The p-value from Wilcoxon test is shown.