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

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

AFFILIATIONS

[1]Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" (IHSM "La Mayora"), Universidad de Málaga-Consejo Superior de Investigaciones Científicas (UMA-CSIC), 29010 Málaga, Spain.

[2]Facultad de Ingeniería, Universidad Nacional de Entre Ríos, 3100 Oro Verde, Argentina.

[3]Department of Agricultural Sciences, Viikki Plant Science Centre, University of Helsinki, Finland.

[4]Instituto de Agrobiotecnología del Litoral (CONICET-UNL), 3000 Santa Fe, Argentina.

[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: ileana.tossolini@uner.edu.ar; pablomanavella@ihsm.uma-csic.es; cmmartin@uma.es

ABSTRACT

Transposable elements (TEs) are major components of plant genomes and powerful drivers of regulatory and phenotypic diversity, yet their mechanistic contributions to gene regulation in non-model crops remain poorly understood. Here, we generate a comprehensive and curated TE annotation for woodland strawberry (*Fragaria vesca*), revealing that TEs comprise ~30% of the genome and uncovering an exceptional enrichment of gene-proximal miniature inverted-repeat transposable elements (MITEs) and short inverted repeats (IRs).

Integrating small RNA, DNA methylation and transcriptomic data across tissues and fruit developmental stages, we show that most gene-associated MITEs and IRs are epigenetically active, producing 24-nt siRNAs and displaying dynamic CHH methylation patterns during fruit ripening. Genome-wide analyses reveal that MITE- and IR-proximal genes occupy a distinct

regulatory landscape compared to genes associated with other TE classes, consistent with a role in context-dependent transcriptional regulation.

As a mechanistic proof of concept, we demonstrate that a Mutator-like MITE located near *FvAAT1*, encoding a key alcohol acyltransferase required for aroma ester biosynthesis, modulates gene expression through siRNA-dependent DNA methylation and ripening-stage-specific chromatin reorganization. Transient perturbation of this MITE epigenetic state alters *FvAAT1* expression, and chromatin conformation capture reveals MITE-associated looping at this locus. Notably, this regulatory MITE is conserved in commercial strawberry species, suggesting an evolutionarily preserved function in fruit quality traits. Extending these findings to natural variation, we characterize thousands of TE insertion and deletion polymorphisms across 210 European *F. vesca* accessions and show that TE-based variation recapitulates population structure. A TE-informed genome-wide association analysis identifies significant associations between MITE and IR polymorphisms and fruit volatile compounds, including key aroma determinants.

Together, our results establish MITEs and short IRs as widespread, epigenetically active regulatory elements that shape gene expression, chromatin architecture and phenotypic diversity in strawberry. This work provides a conceptual and methodological framework to integrate TE-driven regulatory variation into population genomics, breeding and genome-editing strategies aimed at improving 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 60% 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, known as DNA transposons, move via a “cut-and-paste” mechanism [9, 10] or, in the case of Helitrons, a “peel-and-paste” replicative mechanism that uses a circular DNA intermediate [11]. 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.

Beyond their contribution to genome size and structure, specific types of TEs are increasingly recognized for their regulatory potential. In this context, miniature inverted-repeat transposable elements (MITEs) represent a prominent group of short (<800 bp), non-autonomous Class II TEs that are highly abundant in eukaryotic genomes [7, 12, 13]. 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 [14, 15]. 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 RNA-dependent RNA Polymerase 2 (RDR2)-independent 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 [16-18]. 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 [19]. Various computational approaches have been developed for TE annotation [20], including *de novo* prediction [21-23], motif-based searches [24, 25], and similarity-based methods [26, 27]. Combining multiple strategies has proven most effective in generating accurate and comprehensive TE annotations [19, 23, 28, 29].

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 [30, 31]. Integrating TE polymorphisms

with Genome-Wide Association Studies (GWAS) has provided valuable insights into the functional implications of TEs in various biological processes [32-35]. For instance, transposon insertion polymorphisms (TIPs) have been shown to improve predictions of agronomic traits beyond what is possible with single-nucleotide polymorphism (SNP)-based approaches alone [34, 36-38].

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 [39]. While previous studies have examined TE diversity in the *F. vesca* genome [40-42], 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 [10, 43]. 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 transposable elements in the *Fragaria vesca* reference genome, generating a curated TE library. Using this resource, we analyzed TE insertion and deletion polymorphisms across 210 re-sequenced European *F. vesca* accessions [44], revealing their contribution to genetic diversity and subpopulation structure. We then integrated TE polymorphisms into a genome-wide association framework, identifying significant associations with economically important fruit traits, including volatile compound profiles.

As a proof of concept, we functionally validated the regulatory impact 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 [45]*. Combining gene expression analyses with chromatin conformation profiling at this locus, we demonstrate that MITE-derived siRNAs trigger DNA methylation that reshapes local chromatin organization, thereby modulating *FvAAT1* expression in a ripening-stage-dependent manner.

While genome-wide TE distribution patterns, such as LTR enrichment in pericentromeric regions and MITE localization near genes, have been described in other plant species, the TE landscape of *F. vesca* has remained poorly annotated and functionally unexplored. Our work provides a comprehensive TE resource for this species and establishes a framework to investigate the regulatory potential of MITEs and short inverted repeats, particularly during fruit development. By integrating TE-driven variation into population-genomic and functional analyses, this study offers new insights into how TEs contribute to gene regulation and phenotypic diversity in strawberry, and highlights their potential as targets for breeding and genome-editing strategies aimed at improving 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 [23, 46, 47]. To address these challenges, we combined two recent bioinformatic pipelines, EDTA [48] and DeepTE [49], and applied TEtrimmer [50], a consensus refinement and de-redundancy tool, to generate a semi-curated de novo TE library in *F. vesca* (Figure S1). This allowed us to comprehensively identify and classify TEs in the *F. vesca* v4.0 reference genome [51, 52] into classes, orders, and superfamilies (Table 1, Figure 1A).

The complete set of identified TEs accounted for approximately 30% of the *F. vesca* genome (Figure 1B). In plants, TE content is typically proportional to genome size [53, 54]. 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 [55-57]. The proportion of TEs identified for *F. vesca* (~30%), which has a genome size of ~220 Mb [58], aligns with these observations (Figure 1B).

In contrast to most plant species, where Class I retrotransposons typically dominate, the genome of *F. vesca* exhibits a relatively balanced proportion between Class I and Class II elements. Class I retrotransposons constitute approximately 15% of the total genomic DNA, whereas Class II DNA transposons represent around 13.4% (Table 1). However, domain-based analyses revealed that only a small fraction of Class II elements retain detectable

177 transposase-associated domains, indicating that most DNA transposon copies correspond to
178 non-autonomous or fragmented elements rather than potentially autonomous transposons.
179 Within Class I, LTR retrotransposons remain the most prevalent, comprising most of these
180 elements, with the Gypsy superfamily being particularly abundant (Figure 1A), and non-LTR
181 retrotransposons contribute only a minor fraction (~0.25%) of the genome. The relatively
182 moderate proportion of Class I retrotransposons observed in *F. vesca* might reflect its smaller
183 genome size compared to other plant genomes and could be further influenced by limitations
184 in accurately assembling centromeric regions, which are enriched with these elements.

185 Interestingly, *Fragaria* species exhibit higher diversity among DNA transposons, compared
186 to other plant genomes, highlighting the unique composition of their genomes [40]. After re-
187 classifying DNA transposons into MITES and non-MITEs with DeepTE, almost 13% of
188 identified TEs belong to subclass 1, including both MITEs and other DNA transposons from
189 the CACTA, Harbinger, TcMar, hAT, and Mutator superfamilies. The remaining 0.62%
190 corresponds to subclass 2, specifically Helitrons (Figure 1A). Despite their known difficulty in
191 identification, the proportion of Helitrons identified in the *F. vesca* genome is consistent with
192 reports in other plant genomes, where they typically constitute below 4% of the genome [40,
193 48, 59].

194 We then analyzed the size distribution of TEs for each class based on the annotation
195 generated for *F. vesca*, including both fragmented and intact TEs (Figure 1C, Figure S2, Table
196 S1). Overall, no clear differences were observed in size distributions among the results
197 obtained from EDTA, DeepTE, and the curated library (Figure 1C). The average length of each
198 intact and fragmented TE type is detailed in Figure S2 and Table S1. The similarity in size
199 distributions across libraries suggests that the final curated TE library preserves the structural
200 characteristics of both intact and fragmented elements, supporting the robustness of the library
201 for subsequent analyses.

202 A high proportion of fragmented TEs was identified across the genome. This explains the
203 relatively low mean length values observed for the complete TE library compared to the intact
204 subset (Figure S2), a pattern consistent with the long-term accumulation of mutations and
205 deletions in ancient TEs. While Class I elements generally have lengths below 4 kb, the
206 longest sequences belong to the LTR order (Figure 1D, Figure S2). In contrast, Class II TEs
207 display a broader size range strongly linked to their respective superfamily classification
208 (Figure 1D). The distinct size patterns observed among TE classes suggest differences in
209 replication strategies and selective pressures acting on different TE families.

210 To better understand the structural preservation of these elements, we partitioned the
211 library into two mutually exclusive sets: intact and fragmented elements. A marked contrast in
212 the distribution of TE classes between these groups was observed. While DNA transposons

213 dominated both fractions, they showed a significant increase of 12.3% in the intact library
214 compared to the fragmented set. This enrichment was driven primarily by MITEs, which
215 accounted for 36.6% of the intact elements, a substantial increase of 16.1% relative to their
216 20.5% representation in the fragmented library. Conversely, LTRs were more prevalent among
217 fragmented sequences than in the intact subset, representing a decrease of 9.5%. This
218 highlights a higher rate of structural decay for these larger elements. The abundance of intact
219 MITEs, which are known for being relatively short, can be attributed to their unique structural
220 nature. Unlike autonomous DNA transposons or LTR retrotransposons, MITEs are
221 characterized by their compact sequence length (Figure 1D) and lack of internal coding
222 regions. From a genomic perspective, a shorter sequence presents a significantly smaller
223 "mutational target" for stochastic fragmentation, deletions, or nested TE insertions. While the
224 integrity of an LTR element is easily compromised by a single internal deletion, MITEs are
225 more likely to remain structurally recognizable as "intact" over evolutionary time. Furthermore,
226 because MITEs do not rely on internal Open Reading Frames (ORFs) for their identification
227 they are less sensitive to the sequence decay that typically fragments autonomous elements.
228 Consequently, while larger autonomous elements are rapidly converted into genomic "noise"
229 or fragmented relics, MITEs maintain higher structural stability.
230
231 **Genome-wide distribution and abundance of Transposable Elements**
232 The distribution, abundance, and genome coverage of TEs vary substantially across
233 species and are strongly influenced by subfamily classification. While some TEs tend to
234 accumulate near protein-coding genes or toward the centromeres, others are more evenly
235 distributed across chromosomes. To assess the distribution of TE types in *F. vesca*, circos
236 ideograms were generated for each TE class, with gene density included as a reference
237 (Figure 2).
238 As shown in Figure 2A, regions with higher Class I TE content tend to have lower gene
239 density, a pattern particularly evident for chromosomes Fvb1, Fvb3, and Fvb4. This TE
240 distribution is predominantly driven by LTR transposons, which are the most abundant among
241 Class I. A large fraction of LTR elements appeared located more than 3,000 bp from genes,
242 suggesting a low likelihood of gene association (Figure S3). This distribution pattern of Class
243 I elements has been previously reported in several plant species, including tomato (*Solanum*
244 *lycopersicum*) and rice (*Oryza sativa) [56, 57, 60]*).
245 In contrast, Class II TEs are more concentrated near pericentromeric regions compared to
246 chromosome arms except for Helitrons, which exhibit a more widespread distribution across
247 the genome (Figure 2B). Helitrons positioning varies considerably between species. For
248 example, Helitrons display a more erratic pattern of distribution in potato and rice [57, 61]

249 compared to Arabidopsis, where they are enriched in gene-poor pericentromeric regions, or
250 in maize (*Zea mays*), where they are more abundant in gene-rich regions [62].
251 MITEs exhibit a distinct distribution pattern characterized by a notable concentration in
252 gene-rich regions, particularly in the immediate vicinity of, or within, coding sequences (Figure
253 S3). This pattern aligns with previous reports describing the regulatory role of MITEs in gene
254 expression [56, 57, 63]. Notably, 46% of *F. vesca* genes harbor a MITE within 3,000 bp of their
255 coding sequence, and 15% of these genes contain MITEs within their coding regions. This
256 proximity to genes reinforces the potential role of MITEs in influencing gene expression
257 through epigenetic and chromatin-level mechanisms encouraging us to focus on the study of
258 these elements.
259
260 **Characterization of MITEs and short TE-derived IRs in *F. vesca***
261 MITEs are frequently found in close proximity to genes, where they influence gene
262 expression and act as evolutionary agents contributing to adaptation and developmental
263 processes [64, 65]. This regulatory impact is often mediated by the production of 24-nt small
264 RNAs through the non-canonical RdDM pathway [64, 66, 67]. Such processes trigger localized
265 DNA methylation and induce structural changes in chromatin organization, leading to the
266 formation of short-range chromatin interactions that modulate the transcriptional activity at
267 neighboring genes [64-67]. A similar effect can be triggered by gene-associated TE-derived
268 short inverted repeats (IRs) elements, which often originate from MITEs that have lost
269 distinctive structural elements. The presence of IRs reflects the progressive fragmentation and
270 degeneration of TEs over evolutionary time [7]. In any case, and for the purpose of the
271 regulatory effects studied here, IRs and MITEs operate similarly.
272 Given the mechanistic similarities between IRs and MITEs, we employed *einverted* from
273 the European Molecular Biology Open Software Suite (EMBOSS) [68] to also annotate all IRs
274 in *F. vesca*. We then classified the putative nature of the identified IRs and compared them
275 with the superfamilies previously assigned to MITEs (Figure 3A). IRs can originate from both
276 MITE-type and DNA transposons, Class I elements, or even non-TE sequences [64].
277 Consistently, 84% of the identified IRs were classified as MITE-derived elements, in line with
278 the matching distribution patterns of MITEs and IRs (Figure S4A). Additionally, the overall
279 superfamily distribution of IRs was consistent with that of the annotated MITEs in our TE library
280 (Figure 3A). However, the proportion of IRs from the MITE Harbinger superfamily was lower
281 than that of the annotated MITEs. This suggests that certain IRs may have originated from
282 more ancient MITE elements that have since diverged from their original sequence patterns.
283 Lastly, 16% of the IRs were classified as elements distinct from MITEs ("Other" category in
284 Figure 3C), predominantly belonging to DNA Mutator/CACTA, LTR Gypsy/Copia, and

285 unknown origin (Figure S4B). Considering this classification, we observed that 58% of the
286 identified IRs overlapped by at least 20% of their length with MITEs annotated in our TE library.
287 Therefore, subsequent analyses were performed separately for MITEs and MITE-independent
288 IRs.
289 Most of these identified IRs (4,374, ~80%) were found near or within annotated protein-
290 coding regions (Figure 3B). Of these, 1,047 IRs are situated 500 bp from a gene, 734 between
291 501 bp and 1,000 bp, and 1,493 between 1,001 bp and 3,000 bp (Figure 3B). In total, 14% of
292 *F. vesca* genes harbor an IR within these ranges. Similarly, of the 26,060 MITEs annotated
293 (including fragmented elements), 18,221 MITEs were located near or within annotated protein-
294 coding regions. Specifically, 5,277 MITEs were located within 500 bp of a gene, 3,154 between
295 501 bp and 1,000 bp, and 6,127 between 1,001 and 3,000 bp (Figure 3A). In total, 51% of *F.*
296 *vesca* genes harbor at least one MITE or IR within 3,000 bp of their coding regions. Notably,
297 1,100 IRs (~20%) and 3,665 MITEs (~14%) are fully located within protein-coding genes,
298 primarily in introns (Figure 3B-C), representing approximately 8% of all coding genes in the *F.*
299 *vesca* genome. TE insertions in introns can have a significant regulatory effect. For example,
300 in maize, insertion of a CACTA TE derived element into the *Par1* gene creates an epiallele
301 whose methylation status directly affects gene expression and plant architecture [69].
302 The detected IRs ranged in size from 100 bp to 700 bp (the length limit of our analysis),
303 with an average length of 445 bp. The annotated MITEs exhibited a broader size distribution,
304 ranging from 51 bp to 2,120 bp, with an average length of 261 bp. An overrepresentation of
305 elements between 100 bp and 500 bp (73%) was observed (Figure 1D), values aligning with
306 the typical size range of MITEs in plants [70].
307
308 **Most MITEs and IRs in *F. vesca* are epigenetically active**
309 Upon transcription, MITEs and IRs can be processed into siRNA that trigger non-canonical
310 RdDM, leading to the regulation of neighboring genes and turning these TEs into autonomous
311 regulatory elements. To identify potentially active regulatory MITEs and IRs, we used sRNA
312 and bisulfite (BS) sequencing to determine whether the identified MITEs and IRs produced
313 24-nt siRNAs and triggered DNA methylation—clear indicators of active and potentially
314 regulatory TEs. Three distinct tissues and stages were examined: leaf, and fruit receptacle at
315 both white (immature) and red (ripe) developmental stages (Figure 3D, Figure S5).
316 Our analysis revealed that 66% of MITEs and 77% of IRs produced siRNAs in at least one
317 of the tested tissues (Table S2). Notably, approximately 80% of these elements are located
318 within 3 kb of a gene, which is consistent with previous reports showing that gene-hosted
319 MITEs and IRs are commonly cotranscribed by RNA polymerase II (Pol II) and are
320 epigenetically active [64]. Metagene analysis of 24-nt siRNAs mapping to the MITEs and IRs

