

1 **Title page:**

2 **Structural variation underlies functional diversity at methyl salicylate loci in tomato**

3 Manoj Sapkota¹, Lara Pereira¹, Yanbing Wang¹, Lei Zhang¹, Yasin Topcu¹, Denise Tieman²,

4 Esther van der Knaap^{1*}

5 ¹Institute of Plant Breeding, Genetics, and Genomics, University of Georgia, Athens, GA 30602,

6 USA

7 ²Horticultural Sciences, University of Florida, Gainesville, FL 32611, USA.

8

9 *Corresponding author:

10 E-mail: esthervanderknaap@uga.edu (EvdK)

11

12 **Full title:** Structural variation underlies functional diversity at methyl salicylate loci in tomato

13 **Short title:** Structural variation at methyl salicylate loci

Abstract

Methyl salicylate is an important inter- and intra-plant signaling molecule, but is deemed undesirable by humans when it accumulates to high levels in ripe fruits. Balancing the tradeoff between consumer satisfaction and overall plant health is challenging as the mechanisms regulating volatile levels have not yet been fully elucidated. In this study, we investigated the accumulation of methyl salicylate in ripe fruits of tomatoes that belong to the red-fruited clade. We determine the genetic diversity and the interaction of four known loci controlling methyl salicylate levels in ripe fruits. In addition to *Non-Smoky Glucosyl Transferase 1 (NSGT1)*, we uncovered extensive genome structural variation (SV) at the *Methylesterase (MES)* locus. This locus contains four tandemly duplicated *Methylesterase* genes and genome sequence investigations at the locus identified nine distinct haplotypes. Based on gene expression and results from biparental crosses, functional and non-functional haplotypes for *MES* were identified. The combination of the non-functional *MES* haplotype 2 and the non-functional *NSGT1* haplotype IV or V in a GWAS panel showed high methyl salicylate levels in ripe fruits, particularly in accessions from Ecuador, demonstrating a strong interaction between these two loci and suggesting an ecological advantage. The genetic variation at the other two known loci, *Salicylic Acid Methyl Transferase 1 (SAMT1)* and tomato *UDP Glycosyl Transferase 5 (SlUGT5)*, did not explain volatile variation in the red-fruited tomato germplasm, suggesting a minor role in methyl salicylate production in red-fruited tomato. Lastly, we found that most heirloom and modern