321 showed a characteristic two-peak profile, indicating that the siRNAs are produced from the
322 terminal inverted repeats (TIR) through DLC3 processing (Figure 3D). This behavior was
323 particularly pronounced for IRs located within 1,000 bp of genes. In contrast, for MITEs, this
324 pattern persisted even for elements located farther away from genes, although their
325 expression levels decreased with increasing distance (Figure S5A).
326 BS-sequencing metagene analysis revealed that annotated TEs, MITEs, and IRs are
327 methylated under all tested conditions (Figure 3E, Figure S5B). Gene-associated MITEs and
328 IRs, particularly those closest to genes, showed a signature two-peak CHH methylation
329 pattern, consistent with the siRNA production profile and RdDM-dependent methylation.
330 Moreover, siRNA levels and CHH methylation decreased as the distance between MITEs or
331 IRs and the associated gene increased (Figure S5B). This distance-dependent epigenetic
332 activity of MITEs and IRs is consistent with previous reports showing greater transcriptional
333 and regulatory activity of elements closer to a gene [64]. This observation supports the idea
334 that 24-nt siRNAs derived from MITEs and IRs near protein-coding genes initiate a non-
335 canonical RdDM pathway likely dependent on Pol II transcription.
336 In contrast, methylation in the CpG and CHG contexts remained relatively stable regardless
337 of the distance between TEs, MITEs, or IRs and nearby genes (Figure S5B). When these
338 elements were located farther from genes, the surrounding genomic context showed higher
339 CpG and CHG methylation levels, likely reflecting the heterochromatic nature of these regions.
340 Furthermore, CpG and CHG methylation levels were higher in leaves compared to ripe fruits,
341 contrasting with the siRNA production profile and consistent with their independence from
342 siRNA activity (Figure 3D-E, Figure S5B). This contrasts with immature and ripe fruits, which
343 showed higher CHH methylation, suggesting that RdDM activity is required for dynamic
344 regulation of TEs during fruit development and maturation. These differences underscore the
345 distinct epigenetic regulation of transposons across tissues, with stable maintenance of CpG
346 and CHG methylation in somatic tissues and active RdDM-driven CHH methylation in
347 reproductive tissues. This tissue-specific regulatory pattern also reflects the dynamic role of
348 TEs in gene regulation and genome plasticity during fruit ripening and development.
349
350 **MITEs and IRs exhibit context-dependent regulatory signatures at the genome-wide**
351 **level**
352 To investigate the potential regulatory roles of gene-hosted MITEs and IRs in gene
353 expression, we performed a differential accumulation analysis of 24-nt siRNAs derived from
354 these elements as a proxy for MITE differential expression and processing. We compared
355 siRNA profiles between leaves and fruits at two ripening stages (immature and ripe) in
356 *F.vesca*. The analysis revealed a distinct production of 24-nt siRNAs derived from MITEs and

357 IRs across these conditions, considering thresholds of |log2FC| ≥1 and adjusted p-value ≤0.05
358 (Figure 4A). For MITEs, we identified 17 upregulated and 12 downregulated elements in ripe
359 fruits compared to immature fruits, 318 upregulated and 89 downregulated in ripe fruits
360 compared to leaves, and 89 upregulated and 349 downregulated in leaves compared to
361 immature fruits. Similarly, for IRs, 4 elements were upregulated and 3 downregulated in ripe
362 fruits compared to immature fruits, 144 upregulated and 28 downregulated in ripe fruits
363 compared to leaves, and 26 upregulated and 119 downregulated in leaves compared to
364 immature fruits. This approach allowed us to identify MITEs and IRs actively producing siRNAs
365 in a specific sample, which may serve as potential regulators of gene expression in a tissue
366 and stage-specific manner.
367 We then identified genes located within 3,000 bp of the differentially expressed (DE) MITEs
368 and IRs and evaluated whether these genes also exhibited differential expression under the
369 same conditions (Table S3, Table S4). More than 24% of DE MITEs or IRs were located near
370 DE genes (Figure 4B), with this proportion increasing to 50% during fruit ripening. We also
371 found a, a comparable proportion of non-differentially expressed (non-DE) MITEs and IRs
372 located near DE genes across all conditions analyzed. This indicate that the MITE/IR specific
373 regulation is not necessarily coupled to the transcriptional behavior of the neighboring gene.
374 These cases may parallel our previous observations at the *HaWRKY6* locus in sunflower,
375 where tissue-specific gene regulation is mediated by a MITE-triggered chromatin loop, but the
376 regulatory outcome is controlled by differential DNA methylation at a second anchoring region
377 rather than at the MITE itself, whose expression remains largely stable across tissues but is a
378 *sine-qua-non* requirement for the loop formation [66].
379 To further assess whether genes associated with MITEs/IRs display distinct transcriptional
380 behavior, we compared the probability of differential expression among genes located near
381 MITEs/IRs, genes associated with non-MITE TEs, and genes in TE-free regions. Genes
382 located in the vicinity (3 kb) of non-MITE TEs showed a significantly lower probability of being
383 differentially expressed compared to genes in TE-free regions (Fisher's exact test, OR = 0.74,
384 $p = 3.7 \times 10^{-15}$), which is expected given that most TE classes are typically located in
385 repressive genomic contexts characterized by reduced transcriptional activity. In contrast,
386 MITE/IR-proximal genes displayed a significantly higher likelihood of differential expression
387 (OR = 1.13, $p = 3.3 \times 10^{-5}$), indicating that MITE/IR elements are associated with a distinct
388 regulatory landscape. One limitation in this analysis is that the MITE effect caused by
389 triggering short-range chromatin loops not necessarily work as an on/off switch but affect
390 transcriptional dynamics, transcript structure, or regulatory responsiveness without

391 necessarily producing large expression fold-changes detectable at the genome-wide level
392 [67].
393 MITEs and IRs have been also reported to act as enhancers or silencers of gene
394 expression by generating cis-regulatory elements [71, 72] or, by modulating local chromatin
395 organization, altering the gene product without changing overall expression levels (e.g., by
396 triggering alternative splicing or termination) [67]. Their regulatory impact depends on multiple
397 factors, including their genomic position relative to the gene, their interaction with regulatory
398 elements such as transcription factors or epigenetic readers, and the way they reshape
399 chromatin architecture, all of which can vary between conditions.
400 Given the established role of 24-nt siRNAs in mediating RdDM and chromatin remodeling,
401 we next integrated siRNA accumulation with DNA methylation and gene expression data to
402 redefine the identification of MITE/IR-gene pairs with regulatory potential. As a proof-of-
403 concept, we focused on fruit ripening and analyzed publicly available mRNA-seq data from
404 strawberries treated with the DNA methylation inhibitor 5-azacytidine (5-AZA) (Figure 4B,
405 Table S5). Over 35% of DE MITEs/IRs were located near genes that also exhibited significant
406 expression changes upon 5-AZA treatment, with this proportion exceeding 50% during the
407 ripening stage. While this overlap alone does not demonstrate causality, the concordant
408 response of siRNA production, DNA methylation perturbation, and gene expression changes
409 supports the prioritization of specific loci for downstream analysis.
410

411 **A MITE Mutator element modulates FvAAT1 expression through epigenetically driven**
412 **chromatin reorganization**

413 Based on above criteria, we examined individual loci in detail as representative case
414 studies. We focused on elements located near genes with established roles in fruit ripening
415 and associated metabolic or developmental processes, which are highlighted in bold in
416 Supplementary Tables S3 and S4. These case studies are intended to illustrate plausible
417 regulatory configurations supported by epigenetic and transcriptional signals, and are not
418 intended to imply global enrichment or uniform behavior across all analyzed loci.
419 During fruit ripening we found 29 up- or downregulated MITEs and 7 up- or downregulated
420 IRs near genes associated with fruit ripening, metabolic or developmental processes (Figure
421 4A, Figure 5A). Among them, we identified a 416 bp element which corresponds to a MITE
422 Mutator, that showed a 12-fold higher siRNA production at the ripening stage (Table S3). This
423 MITE is located 1,915 bp upstream of the gene *FvAAT1* (FvH4_7g18570) (Figure S6), which
424 encodes a putative ortholog to alcohol acyl transferase SAAT, a key enzyme responsible for
425 the biosynthesis of aroma volatile esters in *F. × ananassa [45]*. Consistent with its functional

426 role, *FvAAT1* displayed substantial upregulation during fruit ripening; at the ripe stage, gene
427 expression was more than 1,700 times higher than at the immature stage (Table S5), aligning
428 with previous findings that underscore its importance in producing characteristic ester-based
429 aroma compounds during ripening [73, 74]. This gene is also upregulated during the turning
430 stage in *F.* × *ananassa* compared to *F. vesca*, emphasizing its importance in commercial fruit
431 quality traits [75]. By contrast, treatment with 5-AZA resulted in a fivefold reduction in *FvAAT1*
432 expression (Table S5), suggesting that the identified epigenetically active MITE can influence
433 its expression.

434 In the same genomic region, we identified another MITE located 1,064 bp downstream of
435 *FvAAT1* and an IR element (IR_9194) located 4,761 bp upstream of *FvAAT1* and 3 bp
436 downstream of FvH4_7g18550, a gene encoding a papain family cysteine protease implicated
437 in protein degradation and defense responses during fruit ripening (Figure S6). This gene
438 showed a marked expression increase (~105-fold) in ripe fruit compared to immature fruit
439 (Table S5), mirroring patterns observed in other plant species in which cysteine proteases
440 have been proposed to regulate ripening processes [76-78].

441 We then evaluated siRNA production and CHH methylation patterns around these
442 elements to gain further insights into potential epigenetic regulation (Figure 5B and 5C). We
443 observed that IR_9194 and especially the MITE Mutator produce 24-nt siRNA with higher
444 levels in fruit ripening, consistent with active elements (Figure 5B, top and middle panel,
445 respectively). This increase in siRNA production in mature fruits coincides with a CHH
446 methylation signal (Figure 5C). Interestingly, the MITE located downstream of FvH4_7g18570
447 displayed the highest methylation levels at its TIRs during immature fruit stage, albeit we did
448 not observe a significant accumulation of siRNA (Figure 5B and 5C, bottom panel). This may
449 indicate that this MITE might be in a permanently silenced state, in contrast to IR_9194 and
450 the MITE Mutator, which remain epigenetically active.

451 Although such MITE appears to be transcriptionally inactive, the increase in DNA
452 methylation at this element may serve as an “anchor point” in conjunction with the more
453 dynamic MITE Mutator element to trigger chromatin looping and control *FvAAT1*, as observed
454 in other systems [66]. To test the hypothesis that the promotion of a chromatin loop between
455 these two elements impact the expression of FvAAT1, we designed an experiment to
456 transiently modulate the epigenetic status of the MITE Mutator. To do so, we infiltrated *F.*
457 *vesca* fruits with a construct either overexpressing the viral silencing suppressor P19, which
458 reduces siRNA activity and thus RdDM, or an additional copy of the MITE Mutator under the
459 35S promoter to increase the production of its specific siRNAs *in trans* and thus induce
460 methylation at this element. We have previously used these strategies to induce or repress
461 MITEs activity [65, 66]. We then measured the expression of *FvAAT1* in infiltrated fruits and

462 mock infiltrated controls. Consistent with the proposed regulatory effect of the MITE Mutator
463 over *FvAAT1,* we observed a correlated expression of the gene when RdDM at the MITE was
464 induced or repressed (Figure 5D). The activation *in trans* of the MITE Mutator correlated with
465 a significant increase in the expression of *FvAAT1* (Figure 5D). Conversely, the epigenetic
466 inactivation of the MITE Mutator, achieved by transiently overexpressing either a MITE Mutator
467 sponge sequence fused to GFP, which sequesters the MITE-derived siRNAs, or the viral
468 silencing suppressor P19, which nonspecifically sequester siRNAs, correlated with a
469 significant repression of the gene *FvAAT1* (Figure 5D). Furthermore, using a chromatin
470 conformation capture (3C) approach, we observed a ripening stage dependent change in
471 chromatin organization, with the MITE Mutator acting most likely as the anchor point (Figure
472 5E). In this sense, a chromatin loop encompassing the entire loci was increasingly detected
473 with the maturation of the fruits, suggesting a positive effect on gene transcription, as
474 previously reported for example during Pol II reclining happening on gene looping events [67].
475 Given the apparent regulatory importance of the MITE Mutator, we assessed its
476 conservation in the orthologous genomic regions of the commercial strawberry species *F.*
477 *chiloensis* and *F.* × *ananassa* through sequence alignment. In *F. chiloensis*, a copy of the
478 element, situated 1,674 bp upstream of the *FvAAT1* homologue was detected, displaying
479 89.4% sequence identity (Figure S7A, left). A similar pattern was observed in *F.* × *ananassa*
480 (Camarosa genome assembly), where a copy of the MITE associated with *FaAAT1* exhibits
481 95.7% sequence identity (Figure S7A, right). Analysis of the predicted secondary structures
482 using RNAfold for the three elements revealed the formation of a long dsRNA hairpin structure,
483 suggesting that they are likely DCL3 substrates and will function similarly to those observed
484 in *F. vesca* (Figure S7B). These findings underscore the evolutionary conservation of this
485 MITE near key ripening-related genes across different *Fragaria* species, potentially reflecting
486 a conserved regulatory function in fruit quality traits.
487 We identified additional IR/MITE elements adjacent to genes with key roles in fruit
488 development and metabolism, suggesting potential regulatory contributions through small
489 RNA production, DNA methylation, and gene expression changes (Table S4 and Table S5).
490 For example, a 439 bp IR element (IR_8440) exhibiting approximately a 4.5-fold increase in
491 expression in red fruit compared to leaf tissue (Table S4) and pronounced CHH methylation
492 in fruit was identified 1,972 bp upstream of the SEPALLATA-like MADS-box gene
493 FvH4_6g46420 (Figure S8A). This IR is adjacent to a MITE Harbinger. A MITE hAT located
494 downstream of FvH4_6g46420 displays relatively constant expression across tissues,
495 suggesting it could serve as a potential “anchoring” region for 3D chromatin configurations.
496 The FvH4_6g46420 gene itself is significantly upregulated in white fruit vs. leaf tissue,
497 consistent with prior studies underscoring the role of its ortholog in *F.* × *ananassa*, *FaMADS9*,

498 which is involved in strawberry fruit development and ripening [79]. Silencing of *FaMADS9* in
499 *F.* × *ananassa* delays or inhibits normal ripening and alters both auxin and ABA metabolism,
500 implicating this MADS-box transcription factor in pleiotropic regulatory networks [80, 81].
501 Furthermore, *FaMADS9* is downregulated under AZA treatment, highlighting the broader
502 interplay between epigenetic modifications and fruit ripening pathways [82].
503 Additional cases reinforce the potential role of IRs and MITEs in regulating nearby gene
504 expression. A MITE hAT (334 bp) is located near FvH4_2g30230, annotated as molybdate
505 transporter 1 (Figure S8B). Molybdenum is an essential trace element in plant nutrition and
506 physiology, and its effects on strawberry flavor and fruit quality are well documented [83]. The
507 upregulation of this gene in ripe fruit and its downregulation in 5-AZA-treated strawberries,
508 along with the distinct expression profiles of the MITE in leaf tissue and in fruits at the white
509 and red ripening stages, together with localized methylation and siRNA mapping, further
510 support a role for IR-associated epigenetic control. Moreover, IR_6046 (448 bp) is located
511 upstream of FvH4_5g23180 (Figure S8C), encoding a probable xyloglucan
512 endotransglucosylase/hydrolase (XTH30), which is critical for cell wall remodeling during fruit
513 softening [84, 85]. Differential expression of both the element and XTH30, alongside distinct
514 siRNA clusters in leaf versus fruit ripening conditions, aligns with a mechanism involving
515 dynamic chromatin organization. Fruit softening is a critical factor affecting the marketability
516 and economic value of fruits; therefore, targeting cell wall-related genes, including XTHs, may
517 offer an effective strategy to delay this process and extend postharvest shelf life.
518 Taken together, these observations underscore the potential of IR and MITE elements to
519 modulate gene activity through epigenetic pathways, thereby influencing fruit maturation and
520 quality traits. An idea strongly supported by the validations made over the *FvAAT1* loci.
521
522 **Discovery of TE Polymorphisms in resequenced *F. vesca* genomes**
523 Due to their inherent mobility, TEs frequently vary between natural populations in copy
524 number and genomic position, producing insertion or deletion polymorphisms [18]. A TE
525 insertion polymorphism (TIP) occurs when a TE is present in an accession but absent from
526 the reference genome, whereas a TE absence polymorphism (TAP) represents a TE found in
527 the reference but missing in other accessions. Because many TEs harbor their own regulatory
528 sequences, they can strongly influence the expression of nearby genes, often more than other
529 types of structural variation [36, 86]. Moreover, our previous results indicate that MITEs and
530 IRs potentially exert a regulatory effect upon insertion into intergenic regions proximal to genes
531 [64]. Historically, genetic diversity analyses have primarily focused on SNPs, while TE
532 insertion/deletion polymorphisms remain an underexplored source of genomic variation [87-
533 90]. TEs activity is often influenced by stress conditions and many causative TE insertions

534 result from recent movements and are not effectively tagged by SNPs, meaning TIPs can be
535 used as direct markers to uncover these adaptive genetic factors [32, 91]. As such, they hold
536 the potential to affect phenotypes and key biotechnological traits in nutritionally and
537 commercially important plant species.

538 We analyzed 210 resequenced *F. vesca* genomes from a European germplasm collection,
539 representing the species' continental diversity [44] (PRJNA1018297), to study natural variation
540 in TEs, MITEs, and IRs using TEPID [90] (Figure S9). A total of 8,184 TE, MITEs and IRs
541 insertions (TIPs) and 11,345 deletions (TAPs) were identified across the population (Table 2).
542 As reported for other plant species [32, 92, 93], the TE landscape was prevalent in rare
543 variants; 84% of TE insertions and 88% of IR insertions exhibited a MAF below 5%. This strong
544 skew toward rare alleles suggests that these insertions are either evolutionarily recent or are
545 being actively purged by purifying selection to maintain genomic integrity. Notably, deletion
546 polymorphisms reached higher frequencies in the population than insertions; for instance,
547 28.7% of TE deletions were classified as common (MAF ≥ 5%) compared to only 15.8% of
548 insertions. This disparity suggests that many deletions represent the loss of ancestral
549 elements present in the reference genome that have not reached fixation in the European
550 germplasm. Among the identified TE insertions, 35% belonged to Class I and 65% to Class II,
551 closely matching the general TE distribution observed in the reference genome (33% and
552 67%, respectively). A similar trend was found for deletion polymorphisms, where Class I
553 elements accounted for 35% and Class II for 65% of the total events Insertion polymorphisms
554 were particularly enriched for LTR/Gypsy, hAT, Mutatorand MITEs elements, suggesting
555 recent bursts of activity that have not yet spread across the population [94] (Figure 6A).
556 Conversely, LTR/Gypsy deletions were notably frequent among common variants, indicating
557 a significant turnover of these retrotransposons in specific lineages.

558 A principal component analysis (PCA) based on the identified TE polymorphisms revealed
559 a clear subpopulation structure (Figure 6B), mirroring the geographical clustering previously
560 reported using SNP-based markers [44]. In the TE-based PCA, the PC1 and 2 captured ~19%
561 of the variance, similar to the SNP-based PCA by Toivainen et al. (2026) [44]. This provides
562 strong confidence in the accuracy and reliability of our detected polymorphisms, as both
563 genomic features capture the same fundamental evolutionary signals. Our TE-based analysis
564 also reflected the differentiation reported in the *F. vesca* volatilome by Urrutia et al. [95], such
565 as the distinct separation of the Icelandic and Baltic Sea clusters. The TE-based PCA also
566 showed a clear resolution for the Iberian and Scandinavian peninsulas, and a separation of
567 Western and Southern European accessions from Northern and Eastern populations. In
568 Scandinavia, this is particularly evident and may be linked to the admixture of western and

569 eastern post-glacial colonization routes [44]. These results suggest that TE-derived structural
570 variation provides a highly sensitive molecular signature for uncovering fine-scale population
571 structure in *F. vesca*.
572 The genome-wide distribution of insertion and deletion polymorphisms in *F. vesca* revealed
573 a clear non-random pattern shaped by genome architecture and selective constraints (Figure
574 6C). Insertions and deletions exhibited partially distinct patterns. Deletions showed a stronger
575 avoidance of exonic regions than insertions, in line with their greater potential to compromise
576 gene integrity (Figure S10A). Across all TEs superfamilies, polymorphisms were strongly
577 depleted within gene bodies, with the most pronounced avoidance observed in exons and
578 untranslated regions (UTRs), consistent with strong purifying selection acting against
579 disruptive structural variation in functional elements (Figure 6C, Figure S10B). Introns showed
580 a comparatively milder depletion, indicating a higher tolerance for structural variation relative
581 to coding and regulatory regions. In contrast, polymorphisms were consistently enriched in
582 gene-flanking regions, where chromatin is typically less condensed, likely facilitating access
583 for the TE mobilization machinery (Figure 6C, Figure S10). Similar patterns have been
584 observed in other species, where TE insertions tend to occur in gene-rich regions, although
585 variations can occur at both the TE family and species levels [36, 94, 96]. Overall, more than
586 half of all identified polymorphisms were located within genes or their immediate flanking
587 sequences (up to 1,000 bp upstream/downstream). Specifically, MITEs and IRs showed a
588 preferential accumulation within promoter regions (Figure 6C, Figure S10B). This high
589 frequency in upstream regions and to some extent downstream, suggests a significant
590 potential for these elements to influence gene expression, as observed in the results described
591 above or in other crop species, like *Brassica rapa* and *Solanum lycopersicum* [36, 37].
592 Taken together, these results suggest that the landscape of TE polymorphisms in *F. vesca*
593 is characterized by a high turnover of elements within regulatory proximity to genes. The
594 robust population structure resolved by these variants underscores the role of TEs as dynamic
595 footprints of post-glacial colonization. The high frequency of rare variants suggests that most
596 transposable element polymorphisms (TEPs) sare subjected to purifying selection or occurred
597 recently in certain lineages. Moreover, these TEPs occur preferentially in promoter regions.
598 Collectively, this represents a significant and underexplored reservoir of genetic diversity with
599 the potential to impact complex phenotypic traits in European woodland strawberry
600 germplasm.
601
602 **Linking TE polymorphisms to phenotypic traits in woodland strawberry fruits**
603 Several phenotypic traits, including fruit dimensions (length, width, volume) and the
604 concentrations of 99 volatile secondary metabolites (the volatilome), were previously

605 measured in a collection of 116 and 154 wild *F. vesca* accessions across two independent
606 harvest seasons (H16 and H17) [95]. Since the genomes of these accessions were included
607 in our polymorphism analysis, we leveraged the detected TEPs to perform a genome-wide
608 association study (GWAS) to investigate the impact of TE and IR natural diversity in those
609 strawberry fruit traits.

610 Several phenotypic traits, including fruit dimensions (length, width, volume) and the
611 concentrations of 99 volatile secondary metabolites (the volatilome), were previously
612 measured in a collection of 116 and 154 wild *F. vesca* accessions across two independent
613 harvest seasons (H16 and H17) [95]. Since the genomes of these accessions were included
614 in our polymorphism analysis, we leveraged the detected TEPs to perform a genome-wide
615 association study (GWAS) to investigate the impact of TE and IR natural diversity in those
616 strawberry fruit traits.

617 For the volatilome dataset, 68 distinct TEPs were significantly associated (5% Bonferroni
618 threshold) with 40 VOCs across harvests H16 and H17 resulting in 78 significant trait–
619 polymorphism associations (Table S7). None of these associations corresponded to TEP
620 directly impacting the coding sequence of genes known to be involved in strawberry VOC
621 metabolism; this is consistent with our observation that only a very small fraction of these
622 variants occur in exonic regions and tend to occur more frequently in intergenic regions
623 proximal to genes where they could exert a regulatory effect (Figure 6C). This is also
624 suggested in the results in this work presented, in which we observed the regulatory effect of
625 different MITEs on the expression of *FvAAT1*.

626 Considering our results, we next focused on statistically significant associations that exhibit
627 potential biological relevance regulating the genes in their vicinity based on current knowledge,
628 specially focusing on MITE or IR polymorphisms. For instance, a MITE Mutator insertion was
629 associated with 6-methyl-5-hepten-2-one (sulcatone) on chromosome Fvb3 ($P = 5.6512 \times 10^{-}$
630 ^{6}; PVE = 9.20%) (Figure 7A-B, Figure S12A). Sulcatone is a common apocarotenoid that
631 originates from lycopene cleavage [97]. Apocarotenoids contribute to the aromas of different
632 flowers and fruits and are highly valued by the flavor and fragrance industry [97]. Unsaturated
633 methyl ketones, such as sulcatone, have been proposed to act as highly reactive VOCs with
634 defensive and pesticidal properties and have been associated with unpleasant taste in fruits
635 affected by powdery mildew disease [98-100]. Accessions carrying this insertion (n = 34)
636 showed lower 6-methyl-5-hepten-2-one levels in harvest H17 (Figure 7C). Interestingly, the
637 insertion is located in a gene-dense genomic region (FvH4_3g02700, FvH4_3g02710,
638 FvH4_3g02720, FvH4_3g02730 and FvH4_3g02731 genes) enriched in stress- and ROS-
639 related signalling components and redox enzymes, including LEA/NDR1-like (LEA-2) proteins,
640 NDR1 a secretory class III peroxidase, a cysteine-rich receptor-like kinase (CRK-like), and an

641 LpxD-like N-acyltransferase. Class III peroxidases act as major apoplastic regulators of H_2O_2
642 homeostasis and oxidative reactions, shaping the local redox environment that feeds into
643 downstream oxidative volatile formation [101]. In parallel, NDR1 is a central component of
644 plant defence signalling, and CRK family kinases are widely implicated in ROS-associated
645 signal transduction during stress responses. The presence of a MITE element in this gene
646 module could impact the regulation of these genes, ultimately affecting the response to
647 oxidative stress and indirectly influencing the oxidative flux required for carotenoid degradation
648 and the sulcanone levels.
649 A Harbinger Class II TE deletion was associated with lower levels of the 6-C aldehyde
650 green leaf volatile (GLV) (E)-2-hexenal in H17 (P = 9.51 x 10^{-8}; PVE = 2%) (Figure 7D-F). GLV
651 are VOCs emitted in response to wounding or plant damage and are responsible for the
652 characteristic "green" aroma of plants (reviewed in [102]). The deletion is located on
653 chromosome Fvb1, 13 kb upstream of a gene annotated as 2-ALKENAL REDUCTASE
654 (FvH4_1g00330). Alkenal reductases facilitate the reduction of double-bonds in aldehydes,
655 including short-chain aldehydes such as (E)-2-hexenal [103, 104]. This gene, or any of its
656 eight downstream paralogs, represents a potential candidate for the conversion of (E)-2-
657 hexenal to hexanal in strawberry. Notably, this locus has not been identified in previous QTL
658 analysis in *F. vesca* or cultivated strawberry [105, 106], highlighting the potential of TEPs-
659 GWAS to identify novel QTLs.
660 Other associations further suggest linkage between the TEP and the underlying candidate
661 gene. A primary example is methyl 2-aminobenzoate (methyl anthranilate), a precursor of
662 tryptophan involved in primary metabolism and associated with a grape-like aroma in fruits
663 [107]. We identified a significant association with the deletion of a MITE Mutator on
664 chromosome Fvb5 (P = 1 x 10^{-14}; PEV = 0) (Figure S13A-C). Interestingly, although this variant
665 appears to explain only a marginal amount of phenotypic variance, this locus has been
666 consistently associated with methyl anthranilate levels in octoploid strawberry [108]. Barbey
667 et al. described a QTL at the beginning of chr5-4, the same region where the TEP is located.
668 This site is 380 kb from an L-tryptophan—pyruvate aminotransferase (TAA1) gene
669 (FvH4_5g05880), which utilizes L-tryptophan for auxin biosynthesis; this gene may be related
670 to varying levels of the intermediate methyl anthranilate.
671 Our analysis also identified a MITE Harbinger deletion on chromosome Fvb3 that was
672 associated with γ-decalactone levels across both harvest's seasons (P = 4.2911 x 10^{-6} in H16
673 and P = 1.4523 x 10^{-6} in H17) (Figure S13D-F, Figure S12B). This variant appears to explain
674 only a marginal amount of phenotypic variance in both harvests. Accessions carrying this
675 deletion (n = 87 and 115 in H16 and H17, respectively number) showed higher levels of γ-
676 decalactone, a key compound previously identified as one was one of the most variable in the

677 *F. vesca* volatilome analysis by Urrutia et al. [95]. In *F.* × *ananassa*, γ-decalactone is the most
678 abundant lactone [79], reaching maximum levels in fully ripe fruits [79] and is positively
679 correlated with consumer perception of favorable organoleptic traits [95, 109]. The deletion
680 falls within an intron of the gene FvH4_3g30910, annotated as a replication factor-A C-terminal
681 domain protein, which shows no evident functional connection to γ-decalactone. Notably,
682 however, the deletion lies ~5 kb downstream of FvH4_3g30901, a gene encoding a serine
683 carboxypeptidase-like (SCPL) enzyme. SCPL enzymes belong to the α/β-hydrolase
684 superfamily, functioning as acyltransferases in plant secondary metabolism which include
685 esterases and acyltransferases that act on diverse lipid substrates [110, 111]. Given the
686 functional plasticity of these enzymes, it is likely that many SCPL members involved in fruit
687 specialized metabolism are yet to be characterized. In this context, structural variation
688 affecting chromatin state or *cis*-regulation near the SCPL gene could indirectly contribute to
689 quantitative differences in γ-decalactone.
690 Taken together, these results suggest that natural polymorphisms in transposable
691 elements, particularly MITEs and IRs, are linked to variation in strawberry fruit volatilome
692 composition, primarily through effects in gene-proximal regions. While these associations do
693 not imply direct causality, they reveal candidate regulatory loci whose effects are shaped by
694 local genomic and chromatin context.
695
696 **DISCUSSION**
697 Our study provides a comprehensive analysis of TEs in *Fragaria vesca*, highlighting their
698 role in genome composition, organization, and potential functional impact. By systematically
699 annotating TEs and analyzing their polymorphisms across 210 *F. vesca* accessions, we have
700 gained valuable insights into their evolutionary significance and their potential contributions to
701 gene regulation and phenotypic diversity.
702 Our TE annotation revealed that TEs occupy ~30% of the *F. vesca* genome, a proportion
703 consistent with expectations for plant species of similar genome size and slightly higher than
704 the 22% reported in the initial 2011 *F. vesca* genome assembly [41]. Unlike most plant
705 genomes, where LTR retrotransposons are dominant [54] *F. vesca* exhibits a comparable
706 proportion of Class II DNA transposons, with MITEs representing a substantial fraction of intact
707 elements. This enrichment is likely explained by their compact structure and lack of coding
708 regions, which reduce their susceptibility to fragmentation over evolutionary time. In contrast,
709 LTR retrotransposons, particularly the larger elements, appear more frequently fragmented,
710 consistent with higher rates of structural decay. The genome-wide distribution of TEs further
711 reveals distinct patterns among classes: LTRs are enriched in gene-poor, pericentromeric
712 regions, although their relatively low abundance in *F. vesca* may reflect incomplete assembly

713 of centromeric regions, whereas MITEs, IRs, and other DNA transposons are concentrated
714 near or within coding sequences, suggesting a potential role in the regulation of adjacent
715 genes. These observations underscore the unique composition of the *F. vesca* genome and
716 provide a foundation for exploring how TE-mediated regulatory variation contributes to
717 phenotypic diversity in strawberry. One of the most significant findings of our study is the
718 widespread production of small RNAs (siRNAs) from MITEs and IRs in *F. vesca*, supporting
719 their potential regulatory role in gene expression. On average, 71% of these non-autonomous
720 elements produced 24-nt siRNAs in at least one tissue type studied, with a characteristic two-
721 peak siRNA profile indicative of processing through the RdDM pathway. The correlation
722 observed between siRNA production and CHH methylation, particularly in reproductive
723 tissues, underscores the functional relevance of these elements in tissue-specific gene
724 regulation. In contrast, CpG and CHG methylation patterns remained relatively stable,
725 supporting the role of siRNA-driven CHH methylation as a key mechanism controlling TE
726 activity and gene expression in a dynamic and reversible manner. These findings highlight the
727 functional significance of MITEs and IRs as autonomous regulatory elements that modulate
728 gene expression through small RNA-mediated epigenetic modifications. These observations
729 position MITEs and IRs not merely as silencing targets, but as active components of regulatory
730 networks capable of modulating local chromatin states.
731 While genome-wide analyses revealed that differentially expressed MITEs and IRs are not
732 significantly more likely than non-DE elements to be located near differentially expressed
733 genes, this finding underscores an important conceptual point: TE proximity alone is
734 insufficient to infer regulatory function. This is not entirely surprising as we previously found
735 that one of the most important features of MITE affecting gene expression is not their
736 sequence but their capacity to trigger chromatin loops, which inevitably require an additional
737 anchor point. In a previous study we showed that, if a MITE is expressed, its regulatory
738 function can be modulated by alterations (methylation, accessibility, etc.) on the second
739 anchor point [66]. The lack of global enrichment of differentially expressed genes near DE
740 MITEs or IRs during ripening further indicates that these elements do not act alone. Instead,
741 their regulatory potential appears to emerge from the convergence of multiple features,
742 including tissue-specific siRNA production, local methylation dynamics, transcriptional
743 responsiveness, and the genomic neighborhood.
744 This concept is exemplified by the detailed analysis of selected loci associated with fruit
745 ripening. At the *FvAAT1* locus, a MITE Mutator element displays coordinated increases in
746 siRNA accumulation and a clear CHH methylation signature during ripening, coinciding with
747 dramatic upregulation of the adjacent gene encoding an alcohol acyltransferase, a key
748 enzyme responsible for volatile ester biosynthesis in strawberry [45, 112, 113]. The

749 upregulation of *FvAAT1* during ripening, coupled with siRNA production from the adjacent
750 MITE, supports a model in which this TE influences gene expression through methylation-
751 dependent chromatin remodeling. Further validation using 5-AZA demethylation assays
752 confirmed that *FaAAT1* expression is methylation-dependent, highlighting the direct regulatory
753 effect of MITEs in *F. vesca* fruit ripening. The conservation of this particular MITE in *F.*
754 *chiloensis* and *F.* × *ananassa*, with high sequence identity and structural similarities, suggests
755 that it plays a conserved role in fruit aroma regulation across strawberry species. This
756 discovery adds to the growing evidence that TEs contribute to the evolution of crop traits by
757 introducing regulatory variation that affects key metabolic pathways [7, 36, 114, 115]. It also
758 positions TEs, especially gene-associated MITEs and short IRs, as promising targets for
759 genome editing. The CRISPR/Cas9-mediated removal of a MITE could significantly alter the
760 expression of adjacent genes and the associated phenotypic traits. Genome editing of MITEs
761 also offers the advantage that these elements are non-coding and naturally variable in
762 populations, potentially simplifying regulatory approval.
763 In addition to their regulatory potential, TE polymorphisms provide an important source of
764 genetic diversity that can influence trait variation [37]. Our GWAS analyses link TE
765 polymorphisms to volatile composition, demonstrating the phenotypic relevance of mobile
766 elements in *F. vesca*. Notably, some of the identified TEPs correspond to previously
767 unreported loci affecting key flavor compounds, such as sulcatone, γ-decalactone, and methyl
768 anthranilate, highlighting the potential of TEP-focused GWAS to uncover novel regulatory loci
769 that are missed by traditional SNP-based approaches. While these associations do not
770 establish direct causality, they provide candidate regulatory elements for functional validation
771 and suggest that integrating TE polymorphisms into breeding programs could expand the
772 spectrum of allelic variation available for trait improvement. Our findings demonstrate that TE-
773 driven variation significantly contributes to the natural diversity of fruit metabolic profiles, with
774 potential implications for breeding strategies aimed at improving fruit aroma and flavor.
775 Furthermore, understanding how TEs modulate gene expression and metabolite production
776 at the molecular level provides valuable insights into the complex regulatory networks
777 underlying crop traits. The high degree of TE polymorphism observed in *F. vesca*, combined
778 with the functional relevance of associated traits, suggests that TEs represent an untapped
779 source of genetic variation for crop improvement.
780
781 **CONCLUSIONS**
782 Our findings demonstrate that, as in other plants, TEs in strawberries are not merely
783 passive genomic elements but rather key contributors to genome function, gene regulation,
784 and phenotypic variation. By integrating TE annotation, small RNA profiling, methylation data,

785 gene expression analyses, and TEP-based GWAS, we identify specific TE–gene
786 configurations with plausible regulatory roles in fruit development and quality traits. Although
787 TE proximity alone does not predict regulatory function at the genome-wide level, selected loci
788 illustrate how epigenetically active MITEs and IRs can modulate gene expression in
789 biologically relevant contexts, such as fruit ripening.
790 This identification of TE-derived regulatory elements and polymorphisms associated with
791 fruit traits provides a valuable resource for future breeding efforts aimed at improving
792 strawberry quality. In particular, the regulatory roles of MITEs and IRs in fruit ripening suggest
793 that targeted epigenetic modifications could be leveraged to optimize fruit quality. The use of
794 genome-editing tools such as CRISPR/Cas9 to precisely manipulate TE-derived regulatory
795 regions may enable the fine-tuning of gene expression without introducing foreign DNA,
796 making it a promising strategy for crop improvement.
797 Additionally, the conservation of key regulatory MITEs in commercial strawberry species
798 suggests that similar TE-driven mechanisms may underlie important agronomic traits in
799 cultivated varieties. Further comparative genomic studies across *Fragaria* species could help
800 elucidate the broader evolutionary impact of TEs and their role in shaping fruit diversity.
801
802 **METHODS**
803 **Transposable element identification and classification**
804 The genome sequence of *Fragaria vesca* and annotation files (v4.0.a2) were downloaded from
805 the Phytozome web portal at https://phytozome-next.jgi.doe.gov/ [116]. TEs were detected
806 and preliminarily classified using the EDTA v2.1.0 [48], executed as follows: `EDTA.pl --`
807 `genome $genome --exclude $annotation_bed --anno 1`. where `$genome` is the *F.*
808 *vesca* reference genome, and `$annotation_bed` is the assembly annotation converted to
809 BED format from the original GFF3 file using the *gff2bed* tool from BEDOPS v2.4.41 [117].
810 EDTA produces a filtered non-redundant TE library for annotation of structurally intact and
811 fragmented elements. TEs are classified into several non-overlapping categories, including
812 LTR (referring to LTR retrotransposons), non-LTR (including SINEs and LINEs), TIR (referring
813 to DNA transposons with TIRs, including MITEs), Helitron, and non-TE repeat sequences.
814 TEs identified by EDTA were reclassified with DeepTE [49] that uses convolutional neural
815 networks and detects domains inside TEs to correct false classification. Within Class II,
816 DeepTE allows for the subclassification of MITEs and non-MITEs (DNA TEs). DeepTE was
817 run with the following command: `DeepTE.py -d /path/intermed_files -o`
818 `/path/results -i $TE_lib -sp P -m_dir /path/Plants_model`. Where `$TE_lib`
819 is the FASTA file with the TE sequences detected by EDTA, the parameter `-sp P` specifies

820 a plant organism, and the `-m_dir` flag indicates the path to the training model. A custom
821 bash script was then used to merge the GFF3 EDTA output files with the DeepTE classification
822 (TE order and superfamily). For TEs categorized as “unknown” by DeepTE, the original EDTA
823 classification was preserved, translating it into the classification used in this work (Figure 1B).
824 The consensus TE library was then curated using TEtrimmer [118], a tool designed to
825 automate the manual curation of TEs, which performs automated consensus sequence
826 curation, clusters related elements, and removes duplicates. This step markedly improves TE
827 library quality and facilitates more accurate repeat masking and classification. TEtrimmer was
828 run with default parameters and the `--classify_all` option. Then, the consensus library
829 file after de-duplication step was used as input for RepeatMasker to finally annotate the
830 curated TE library.
831 Two GFF3 files were generated with the genomic coordinates and assigned TE
832 class/superfamily. One includes all TEs (including fragments), while the other contains only
833 structurally intact TEs, that is, those containing all domains corresponding to their superfamily.
834 TE-derived IRs were classified by grouping sequences based on their sequence identity, using
835 the CD-HIT-EST module of CD-HIT v4.8.1 [119], as previously reported for *Arabidopsis*
836 *thaliana* [63], running the following command: `cd-hit-est -i $fa -o $output -c 0.8`
837 `-A 0.8 -T 2 -n 5`. A minimum sequence identity threshold of 80% (`-c`), a minimum
838 alignment coverage of 80% (`-A`), and a word size of 5 (`-n`) were set. Next, DeepTE was run
839 as previously described for TE annotation, using the representative sequences from each
840 cluster to classify the IRs. When IRs overlapped with annotated TEs but were not initially
841 classified as Class II, they were reclassified based on the overlapping element's category
842 using a custom bash script. The presence of transposase-associated domains in Class II
843 elements was evaluated using the DeepTE_domains.py module, which detects conserved
844 protein domains in TE sequences, applied to both Class II consensus sequences and genome-
845 wide annotated copies.
846
847 **IR and MITE analysis**
848 Inverted repeats were identified in the *F. vesca* v4 genome using *einverted* from the EMBOSS
849 program suite v6.6.0.0 [68]. The following parameters were used: maximum repeat of 700 bp,
850 a gap penalty of 8, a minimum score threshold of 150, a match score of 3, and a mismatch
851 score of −4. A custom Perl script was used to parse the output and generate a GFF3 file with
852 the IR annotation. To classify IRs as MITE-independent elements, we used bedtools intersect

853 v2.26.0 [59], discarding elements that exhibited more than 20% overlap between MITEs and
854 IRs.
855 To assess the conservation of the MITE Mutator, the Tripal MegaSearch tool from the Genome
856 Database for Rosaceae (GDR) website (https://www.rosaceae.org/) was first used to identify
857 orthologous genes in the strawberry species *Fragaria chiloensis* (KIB CAS genome assembly)
858 and *Fragaria* × *ananassa* (Camarosa genome assembly). A 3,000 bp sequence upstream of
859 each gene was retrieved from JBrowse (GDR) for the respective genomes and locally aligned
860 with the *F. vesca* MITE Mutator sequence using EMBOSS Water
861 (https://www.ebi.ac.uk/jdispatcher/psa/emboss_water), which employs the Smith-Waterman
862 algorithm for pairwise alignment.
863 The predicted secondary structures with minimum free energy for each MITE were generated
864 using the RNAfold web server (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi)
865 with default parameters.
866
867 **Plant material and growth conditions**
868 Plants from the *F. vesca* accession 'Reine des Vallées' used in this study were grown and
869 maintained in a greenhouse under natural sunlight conditions (IHSM "La Mayora", Málaga,
870 Spain), using a mixture of universal substrate and river sand (3:1 ratio by volume). Fruits at
871 the immature (white) and ripe (red) stages were collected for small RNA-seq (sRNA-seq) and
872 RNA-seq transcriptome analyses. Strawberry fruits infiltrations were performed as previously
873 described [120]. MITE-Mutator siRNA sponge was created by fussing half of its sequence
874 (only one of the TIR to prevent dsRNA hairpin formation) with the GFP cloning sequence.
875
876 **sRNA-seq**
877 RNA was extracted as previously described [121] from fruit receptacles at two developmental
878 stages, immature (white) and ripe (red). In detail, for each stage, two biological replicates were
879 prepared; each biological replicate was composed of 30 receptacles after the removal of the
880 achenes. RNA quality of the samples was measured on a Bioanalyzer 2100 (Agilent
881 Technologies, Santa Clara, CA). RNA purity was checked using the NanoPhotometer®
882 spectrophotometer (IMPLEN, CA, USA). RNA integrity was analyzed by lab-on-a-chip
883 electrophoresis using the RNA Nano 6000 Assay Kit with the Agilent Bioanalyzer 2100 system
884 (Agilent Technologies, CA, USA). Quantification of samples was performed with Qubit® RNA
885 Assay Kit in the Qubit® 3.0 Flurometer (Life Technologies, CA, USA). Subsequently, 1 µg of
886 total RNA per sample was used to generate small RNA libraries using the Truseq® Small RNA
887 Library Prep Kit according to manufacturer's instructions (Illumina Inc., San Diego, CA, USA).
888 Library pool was sequenced on the Illumina Nextseq550 instrument at the University of Málaga

889 (Spain), with 1 × 75 cycles according to the manufacturer's protocol. Publicly available sRNA-
890 Seq reads from *F. vesca* leaves were retrieved from GEO under accession number
891 GSM1513956 and analyzed together with immature and ripe reads.
892 The small RNA reads were first cut to remove 3' adapters using cutadapt v4.9 and their quality
893 checked using FastQC v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
894 and MultiQC v1.25.1 [122]. They were then mapped with ShortStack v4.1.1 [123] to *F. vesca*
895 v4 genome using the following parameters: `--locifile` (with MITE or IR annotation files), `--nohp`
896 (prevents search for microRNAs) and `--make_bigwigs`. The DESeq2 Bioconductor
897 package ([124], R Core Team 2022) was used to normalize the total raw read counts from
898 ShortStack for differential expression analysis. MITEs/IRs with 10 or more 24-nucleotide reads
899 across their entire region in all samples were considered to have potential siRNA production.
900 After count normalization, the expression of IR/MITE-derived siRNA was compared between
901 leaf tissue vs. immature stage, ripe vs. immature stages, and ripe stage vs. leaf tissue, using
902 the negative binomial test. Differentially expressed MITEs/IRs were determined by adjusting
903 p-values with the Benjamini–Hochberg false discovery rate, and those with an absolute log2
904 fold change ≥1 and an adjusted p-value ≤0.05 were defined as differentially expressed
905 elements.
906
907 **RNA-seq**
908 Total RNA was isolated as described previously [120] using the same protocol applied to the
909 fruit receptacles used for sRNA-seq analysis. For the RNA-seq experiment, however, three
910 independent biological replicates were prepared per developmental stage. After validating the
911 quality and quantity of the total RNA using Bioanalyzer 2100 (Agilent Technologies Santa
912 Clara, CA, USA) and Qubit 2.0, libraries were prepared using Illumina's TruSeq RNA Sample
913 Prep Kit v2. Pools of libraries were sequenced by paired-end sequencing (100x2) in an
914 Illumina HiSeq 2500 sequencer by Sistemas Genómicos (Valencia, Spain). Also,
915 transcriptome data for leaf tissue (GEO: GSM3243286 and GSM3243287), fruits treated with
916 5-AZA and its control (GEO: GSE151534) produced by previous studies were used [82, 125].
917 The analysis started by quality trimming and filtering the raw reads with Trimmomatic v0.39
918 [126], the reads were then aligned to the *F. vesca* v4 reference genome using STAR v2.7.11b

919 [127] which was guided by the gene and exon annotation from *F. vesca* v4.0.a2. *Sort* and
920 *bamCoverage* tools from Samtools v1.21 [128] were then used to sort BAM files by genomic
921 coordinates and generate normalized coverage tracks for visualization and further analysis,
922 respectively. Read quality before and after trimming was analyzed with FastQC v0.12.1
923 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and, together with mapping
924 efficiency, they were summarized with MultiQC v1.25.1 [122]. Read counts on each gene were
925 then calculated with featureCounts v2.0.8 [129]. This pipeline was run with the aid of the
926 Snakemake workflow engine [130]. Differential expression analysis was performed by the
927 DESeq2 package [124], filtering out genes with counts below 10 in all samples. Genes were
928 then annotated using the information (BLAST, InterPro, GO term, GO accession) provided by
929 the GDR database (https://www.rosaceae.org/).
930
931 **BS-seq**
932 Publicly available Bisulfite-Seq data [125] from *F. vesca* leaf tissue, immature receptacle and
933 ripe receptacle were retrieved from GEO under accession numbers GSM3243379,
934 GSM3243380 and GSM3243381, respectively. Although it is not specified whether tissues
935 were collected from the same individual plants, the observed methylation trends, including
936 global hypomethylation and consistent DMRs across replicates, support their reliability for
937 integrative methylome analysis. The EpiDiverse WGBS v1.0 pipeline for bisulfite read
938 trimming, mapping and methylation calling was implemented
939 (https://github.com/EpiDiverse/wgbs), which was specifically designed for non-model plant
940 species [131]. For each sample, bedGraph files corresponding to each sequence context
941 (CpG, CHG, CHH) were obtained. Then *bedGraphToBigWig* v.2.10 from ENCONDE
942 (https://www.encodeproject.org/software/bedgraphtobigwig/) was used to generate coverage
943 tracks for visualization and plotting.
944
945 **3C assay and RT–qPCR**
946 The 3C assay was conducted following a published procedure [66]. Approximately 2 g of tissue
947 was collected per sample and crosslinked with 1% formaldehyde. Crosslinking was stopped
948 by adding glycine to a final concentration of 0.125 M, after which nuclei were isolated. To
949 detect FvAAT1 loops DNA was digested with XhoI (Thermo Fisher Scientific). DNA fragments
950 were ligated with 25 U of high concentrated T4 DNA ligase (Thermo Fisher Scientific) at 22
951 °C for 5 h in a 2-ml volume. Subsequent steps included reverse crosslinking, proteinase K

952 treatment (Qiagen) and DNA purification with the phenol–chloroform. To measure interaction
953 frequency, qPCR was performed, and results were analyzed with the 2−ΔΔCt formula using
954 ACT2 as a control. The final qPCR products were subjected to agarose gel electrophoresis
955 (2%) and bands corresponding to the expected size were purified and sequenced using the
956 Sanger method to confirm ligation products in the case of 3C experiments. For RT–qPCR, 500
957 ng of total RNA underwent RT using the RevertAid RT kit (Thermo Fisher Scientific) after
958 treatment with DNase (Thermo Fisher Scientific). Subsequently, qPCR was conducted using
959 SYBR green (Thermo Fisher Scientific Maxima SYBR Green qPCR Master Mix (2×)). Relative
960 expression levels were determined using the 2−ΔΔCt method with *FvCHP1* as the
961 housekeeping gene control.
962
963 **Identification of TE Polymorphisms**
964 A set of 217 resequenced woodland strawberry genomes from a European germplasm
965 collection (PRJNA1018297), representing the species' continental diversity, was used for the
966 polymorphism study [44]. Detailed information on each accession's geographic origin, read
967 depth, and chromosomal coverage can be found in Table S6. Since the study of TE
968 polymorphisms requires a coverage over 15X and no less than 10X [90], resequencing data
969 was analyzed using SAMtools *coverage* v1.17 [132]. The results were processed using
970 custom bash and R scripts. Next, the accessions were filtered based on a mean per-
971 chromosome coverage exceeding 10X. Consequently, seven of the 217 available accessions
972 (Alta, FIN55, NOR2, NOR-P1-10, RUS3, UK3, and UK9) were excluded from subsequent
973 analyses, resulting in a final set of 210 accessions.
974 TEPID v0.10 [90] was employed for the detection of polymorphisms, using the curated non-
975 redundant TE library, and the IR annotation. The three stages of the TEPID algorithm were
976 executed using the commands listed below:
977 <u>Initial Mapping Stage:</u>
978 `tepid-map -x $bowtie_index -y $yaha_index -p $threads -s $insert_size`
979 `-n $reseqid -1 $r1 -2 $r2 -z`
980 Where `$bowtie_index` and `$yaha_index` indicate the paths to the reference indices
981 generated by each program; `$insert_size` refers to the average insert size of the
982 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 `$split_BAM` are the Bowtie2 [133] and YAHA [134] alignment files resulting from the mapping stage, respectively; and `$TE_bed` is the BED file with TE annotations in the format specified by TEPID.

After the discovery stage, an intermediate step involved merging the TE insertions and deletions detected in the population into a single BED file using the TEPID scripts `merge_insertions.py` and `merge_deletions.py`. Briefly, insertions of the same TE were merged if their coordinates were within ±100 bp, generating a list of accessions containing that insertion. Similarly, TE deletions were merged to create a file representing all deletions detected in the population, along with the list of accessions containing each deletion.

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.

This step, designed to reduce false-negative calls, re-evaluates genomic regions with less stringent settings where a TE variant was identified in at least one sample of the population but not in the focal sample. It produces two types of BED files for each accession: (i) second_pass, referring to polymorphisms not detected in the discovery step but rescued after re-evaluating the region, and (ii) ambiguous, referring to TE presence/absence calls that could not be reliably determined during the refinement step. For each accession, second_pass insertions and deletions were combined with the variants discovered in step 2. The resulting final insertion and deletion files for each accession were then merged using the scripts `merge_insertions.py` and `merge_deletions.py`.

The resulting BED files were processed using custom R scripts.

1017 To determine the genomic distribution of TE polymorphisms (Figure 6C and Figure S10) we
1018 developed a custom script using R packages GenomicRanges and GenomicFeatures [135].
1019 Insertions and deletions were mapped against the *Fragaria vesca* v4.0.a2 reference
1020 annotation. Genomic features were defined hierarchically to resolve overlapping annotations:
1021 regions 1 kb upstream of the transcription start site (TSS) were classified as promoters
1022 ("Upstream"), followed by 1 kb downstream of the transcription termination site (TTS). Within
1023 gene bodies, determining priority was given to Exons, followed by UTRs (5' and 3'), and finally
1024 Introns. Polymorphisms falling outside these regions were classified as distal intergenic (>1
1025 kb or <-1 kb). To assess whether the observed distribution deviated from random expectation,
1026 we calculated the "Genomic fraction" occupied by each feature type. This was calculated by
1027 collapsing the reference genome into non-overlapping unique base pairs (unstranded) to
1028 represent the physical target size available for mutation. Enrichment or depletion was
1029 quantified using the Log_2 ratio of the Observed frequency to the Expected frequency
1030 (percentage of the genome occupied by a specific feature).
1031
1032 **Genome-Wide Association Study**
1033 GWAS was conducted on the 4395 and 4340 polymorphisms present in 116 and 154
1034 accessions for harvests H16 and H17 [91], respectively, with a MAF ≥ 5%. The association
1035 analyses were performed by GAPIT (version 3.5) [136] using a Multi-Locus Mixed Model
1036 (MLMM) [137]. The model included the three first principal components derived from all the
1037 polymorphisms used for each group of accessions (H16 and H17) (Figure S11) as covariates
1038 to reduce false positives due to population stratification. Additionally, a SNP-based kinship
1039 matrix capturing the relationship between accessions was included as random effects. Biallelic
1040 variants with a minor allele frequency greater than 0.01 were retained for kinship estimation
1041 using the VanRaden method [138]. The Bonferroni correction was used to establish a
1042 threshold of significance for all the associations.
1043
1044 **Data processing and visualization**
1045 To calculate the number of genes near each type of TE and differentially expressed
1046 MITEs/IRs, the tool *slop* and *intersect* from bedtools v2.26.0 [59] were used for distances of
1047 500, 1000, and 3000 bp both upstream and downstream of each element. Additionally,
1048 *intersect* was used to count elements located within coding regions, utilizing different
1049 annotation files for each feature.
1050 To evaluate whether MITEs or IRs are preferentially associated with gene expression changes
1051 relative to other genomic contexts, genes were classified into mutually exclusive categories
1052 based on their proximity (≤ 3 kb) to: (i) MITEs or IRs, (ii) other TE classes, or (iii) no annotated

TE (TE-free regions). Each gene was assigned to a single category based on genomic overlap identified using bedtools intersect. For each comparison, genes were further classified as DE or non-DE using the list of DE genes identified from the RNA-seq analysis comparing mature versus immature fruit stages. Contingency tables were constructed, and statistical significance was assessed using Fisher's exact test. Odds ratios (OR) and p-values were used to estimate the relative likelihood of differential expression across genomic contexts.

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.

Principal Component Analysis (PCA) was performed using the R package prcomp [139]. 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), MetBrewer (https://github.com/BlakeRMills/MetBrewer), rtracklayer [140], and ggman (https://github.com/drveera/ggman).

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. The curated TE, MITE, and IR libraries, as well as the code for TE polymorphism and GWAS analyses, are available at: https://github.com/spriego/PriegoCubero_Tolley_et_al_Fvesca

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

R.T. carried out TE identification and classification. **I.T.** conducted MITEs and IRs bioinformatic analysis. **S.P.C.** performed TE polymorphism identification and GWAS analysis. **J.L.G**. and **C.Z** conducted 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.**, and **T.T.** provided genomic data of *F. vesca* accessions (bam files). **I.T.**, **I.T., C.M.P.**, and **P.A.M.** conceived the study. **P.A.M.** secured project funding. **I.T.** and **P.A.M.** supervised the work and 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. We thankfully acknowledge the computer resources (Picasso Supercomputer), technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Malaga. This work was supported by grants from the Spanish Ministry of Science and Innovation (PID2022-137037NB-I00 and CNS2023-145312) to P.A.M.

REFERENCES

1. Matsushima W, Planet E, Trono D: **Ancestral genome reconstruction enhances transposable element annotation by identifying degenerate integrants.** *Cell Genom* 2024, **4:**100497.
2. International Wheat Genome Sequencing C: **Shifting the limits in wheat research and breeding using a fully annotated reference genome.** *Science* 2018, **361**.
3. Hirsch CD, Springer NM: **Transposable element influences on gene expression in plants.** *Biochim Biophys Acta Gene Regul Mech* 2017, **1860:**157-165.
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. 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.
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. Kumar A, Bennetzen JL: **Plant retrotransposons.** *Annu Rev Genet* 1999, **33:**479-532.
10. 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.
11. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS, et al: **Ten things you should know about transposable elements.** *Genome Biol* 2018, **19:**199.
12. 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.
13. 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.
14. 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.
15. 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.
16. Cuerda-Gil D, Slotkin RK: **Non-canonical RNA-directed DNA methylation.** *Nat Plants* 2016, **2:**16163.
17. Vaucheret H, Voinnet O: **The plant siRNA landscape.** *Plant Cell* 2024, **36:**246-275.
18. Yang DL, Huang K, Deng D, Zeng Y, Wang Z, Zhang Y: **DNA-dependent RNA polymerases in plants.** *Plant Cell* 2023, **35:**3641-3661.
19. 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.
20. 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.
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.
22. Shi J, Liang C: **Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection.** *Plant Physiol* 2019, **180:**1803-1815.

1168 23. Storer JM, Hubley R, Rosen J, Smit AFA: **Methodologies for the De novo Discovery of**
1169 **Transposable Element Families.** *Genes (Basel)* 2022, **13**.
1170 24. Mao H, Wang H: **SINE_scan: an efficient tool to discover short interspersed nuclear**
1171 **elements (SINEs) in large-scale genomic datasets.** *Bioinformatics* 2017, **33:**743-745.
1172 25. Ou S, Jiang N: **LTR_retriever: A Highly Accurate and Sensitive Program for**
1173 **Identification of Long Terminal Repeat Retrotransposons.** *Plant Physiol* 2018,
1174 **176:**1410-1422.
1175 26. Huda A, Jordan IK: **Analysis of transposable element sequences using CENSOR and**
1176 **RepeatMasker.** *Methods Mol Biol* 2009, **537:**323-336.
1177 27. Kennedy RC, Unger MF, Christley S, Collins FH, Madey GR: **An automated homology-**
1178 **based approach for identifying transposable elements.** *BMC Bioinformatics* 2011,
1179 **12:**130.
1180 28. Baril T, Galbraith J, Hayward A: **Earl Grey: A Fully Automated User-Friendly**
1181 **Transposable Element Annotation and Analysis Pipeline.** *Mol Biol Evol* 2024, **41**.
1182 29. Riehl K, Riccio C, Miska EA, Hemberg M: **TransposonUltimate: software for**
1183 **transposon classification, annotation and detection.** *Nucleic Acids Res* 2022, **50:**e64.
1184 30. Castanera R, Morales-Díaz N, Gupta S, Purugganan M, Casacuberta JM: **Transposons**
1185 **are important contributors to gene expression variability under selection in rice**
1186 **populations.** *eLife* 2023, **12:**RP86324.
1187 31. Rohilla M, Mazumder A, Saha D, Pal T, Begam S, Mondal TK: **Genome-wide**
1188 **identification and development of miniature inverted-repeat transposable elements**
1189 **and intron length polymorphic markers in tea plant (Camellia sinensis).** *Sci Rep*
1190 2022, **12:**16233.
1191 32. Akakpo R, Carpentier MC, Ie Hsing Y, Panaud O: **The impact of transposable elements**
1192 **on the structure, evolution and function of the rice genome.** *New Phytol* 2020,
1193 **226:**44-49.
1194 33. Liang Y, Liu HJ, Yan J, Tian F: **Natural Variation in Crops: Realized Understanding,**
1195 **Continuing Promise.** *Annu Rev Plant Biol* 2021, **72:**357-385.
1196 34. Yan H, Haak DC, Li S, Huang L, Bombarely A: **Exploring transposable element-based**
1197 **markers to identify allelic variations underlying agronomic traits in rice.** *Plant*
1198 *Commun* 2022, **3:**100270.
1199 35. Yin Z, Yang Q, Shen D, Liu J, Huang W, Dou D: **Online data resource for exploring**
1200 **transposon insertion polymorphisms in public soybean germplasm accessions.**
1201 *Plant Physiol* 2023, **193:**1036-1044.
1202 36. Cai X, Lin R, Liang J, King GJ, Wu J, Wang X: **Transposable element insertion: a**
1203 **hidden major source of domesticated phenotypic variation in Brassica rapa.** *Plant*
1204 *Biotechnol J* 2022, **20:**1298-1310.
1205 37. Dominguez M, Dugas E, Benchouaia M, Leduque B, Jimenez-Gomez JM, Colot V,
1206 Quadrana L: **The impact of transposable elements on tomato diversity.** *Nat Commun*
1207 2020, **11:**4058.
1208 38. Vourlaki IT, Castanera R, Ramos-Onsins SE, Casacuberta JM, Perez-Enciso M:
1209 **Transposable element polymorphisms improve prediction of complex agronomic**
1210 **traits in rice.** *Theor Appl Genet* 2022, **135:**3211-3222.
1211 39. Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD, Teresi
1212 SJ, Nelson ADL, Wai CM, et al: **Origin and evolution of the octoploid strawberry**
1213 **genome.** *Nat Genet* 2019, **51:**541-547.
1214 40. Lyu K, Xiao J, Lyu S, Liu R: **Comparative Analysis of Transposable Elements in**
1215 **Strawberry Genomes of Different Ploidy Levels.** *Int J Mol Sci* 2023, **24**.
1216 41. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P,
1217 Mockaitis K, Liston A, Mane SP, et al: **The genome of woodland strawberry (Fragaria**
1218 **vesca).** *Nat Genet* 2011, **43:**109-116.

42. 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.
43. 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.
44. Toivainen T, Salonen JS, Kirshner J, Lembinen S, De Kort H, Lyyski A, Edger PP, Hilmarsson HS, Hallsson JH, Sargent DJ, et al: **The Late Quaternary climate impact on the genome of the woodland strawberry (Fragaria vesca), a perennial herb.** *Commun Biol* 2026.
45. 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.
46. Goerner-Potvin P, Bourque G: **Computational tools to unmask transposable elements.** *Nat Rev Genet* 2018, **19:**688-704.
47. Mhiri C, Borges F, Grandbastien MA: **Specificities and Dynamics of Transposable Elements in Land Plants.** *Biology (Basel)* 2022, **11**.
48. 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.
49. 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.
50. Qian J, Xue H, Ou S, Storer J, Fürtauer L, Wildermuth MC, Kusch S, Panstruga R: **TEtrimmer: a novel tool to automate the manual curation of transposable elements.** *bioRxiv* 2024**:**2024.2006.2027.600963.
51. 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.
52. 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.
53. 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**.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.

1271 59. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic**
1272 **features.** *Bioinformatics* 2010, **26:**841-842.
1273 60. Jouffroy O, Saha S, Mueller L, Quesneville H, Maumus F: **Comprehensive repeatome**
1274 **annotation reveals strong potential impact of repetitive elements on tomato**
1275 **ripening.** *BMC Genomics* 2016, **17:**624.
1276 61. Yang L, Bennetzen JL: **Structure-based discovery and description of plant and**
1277 **animal Helitrons.** *Proc Natl Acad Sci U S A* 2009, **106:**12832-12837.
1278 62. Yang L, Bennetzen JL: **Distribution, diversity, evolution, and survival of Helitrons in**
1279 **the maize genome.** *Proc Natl Acad Sci U S A* 2009, **106:**19922-19927.
1280 63. Guo C, Spinelli M, Ye C, Li QQ, Liang C: **Genome-Wide Comparative Analysis of**
1281 **Miniature Inverted Repeat Transposable Elements in 19 Arabidopsis thaliana**
1282 **Ecotype Accessions.** *Sci Rep* 2017, **7:**2634.
1283 64. Arce AL, Mencia R, Cambiagno DA, Lang PL, Liu C, Burbano HA, Weigel D, Manavella
1284 PA: **Polymorphic inverted repeats near coding genes impact chromatin topology**
1285 **and phenotypic traits in Arabidopsis thaliana.** *Cell Rep* 2023, **42:**112029.
1286 65. Mencia R, Arce AL, Houriet C, Xian W, Contreras A, Shirsekar G, Weigel D, Manavella
1287 PA: **Transposon-triggered epigenetic chromatin dynamics modulate EFR-related**
1288 **pathogen response.** *Nat Struct Mol Biol* 2025, **32:**199-211.
1289 66. Gagliardi D, Cambiagno DA, Arce AL, Tomassi AH, Giacomelli JI, Ariel FD, Manavella
1290 PA: **Dynamic regulation of chromatin topology and transcription by inverted repeat-**
1291 **derived small RNAs in sunflower.** *Proc Natl Acad Sci U S A* 2019, **116:**17578-17583.
1292 67. Gagliardi D, Manavella PA: **Short-range regulatory chromatin loops in plants.** *New*
1293 *Phytol* 2020, **228:**466-471.
1294 68. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open**
1295 **Software Suite.** *Trends Genet* 2000, **16:**276-277.
1296 69. Li R, Yao J, Cai S, Fu Y, Lai C, Zhu X, Cui L, Li Y: **Genome-wide characterization and**
1297 **evolution analysis of miniature inverted-repeat transposable elements in Barley**
1298 **(Hordeum vulgare).** *Front Plant Sci* 2024, **15:**1474846.
1299 70. Lu C, Chen J, Zhang Y, Hu Q, Su W, Kuang H: **Miniature inverted-repeat transposable**
1300 **elements (MITEs) have been accumulated through amplification bursts and play**
1301 **important roles in gene expression and species diversity in Oryza sativa.** *Mol Biol*
1302 *Evol* 2012, **29:**1005-1017.
1303 71. Sun Q, He Z, Wei R, Ye J, Chai L, Cheng Y, Xu Q, Deng X: **Red peel regulator 1 links**
1304 **ethylene response factor 25 and beta-citraurin biosynthetic genes to regulate**
1305 **ethylene-induced peel reddening in citrus.** *Plant Cell* 2024, **37**.
1306 72. Yang X, Liu Q, Wang MM, Wang XY, Han MQ, Liu FP, Lu TF, Liu J, Wang YZ: **A single**
1307 **dominant GLOBOSA allele accounts for repeated origins of hose-in-hose flowers**
1308 **in Sinningia (Gesneriaceae).** *Plant Cell* 2024, **37**.
1309 73. Defilippi BG, Kader AA, Dandekar AM: **Apple aroma: alcohol acyltransferase, a rate**
1310 **limiting step for ester biosynthesis, is regulated by ethylene.** *Plant Science* 2005,
1311 **168:**1199-1210.
1312 74. Zhang Z, Yu S, Zhang Z, Zhang J, Li H: **Comparative Characterization of Fruit Volatiles**
1313 **and Volatile-Related Genes Expression of 'Benihoppe' Strawberry and Its**
1314 **Somaclonal Mutant.** *Plants (Basel)* 2023, **12**.
1315 75. Li Y-p, Liu T-j, Luo H-f, Liu S-c: **The transcriptional landscape of cultivated strawberry**
1316 **(Fragaria×ananassa) and its diploid ancestor (Fragaria vesca) during fruit**
1317 **development.** *Journal of Integrative Agriculture* 2021, **20:**1540-1553.
1318 76. Nieuwenhuizen NJ, Maddumage R, Tsang GK, Fraser LG, Cooney JM, De Silva HN,
1319 Green S, Richardson KA, Atkinson RG: **Mapping, complementation, and targets of the**
1320 **cysteine protease actinidin in kiwifruit.** *Plant Physiol* 2012, **158:**376-388.
1321 77. Wang W, Wang Y, Chen T, Qin G, Tian S: **Current insights into posttranscriptional**
1322 **regulation of fleshy fruit ripening.** *Plant Physiol* 2023, **192:**1785-1798.

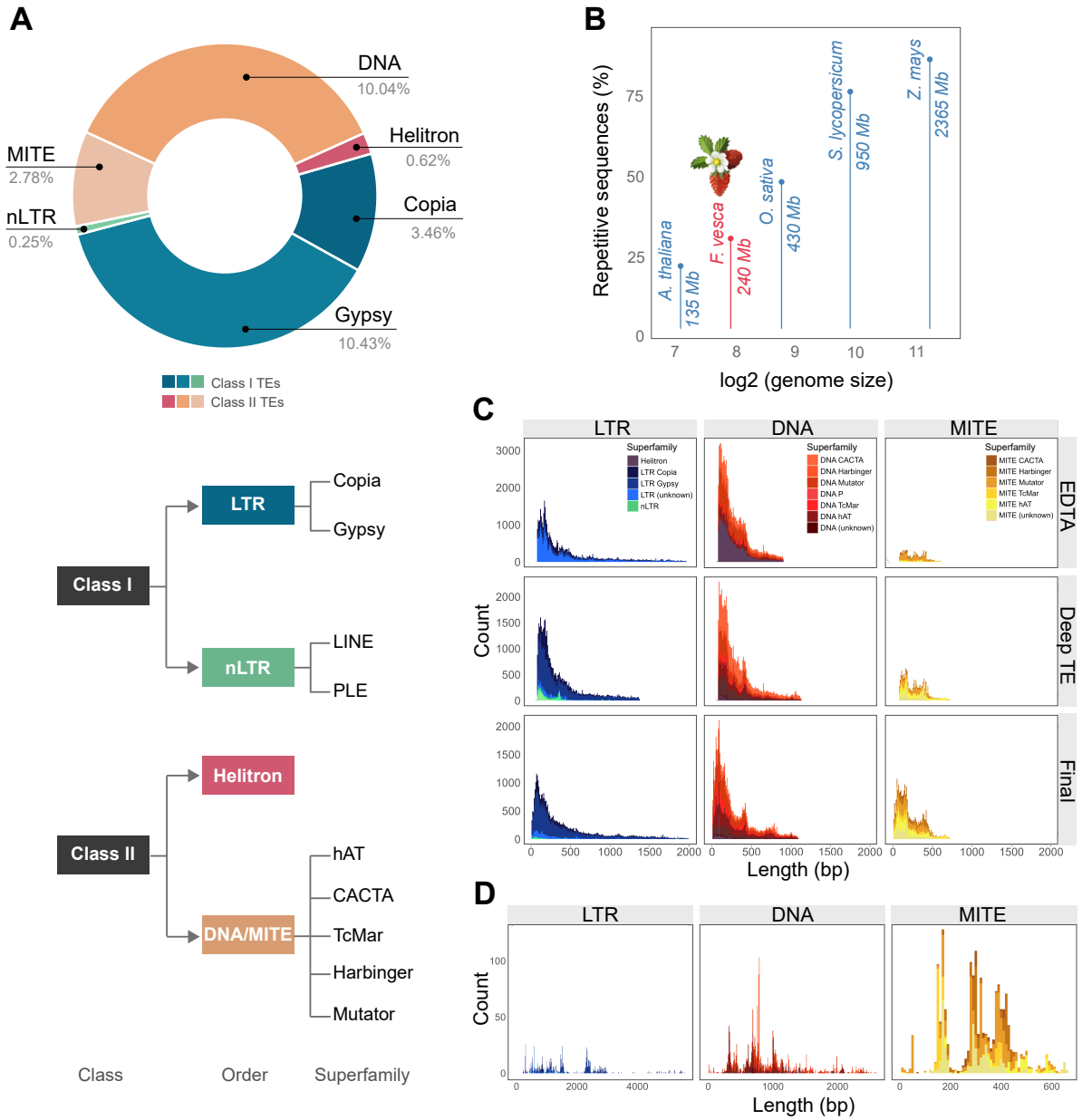
78. Zhai Y, Cui Y, Song M, Vainstein A, Chen S, Ma H: **Papain-Like Cysteine Protease Gene Family in Fig (Ficus carica L.): Genome-Wide Analysis and Expression Patterns.** *Front Plant Sci* 2021, **12:**681801.
79. Sanchez-Sevilla JF, Cruz-Rus E, Valpuesta V, Botella MA, Amaya I: **Deciphering gamma-decalactone biosynthesis in strawberry fruit using a combination of genetic mapping, RNA-Seq and eQTL analyses.** *BMC Genomics* 2014, **15:**218.
80. Seymour GB, Ryder CD, Cevik V, Hammond JP, Popovich A, King GJ, Vrebalov J, Giovannoni JJ, Manning K: **A SEPALLATA gene is involved in the development and ripening of strawberry (Fragaria x ananassa Duch.) fruit, a non-climacteric tissue.** *J Exp Bot* 2011, **62:**1179-1188.
81. Vallarino JG, Merchante C, Sanchez-Sevilla JF, de Luis Balaguer MA, Pott DM, Ariza MT, Casanal A, Pose D, Vioque A, Amaya I, et al: **Characterizing the involvement of FaMADS9 in the regulation of strawberry fruit receptacle development.** *Plant Biotechnol J* 2020, **18:**929-943.
82. Martinez-Rivas FJ, Blanco-Portales R, Molina-Hidalgo FJ, Caballero JL, Perez de Souza L, Alseekh S, Fernie AR, Munoz-Blanco J, Rodriguez-Franco A: **Azacytidine arrests ripening in cultivated strawberry (Fragaria x ananassa) by repressing key genes and altering hormone contents.** *BMC Plant Biol* 2022, **22:**278.
83. Liu L, Xiao W, Ji M-l, Yang C, Li L, Gao D-s, Fu X-l: **Effects of molybdenum on nutrition, quality, and flavour compounds of strawberry (Fragaria×ananassa Duch. cv. Akihime) fruit.** *Journal of Integrative Agriculture* 2017, **16:**1502-1512.
84. Cosgrove DJ: **Growth of the plant cell wall.** *Nat Rev Mol Cell Biol* 2005, **6:**850-861.
85. Del Olmo I, Romero I, Alvarez MD, Tarradas R, Sanchez-Ballesta MT, Escribano MI, Merodio C: **Transcriptomic analysis of CO(2)-treated strawberries (Fragaria vesca) with enhanced resistance to softening and oxidative stress at consumption.** *Front Plant Sci* 2022, **13:**983976.
86. Ariel FD, Manavella PA: **When junk DNA turns functional: transposon-derived non-coding RNAs in plants.** *J Exp Bot* 2021, **72:**4132-4143.
87. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al: **Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato.** *Cell* 2020, **182:**145-161 e123.
88. Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui MA, Guy J, Bortolini Silveira A, Engelen S, Baillet V, Wincker P, et al: **Transposition favors the generation of large effect mutations that may facilitate rapid adaption.** *Nat Commun* 2019, **10:**3421.
89. Saxena RK, Edwards D, Varshney RK: **Structural variations in plant genomes.** *Brief Funct Genomics* 2014, **13:**296-307.
90. Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R: **Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation.** *Elife* 2016, **5**.
91. Tossolini I, Mencia R, Arce AL, Manavella PA: **The genome awakens: transposon-mediated gene regulation.** *Trends Plant Sci* 2025, **30:**857-871.
92. Contreras-Garrido A, Galanti D, Movilli A, Becker C, Bossdorf O, Drost HG, Weigel D: **Transposon dynamics in the emerging oilseed crop Thlaspi arvense.** *PLoS Genet* 2024, **20:**e1011141.
93. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, Colot V: **The Arabidopsis thaliana mobilome and its impact at the species level.** *Elife* 2016, **5**.
94. Catlin NS, Josephs EB: **The important contribution of transposable elements to phenotypic variation and evolution.** *Curr Opin Plant Biol* 2022, **65:**102140.
95. Urrutia M, Meco V, Rambla JL, Martin-Pizarro C, Pillet J, Andres J, Sanchez-Sevilla JF, Granell A, Hytonen T, Pose D: **Diversity of the volatilome and the fruit size and shape in European woodland strawberry (Fragaria vesca).** *Plant J* 2023, **116:**1201-1217.

96. Baduel P, Leduque B, Ignace A, Gy I, Gil J, Jr., Loudet O, Colot V, Quadrana L: **Genetic and environmental modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana.** *Genome Biol* 2021, **22:**138.
97. Shi J, Cao C, Xu J, Zhou C: **Research Advances on Biosynthesis, Regulation, and Biological Activities of Apocarotenoid Aroma in Horticultural Plants.** *Journal of Chemistry* 2020, **2020:**2526956.
98. Riikonen J, Ruhanen H, Uimari A, Poteri M, Toljamo A, Kokko H, Blande JD, Kumpula R, Kivimäenpää M: **Impact of pre-harvest UVC treatment on powdery mildew infection and strawberry quality in tunnel production in Nordic conditions.** *Scientia Horticulturae* 2024, **338:**113706.
99. Achimon F, Peschiutta ML, Brito VD, Ulla SB, Pizzolitto RP: **Sulcatone as a Plant-Derived Volatile Organic Compound for the Control of the Maize Weevil and Its Associated Phytopathogenic Fungi in Stored Maize.** *Plants (Basel)* 2024, **13**.
100. Antonious GF, Dahlman DL, Hawkins LM: **Insecticidal and acaricidal performance of methyl ketones in wild tomato leaves.** *Bull Environ Contam Toxicol* 2003, **71:**400-407.
101. Shigeto J, Tsutsumi Y: **Diverse functions and reactions of class III peroxidases.** *New Phytol* 2016, **209:**1395-1402.
102. Matsui K: **How and Why Plants Came to Smell Green: The Origins, Biosynthesis, and Roles of Green Leaf Volatiles.** *J Exp Bot* 2025.
103. Youn B, Kim S-J, Moinuddin SGA, Lee C, Bedgar DL, Harper AR, Davin LB, Lewis NG, Kang C: **Mechanistic and Structural Studies of Apoform, Binary, and Ternary Complexes of the Arabidopsis Alkenal Double Bond Reductase At5g16970*.** *Journal of Biological Chemistry* 2006, **281:**40076-40088.
104. Zhang Y, Teoh KH, Reed DW, Covello PS: **Molecular cloning and characterization of Dbr1, a 2-alkenal reductase from Artemisia annuaThe nucleotide sequence reported in this article has been deposited in the GenBank database under accession No. FJ750460.This paper is one of a selection of papers published in a Special Issue from the National Research Council of Canada – Plant Biotechnology Institute.** *Botany* 2009, **87:**643-649.
105. Jiménez-Muñoz R, Urrutia M, Meco V, Rambla JL, Toivainen T, Pérez-Hedo M, Sánchez-Sevilla JF, Urbaneja A, Salas JJ, Hytönen T, et al: **Uncovering the genetic basis of fruit volatiles in Fragaria vesca through GWAS reveals FvJMT2 as a methyl benzoate biosynthesis gene with insect-repellent function.** *bioRxiv* 2025**:**2025.2011.2018.689018.
106. Urrutia M, Rambla JL, Alexiou KG, Granell A, Monfort A: **Genetic analysis of the wild strawberry (Fragaria vesca) volatile composition.** *Plant Physiol Biochem* 2017, **121:**99-117.
107. Holland CK, Watson AP, Chiang E: **Anthranilate at the interface of tryptophan and specialized metabolite biosynthesis.** *Front Plant Sci* 2025, **16:**1625337.
108. Barbey CR, Hogshead MH, Harrison B, Schwartz AE, Verma S, Oh Y, Lee S, Folta KM, Whitaker VM: **Genetic Analysis of Methyl Anthranilate, Mesifurane, Linalool, and Other Flavor Compounds in Cultivated Strawberry (Fragaria x ananassa).** *Front Plant Sci* 2021, **12:**615749.
109. Schwieterman ML, Colquhoun TA, Jaworski EA, Bartoshuk LM, Gilbert JL, Tieman DM, Odabasi AZ, Moskowitz HR, Folta KM, Klee HJ, et al: **Strawberry flavor: diverse chemical compositions, a seasonal influence, and effects on sensory perception.** *PLoS One* 2014, **9:**e88446.
110. Bontpart T, Cheynier V, Ageorges A, Terrier N: **BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds.** *New Phytol* 2015, **208:**695-707.
111. Mugford ST, Milkowski C: **Serine carboxypeptidase-like acyltransferases from plants.** *Methods Enzymol* 2012, **516:**279-297.

1428 112. Beekwilder J, Alvarez-Huerta M, Neef E, Verstappen FW, Bouwmeester HJ, Aharoni
1429 A: **Functional characterization of enzymes forming volatile esters from strawberry**
1430 **and banana.** *Plant Physiol* 2004, **135:**1865-1878.
1431 113. Gonzalez M, Gaete-Eastman C, Valdenegro M, Figueroa CR, Fuentes L, Herrera R,
1432 Moya-Leon MA: **Aroma development during ripening of Fragaria chiloensis fruit and**
1433 **participation of an alcohol acyltransferase (FcAAT1) gene.** *J Agric Food Chem* 2009,
1434 **57:**9123-9132.
1435 114. Gill RA, Scossa F, King GJ, Golicz AA, Tong C, Snowdon RJ, Fernie AR, Liu S: **On**
1436 **the Role of Transposable Elements in the Regulation of Gene Expression and**
1437 **Subgenomic Interactions in Crop Genomes.** *Critical Reviews in Plant Sciences* 2021,
1438 **40:**157-189.
1439 115. Studer A, Zhao Q, Ross-Ibarra J, Doebley J: **Identification of a functional**
1440 **transposon insertion in the maize domestication gene tb1.** *Nat Genet* 2011, **43:**1160-
1441 1163.
1442 116. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W,
1443 Hellsten U, Putnam N, Rokhsar DS: **Phytozome: a comparative platform for green**
1444 **plant genomics.** *Nucleic Acids Res* 2012, **40:**D1178-1186.
1445 117. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
1446 Maurano MT, Vierstra J, Thomas S, et al: **BEDOPS: high-performance genomic**
1447 **feature operations.** *Bioinformatics* 2012, **28:**1919-1920.
1448 118. Qian J, Xue H, Ou S, Mann L, Storer J, Furtauer L, Heitkam T, Wildermuth MC, Kusch
1449 S, Panstruga R: **TEtrimmer: a tool to automate the manual curation of transposable**
1450 **elements.** *Nat Commun* 2025, **16:**8429.
1451 119. Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next-**
1452 **generation sequencing data.** *Bioinformatics* 2012, **28:**3150-3152.
1453 120. Li X, Martin-Pizarro C, Zhou L, Hou B, Wang Y, Shen Y, Li B, Pose D, Qin G:
1454 **Deciphering the regulatory network of the NAC transcription factor FvRIF, a key**
1455 **regulator of strawberry (Fragaria vesca) fruit ripening.** *Plant Cell* 2023, **35:**4020-4045.
1456 121. Gambino G, Perrone I, Gribaudo I: **A Rapid and effective method for RNA**
1457 **extraction from different tissues of grapevine and other woody plants.** *Phytochem*
1458 *Anal* 2008, **19:**520-525.
1459 122. Ewels P, Magnusson M, Lundin S, Kaller M: **MultiQC: summarize analysis results**
1460 **for multiple tools and samples in a single report.** *Bioinformatics* 2016, **32:**3047-3048.
1461 123. Axtell MJ: **ShortStack: comprehensive annotation and quantification of small**
1462 **RNA genes.** *RNA* 2013, **19:**740-751.
1463 124. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion**
1464 **for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.
1465 125. Lu P, Yu S, Zhu N, Chen YR, Zhou B, Pan Y, Tzeng D, Fabi JP, Argyris J, Garcia-Mas
1466 J, et al: **Genome encode analyses reveal the basis of convergent evolution of fleshy**
1467 **fruit ripening.** *Nat Plants* 2018, **4:**784-791.
1468 126. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina**
1469 **sequence data.** *Bioinformatics* 2014, **30:**2114-2120.
1470 127. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
1471 Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-
1472 21.
1473 128. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
1474 Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format**
1475 **and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
1476 129. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program**
1477 **for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30:**923-930.
1478 130. Koster J, Rahmann S: **Snakemake--a scalable bioinformatics workflow engine.**
1479 *Bioinformatics* 2012, **28:**2520-2522.

131. 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.
132. 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**.
133. 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.
134. Faust GG, Hall IM: **YAHA: fast and flexible long-read alignment with optimal breakpoint detection.** *Bioinformatics* 2012, **28:**2417-2424.
135. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for Computing and Annotating Genomic Ranges.** *PLOS Computational Biology* 2013, **9:**e1003118.
136. Wang J, Zhang Z: **GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction.** *Genomics Proteomics Bioinformatics* 2021, **19:**629-640.
137. Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M: **An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations.** *Nat Genet* 2012, **44:**825-830.
138. VanRaden PM: **Efficient methods to compute genomic predictions.** *J Dairy Sci* 2008, **91:**4414-4423.
139. Sigg CD, Buhmann JM: **Expectation-maximization for sparse and non-negative PCA.** In *Proceedings of the 25th international conference on Machine learning*. pp. 960–967. Helsinki, Finland: Association for Computing Machinery; 2008:960–967.
140. Lawrence M, Gentleman R, Carey V: **rtracklayer: an R package for interfacing with genome browsers.** *Bioinformatics* 2009, **25:**1841-1842.

1509 **FIGURE LEGENDS**



1510

1511 **Figure 1. Transposable Element (TE) diversity in *Fragaria vesca*. (A)** Total coverage of
1512 the *F. vesca* genome assembly by the main TE classes, shown as percentages relative to the
1513 entire genome. Representation of TE classification scheme generated in this study is
1514 displayed. **(B)** Repetitive sequence content as a proportion of total genome size in several
1515 plant species, listed in ascending order: *Arabidopsis thaliana*, woodland strawberry (*Fragaria*
1516 *vesca*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), and maize (*Zea mays*). **(C)**
1517 Length distribution of TEs (including fragmented TEs) in the *F. vesca* v4.0.a2 annotation. The
1518 classifications are listed in descending order for EDTA, DeepTE, and the curated annotation.
1519 **(D)** Length distribution of intact TEs in the curated *F. vesca* v4 annotation.

1520

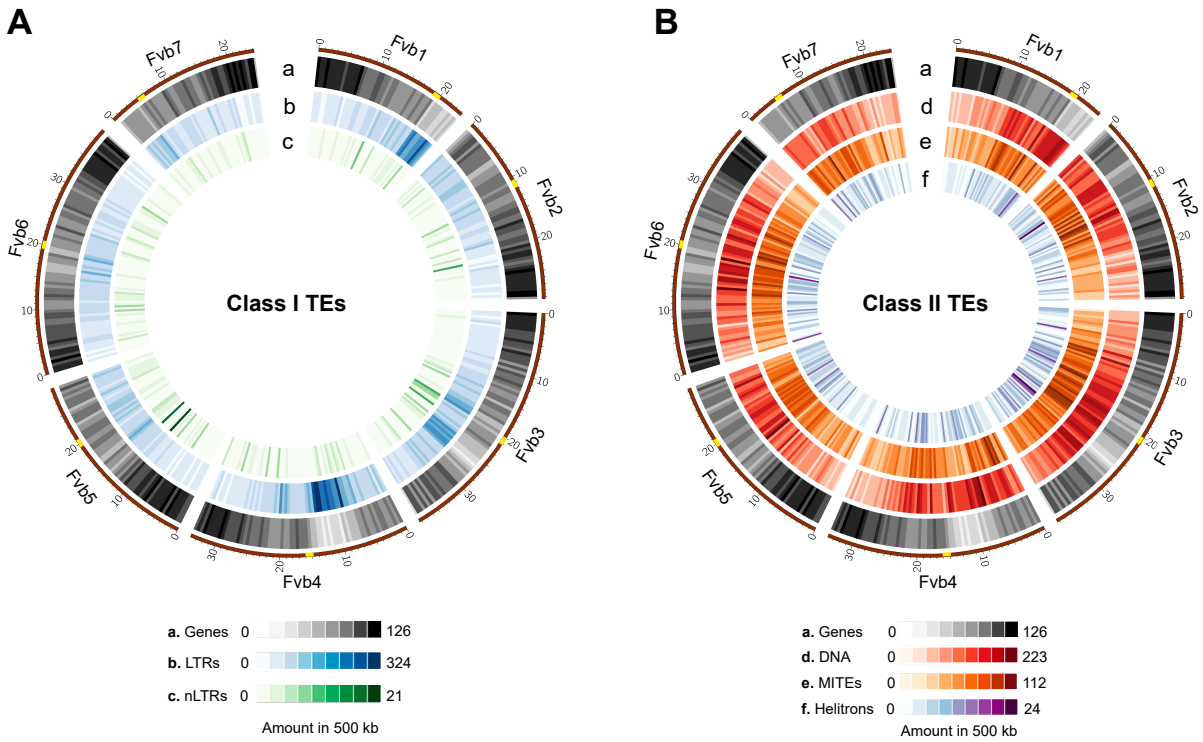


Figure 2. Distribution of TEs in the *F. vesca* genome, represented as density plots within 500 kb windows. (A) Class I TEs: LTR (Track b in blue), non-LTR (Track c in green). **(B)** Class II TEs: DNA (Track “d” red), MITE (Track “e” in orange), Helitron (Track “f” in purple). The color scale indicates 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.

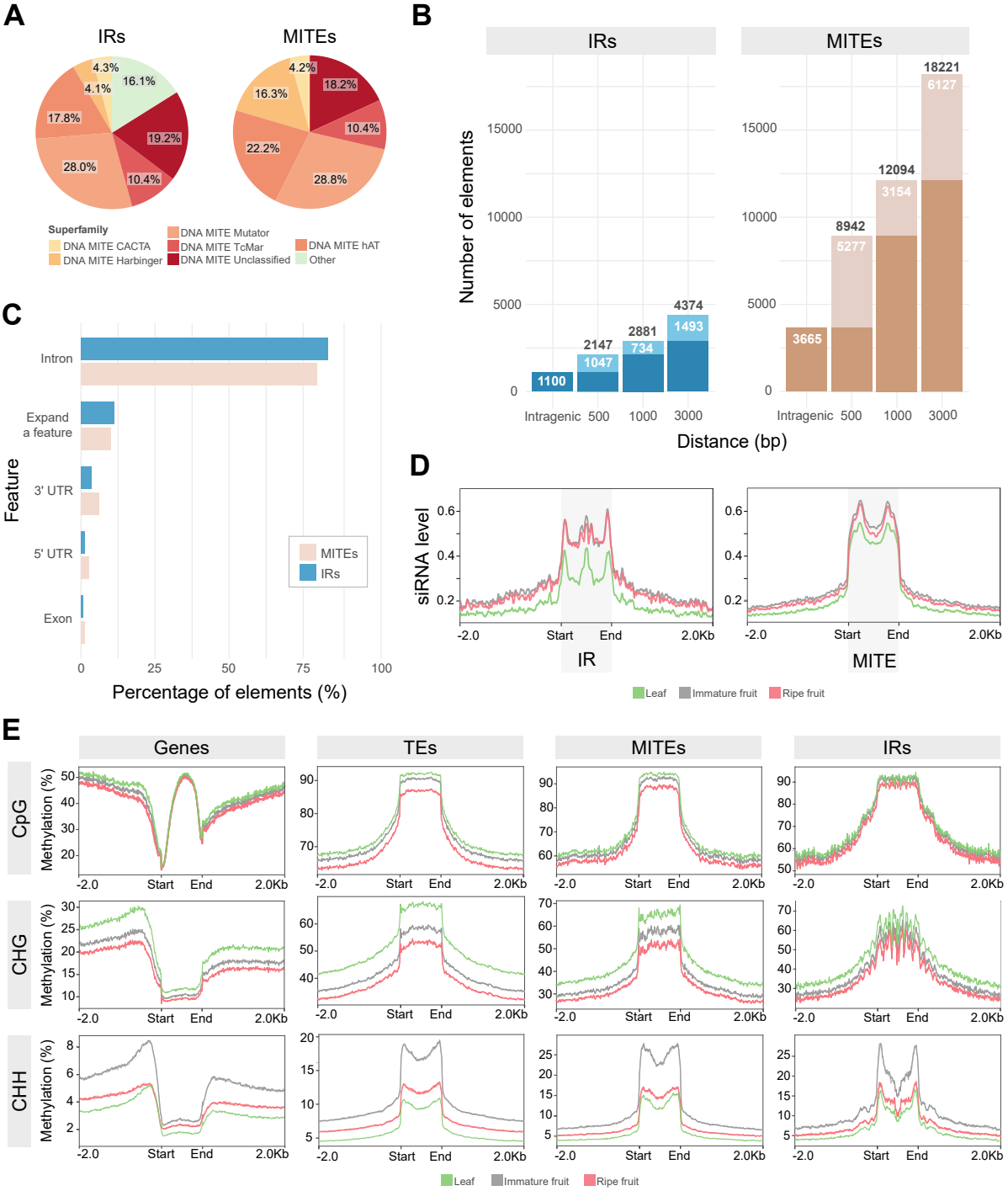


Figure 3. Characterization of TE-Derived IRs and MITEs in *F. vesca*.

(A) Proportion of the identified IRs (left) and annotated MITEs (right) classified by TE superfamily in *F. vesca*. **(B)** 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. **(C)** Number of IRs and MITEs 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. **(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 (shades of gray), and ripe fruits (shades of 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.

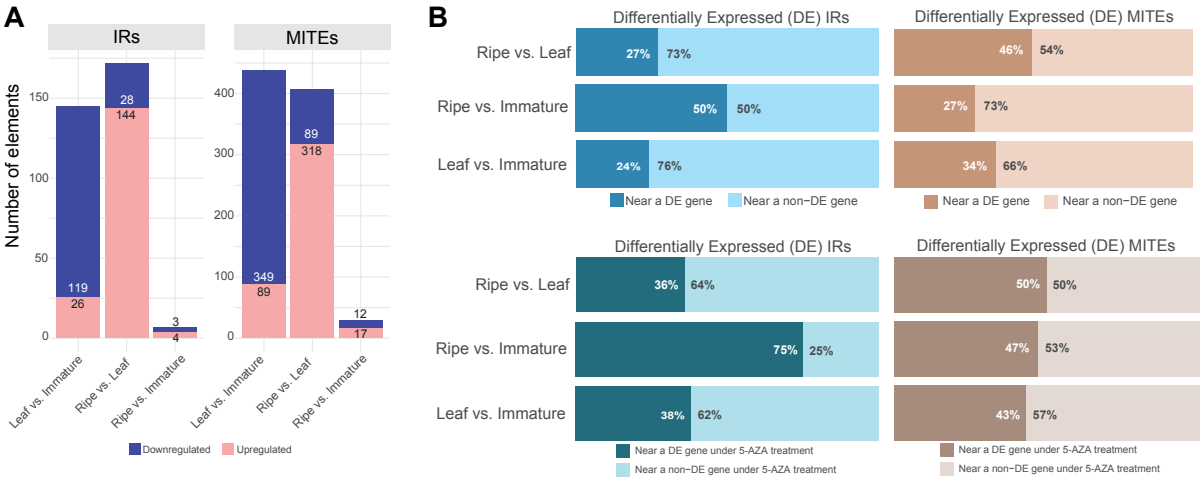


Figure 4. Regulatory role of MITEs and IRs in gene expression.
(A) Differential accumulation analysis of 24-nt siRNAs derived from IRs and MITEs across leaf tissue, immature and ripe fruits, using thresholds of |log2-fold change| ≥ 1 and an adjusted p-value (padj) ≤ 0.05. **(B)** Differential accumulation analysis of 24-nt siRNAs derived from IRs (left) and MITEs (right) elements across leaf tissue, immature and ripe fruits, using thresholds of log2-fold change (FC) ≥0.5 and adjusted p-value (padj) ≤0.05. The percentages of elements located within 3,000 bp of either differentially expressed genes (DE) or non-DE genes are shown.

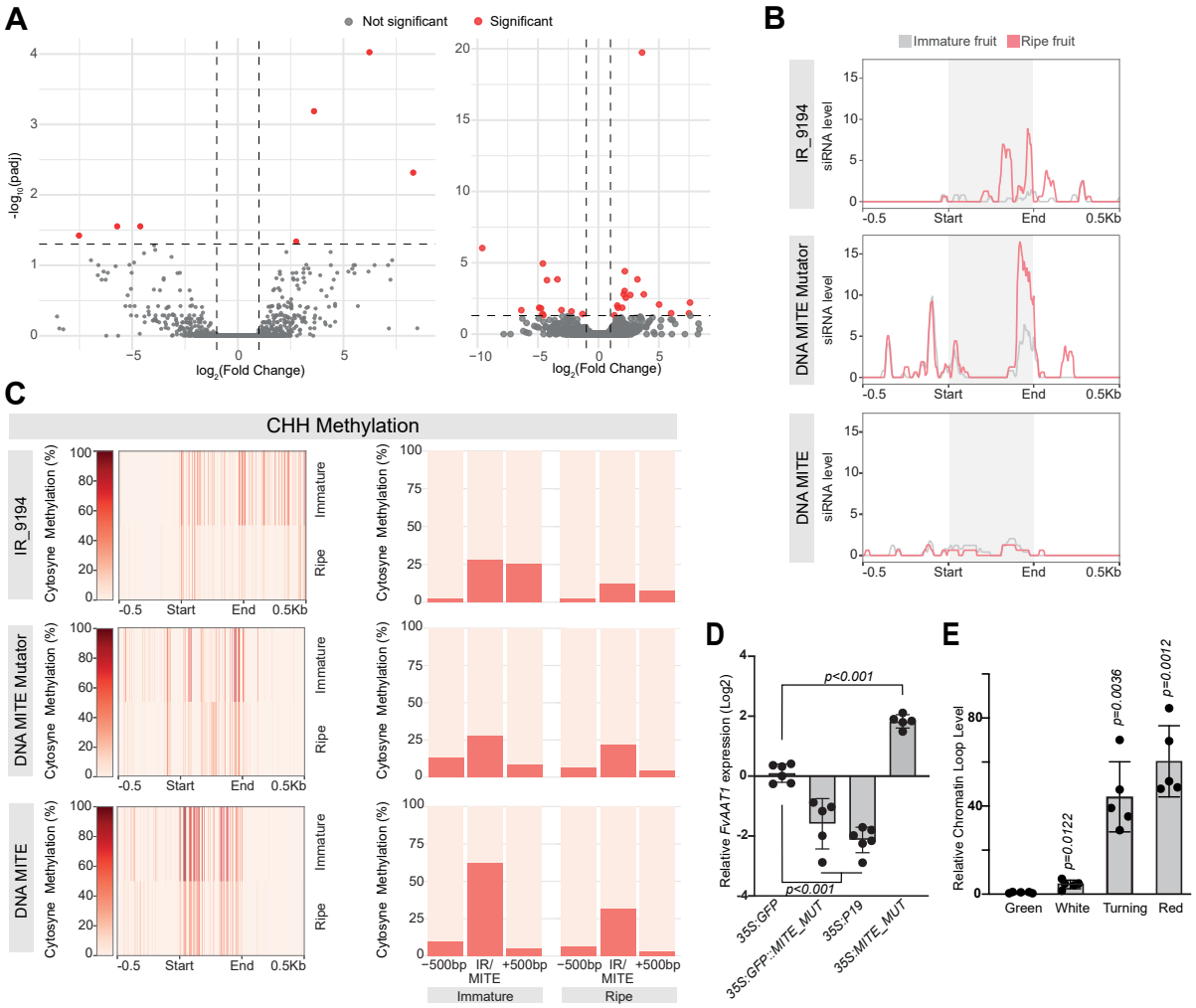


Figure 5. A *FvAAT1*-hosted MITE control the gene expression through epigenetically driven chromatin reorganization

(A) Volcano plots showing the log2FC of the differential expression analysis for IRs (left) and MITEs (right) between immature and ripe stages. Red points represent elements with an adjusted p-value ≤0.05. **(B)** 24-nt siRNAs mapping to IR_9194 and the two MITEs near FvH4_7g18550 and FvH4_7g18570 genes (highlighted by a gray box) in immature (gray) and ripe (pink) fruits. Each element is scaled from start to end plus 500 bp on each side, and all sRNA-seq replicates are plotted together. **(C)** Cytosine methylation in the CHH context for *IR_9194*, MITE Mutator, and MITE, including 500-bp flanking regions, in immature and ripe fruits. The average of individual BS-seq replicates is plotted for each condition. Average percentage of CHH methylation in each element 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 MITE Mutator 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.

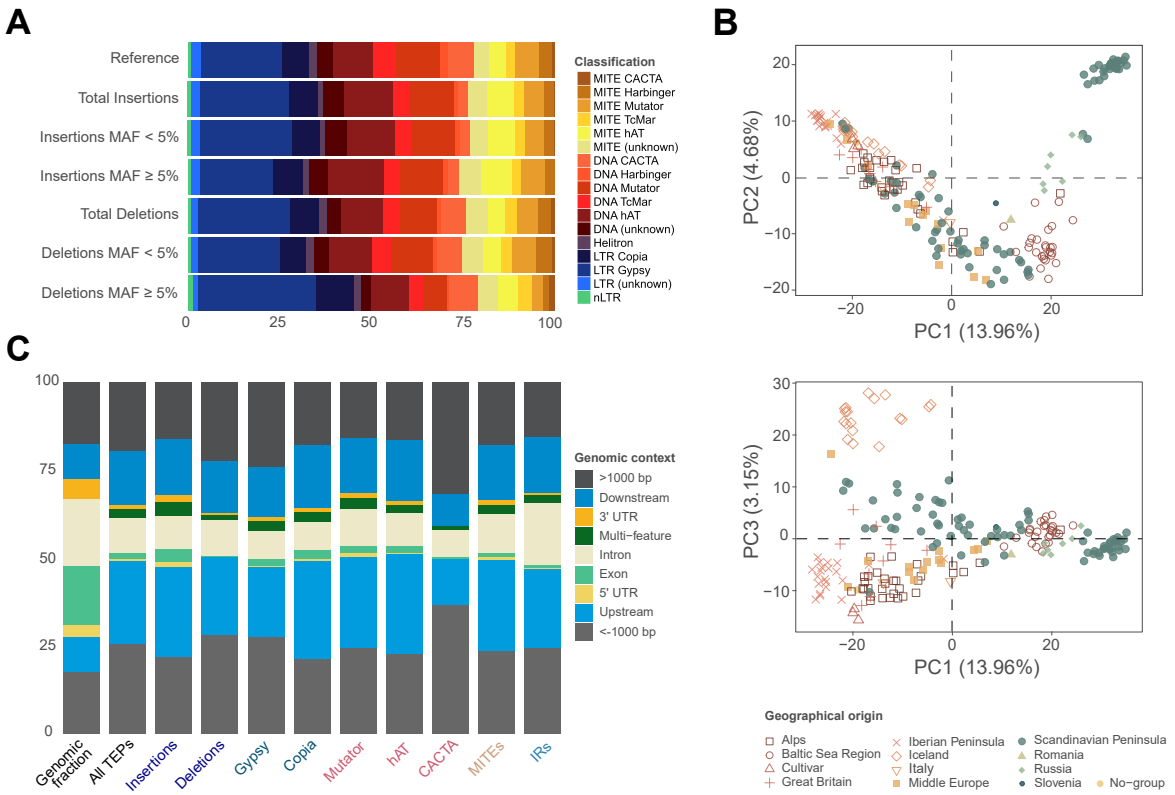


Figure 6. Distribution of insertion and deletion polymorphisms in the *F. vesca* v4.0.a2 annotation.

(A) Proportion (%) of identified polymorphisms by TE superfamily. Rare TE polymorphisms are indicated by minor allele frequency (MAF) < 5% and frequent TE polymorphisms by MAF ≥ 5%. (B) Principal Component Analysis (PCA) based on TE insertion and deletion polymorphisms (MAF ≥ 5%). The geographic origin of the analyzed F. vesca accessions is indicated according to Urrutia et al [95]. **(C)** Distribution of polymorphic elements over genic features, including 5' UTR, Exons, Introns, and 3' UTR (UTR: unstranslated transcribed region; Multi-feature: polymorphisms overlpaping multiple annotations). Stacked bar plots represent the relative proportion of Insertions, Deletions, and specific TE superfamilies. Polymorphisms located within 1000 bp of a gene are shown in shades of blue, representing Upstream and Downstream regions. Intergenic polymorphisms located more than 1000 bp away from those regions are indicated in grey (<-1000 bp and >1000 bp). The 'Genomic fraction' bar represents the percentage of base pairs occupied by each feature in the reference genome.

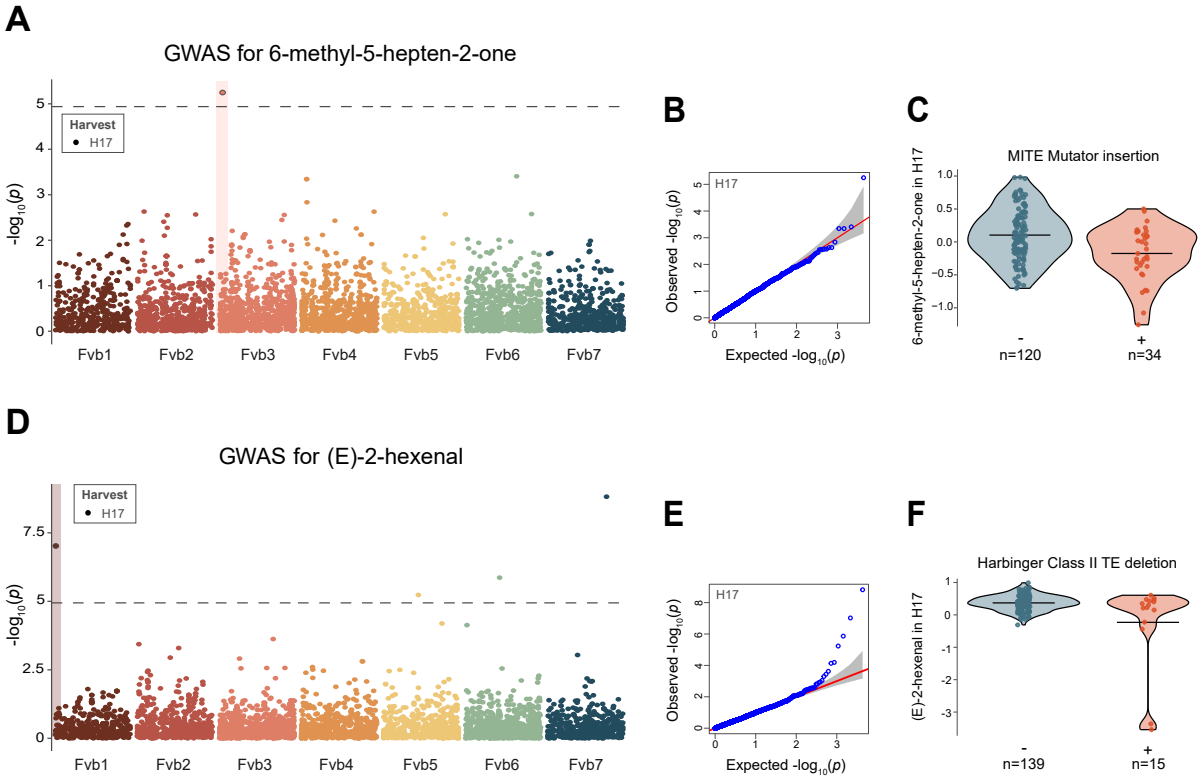


Figure 7. TE polymorphisms linked to volatile compound variation in a collection of *F. vesca* accessions. (A) Manhattan plot of TEP-GWAS for 6-methyl-5-hepten-2-one in harvest season H17. The dotted gray line marks the significance threshold determined by the Bonferroni correction, and the significantly associated polymorphism is highlighted. **(B)** Observed and expected distribution of *p* values for TEP-GWAS for 6-methyl-5-hepten-2-one. **(C)** Violin plot showing the relative abundance of 6-methyl-5-hepten-2-one in accessions carrying the MITE insertion (+, coral) versus non-carriers (-, steel blue). Horizontal lines in (-) and (+) indicate the mean. **(D)** Manhattan plot of TEP-GWAS for (E)-2-hexenal in harvest season H17. The dotted gray line marks the significance threshold determined by the Bonferroni correction, and the significantly associated polymorphism is highlighted. **(E)** Observed and expected distribution of *p* values for TEP-GWAS for (E)-2-hexenal. **(F)** Violin plot showing the relative abundance of (E)-2-hexenal in accessions carrying the MITE insertion (+, coral) versus non-carriers (-, steel blue). Horizontal lines in (-) and (+) indicate the mean.

TABLES

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

Classification	N° of TEs	N° of Copies	Genome Coverage (%)
Class I			
LTR			
Copia	131	8,647	3.46
Gypsy	262	25,977	10.43
Unclassified	35	3,536	0.91
Total LTRs	428	38,160	14.80
non-LTR			
LINE (L1)	8	292	0.07
PLE	7	548	0.18
Total non-LTRs	15	840	0.25
Total Class I	443	39,000	15.05
Class II			
Subclass 1			
MITE			
CACTA	7	1,084	0.08
Harbinger	25	4,243	0.46
Mutator	56	7,495	0.77
TcMar	23	2,721	0.28
hAT	53	5,780	0.68
Unclassified	28	4,737	0.51
Total MITEs	192	26,060	2.78
DNA (non-MITE)			
CACTA	50	8,495	2.97
Harbinger	20	2,376	0.49
Mutator	166	13,946	1.95
TcMar	63	7,391	1.66
hAT	134	12,237	2.00
Unclassified	63	6,184	0.96
Total DNA	496	50,629	10.03
Subclass 2			
Helitron	13	2,291	0.62
Total Class II	701	78,981	13.43

Table 2. Polymorphisms detected by TEPID for each element annotation. MAF ≥ 5% refers to polymorphisms with a minor allele frequency above 5%. The percentage shown represents

1624 the proportion of each value relative to the total number of polymorphic elements detected for

1625 each element and each type (insertion/deletion).

1626

		TEs		MITEs		IRs	
Polymorphisms		**Amount**	**%**	**Amount**	**%**	**Amount**	**%**
Insertions	Total	5974	100.0	1859	100.0	351	100.0
	MAF ≥ 5%	946	15.8	336	18.1	42	12.0
	MAF < 5%	5028	84.2	1523	81.9	309	88.0
Deletions	Total	7855	100.0	2507	100.0	983	100.0
	MAF ≥ 5%	2256	28.7	601	24.0	197	20.0
	MAF < 5%	5599	71.3	1906	76.0	786	80.0

1627 **SUPPLEMENTARY MATERIAL**

1628 **Table S1.** Size distribution of Transposable Elements by Class annotated in *Fragaria vesca*,
1629 including both fragmented and intact TEs.

1630

1631 **Table S2.** ShortStack analysis of sRNA-seq data from *Fragaria vesca* leaves, immature (white
1632 stage), and ripe (red stage) fruits using annotated loci files for MITEs and IRs.

1633

1634 **Table S3.** Differential expression analysis of 24-nt siRNA-producing MITEs and adjacent
1635 genes (< 3 kb) in *Fragaria vesca*. Genomic coordinates, functional annotation and expression
1636 changes between leaf tissue, and immature and ripe stages are shown. Gene IDs associated
1637 with key roles in fruit development and metabolism are indicated in bold.

1638

1639 **Table S4.** Differential expression analysis of 24-nt siRNA-producing IRs and adjacent genes
1640 (< 3 kb) in *Fragaria vesca*. Genomic coordinates, functional annotation and expression
1641 changes between leaf tissue, and immature and ripe stages are shown. Gene IDs associated
1642 with key roles in fruit development and metabolism are indicated in bold.

1643

1644 **Table S5.** Differential gene expression analysis of all *F. vesca* genes across conditions: leaf
1645 tissue vs. immature stage, ripe stage vs. leaf tissue, ripe vs. immature stages, and AZA-treated
1646 vs. control strawberries.

1647

1648 **Table S6.** Resequenced genomes of *Fragaria vesca* analyzed in the present study, including
1649 accession identifiers, country and origin coordinates, read counts, and average chromosome
1650 coverage (excluding unassembled contigs). Accessions in italics were excluded from analysis
1651 due to low coverage; accessions in bold were included in the GWAS study.

1652

1653 **Table S7**. Summary of significant genome-wide associations between Transposable Element
1654 Polymorphisms (TEPs) and agronomic and volatilome traits across two harvest seasons (H16
1655 and H17) in *Fragaria vesca*. Associations described in the Results section are highlighted in
1656 bold.

1657

1658 **Figure S1. Annotation and classification of transposable elements (TEs) in the *Fragaria***
1659 ***vesca* v4.0.a2 reference genome. (A)** Representation of the pipeline used in this work for
1660 transposon annotation. Software employed are highlighted in blue boxes, while input and
1661 output files are shown in pink boxes, with their formats in parentheses. **(B)** Distribution of TE

1662 superfamilies as annotated by EDTA (left), by DeepTE (center), and by TEtrimmer (right) for
1663 the *F. vesca* v4.0 genome.
1664
1665 **Figure S2. The average length of intact TEs and all annotated TEs in *F. vesca* genome**
1666 **(including fragmented TEs).**
1667
1668 **Figure S3. Distribution of intact TE superfamilies in the *Fragaria vesca* v4.0.a2**
1669 **annotation based on their distance from a gene**. Categories include: within a gene
1670 ('intragenic'), within 1 kb ('1000'), between 1 and 3 kb ('3000'), and more than 3 kb ('>3000').
1671 Left: Total counts by superfamily. Right: Proportion by superfamily.
1672
1673 **Figure S4. (A)** Genome distribution of IRs and MITEs. **(B)** Proportion of IRs classified as
1674 'Other', excluding MITE-type elements. Values are shown as percentages relative to the total
1675 number of IRs in the 'Other' category (16.1%).
1676
1677 **Figure S5. (A)** Metagene profile of 24-nt siRNAs mapping to MITEs and IRs located at
1678 different ranges of distance to the closest protein-coding gene, in leaf (green), and immature
1679 (gray) and ripe fruits (pink). Plots show MITEs/IRs scaled from the start to the end plus 2,000
1680 bp to each side. sRNA-seq replicates are plotted together. **(B)** Metagene profile of CpG, CHG,
1681 and CHH DNA methylation at MITEs/IRs located at different ranges of distance to the closest
1682 protein-coding gene, in leaf (green), and immature (gray) and ripe fruits (pink). Plots show
1683 MITEs/IRs scaled from the start to the end plus 2,000 bp to each side. Individual BS-seq
1684 replicates are plotted.
1685
1686 **Figure S6.** Genomic region of the *F. vesc*a v4 assembly containing the MITE Mutator locus
1687 near FvH4_7g18550 and FvH4_7g18570 genes. IRs and MITEs identified in this region are
1688 also displayed.
1689
1690 **Figure S7. Conservation of the MITE Mutator from *Fragaria vesca* in the orthologous**
1691 **genomic regions of commercial strawberry species *Fragaria chiloensis* and *Fragaria* ×**
1692 ***ananassa*. (A)** Local alignment of *F. vesca* MITE Mutator with the MITE located near the
1693 FvH4_7g18570 ortholog in *F. chiloensis* (left) and *F.* × *ananassa* (right). **(B)** Minimum free
1694 energy (MFE) RNA secondary structure predicted by RNAfold server of each MITE. The color
1695 scale shows base-pairing probability 0 (purple) to 1 (red).
1696

1697 **Figure S8.** Region of the *F. vesca* v4.0.a2 reference genome containing the FvH4_6g46420
1698 **(A)**, FvH4_2g30230 **(B),** and FvH4_5g23180 **(C)** loci displaying the IR and MITE elements
1699 annotated in the region and expression and epigenetic profiles. (I) 24-nt siRNAs mapping to
1700 the genomic regions as determined by sRNA sequencing of leaf (green), immature (gray) and
1701 ripe (pink) tissues. Replicates are plotted together. (II) Expression of each gene in leaf (green),
1702 immature (gray) and ripe (pink) tissues as well as strawberries treated with 5-azacytidine (5-
1703 AZA) and its control, measured by RNA-seq. (III) Cytosine DNA methylation in CG, CHG, and
1704 CHH contexts in leaf (green), immature (gray) and ripe (pink) fruits. The average of individual
1705 BS-seq replicates is plotted for each condition.
1706
1707 **Figure S9. (A)** Representation of the pipeline used in this study. The main inputs are shown
1708 at the top, while the main outputs are placed at the bottom. File formats are indicated in
1709 parentheses. **(B)** Representation of the TEPID pipeline used for identifying transposable
1710 element insertion polymorphisms (TIPs) and absence polymorphisms (TAPs).
1711
1712 **Figure S10. Enrichment analysis of TE insertion and deletion polymorphisms across**
1713 **genomic features in *Fragaria vesca* v4.0.a2. (A)** Genome-wide enrichment of all
1714 polymorphisms, insertions, and deletions across annotated genomic features. **(B)** Enrichment
1715 profiles of TE-associated polymorphisms across major TE superfamilies. Values are shown
1716 as log2(observed/expected), where 0 indicates a random distribution relative to genomic
1717 fraction. Positive values (red) indicate enrichment, whereas negative values (blue) indicate
1718 depletion. Bubble size reflects the absolute magnitude of enrichment or depletion.
1719
1720 **Figure S11. Principal Component Analysis (PCA) of the subset of *Fragaria vesca***
1721 **accessions used for GWAS analysis.** Each point represents a single accession, with colors
1722 indicating their respective country of origin.
1723
1724 **Figure S12. Genomic region of the *F. vesca* v4.0.a2 reference genome harboring**
1725 **selected TEP polymorphisms on chromosome Fvb3 associated with variation in: (A)** 6-
1726 methyl-5-hepten-2-one levels. (I) Coverage of mapped reads (light green) and split reads (dark
1727 green) in some accessions carrying the polimorphism (from top to bottom: ES14, ES-13,
1728 UK11, and IT15). (II) Coverage of mapped reads (light blue) and split reads (dark blue) in non-
1729 carriers accesions (from top to bottom: FIN10, FIN12, FIN6, and FIN51). **(B)** γ-decalactone
1730 levels. (I) Coverage of mapped reads (light green) and split reads (dark green) in some
1731 accessions carrying the polimorphism (from top to bottom: DK1, ES14, IT6, and IT15). (II)

1732 Coverage of mapped reads (light blue) and split reads (dark blue) in non-carriers accesions
1733 (from top to bottom: FR1, FIN25, FIN51, and NOR131-2).
1734
1735 **Figure S13. Addition TE polymorphisms linked to volatile compound variation in a**
1736 **collection of *F. vesca* accessions.** **(A)** Manhattan plot of TEP-GWAS for methyl 2-
1737 aminobenzoate in harvest season H16. The dotted gray line marks the significance threshold
1738 determined by the Bonferroni correction, and the significantly associated polymorphism is
1739 highlighted. **(E)** Observed and expected distribution of *p* values for TEP-GWAS for methyl 2-
1740 aminobenzoate. **(F)** Violin plot showing the relative abundance of methyl 2-aminobenzoate in
1741 accessions carrying the MITE Mutator deletion (+, coral) versus non-carriers (-, steel blue).
1742 Horizontal lines in (-) and (+) indicate the mean. **(D)** Manhattan plot of TEP-GWAS for γ-
1743 decalactone across both harvest seasons (H16 and H17, represented by circles and triangles,
1744 respectively). The dotted gray line marks the significance threshold determined by the
1745 Bonferroni correction, and the significantly associated polymorphisms are highlighted. **(E)**
1746 Observed and expected distribution of *p* values for TEP-GWAS for γ-decalactone. **(F)** Violin
1747 plot showing the relative abundance of γ-decalactone in accessions carrying the MITE deletion
1748 (+, coral) versus non-carriers (-, steel blue). Horizontal lines in (-) and (+) indicate the mean.
1749
1750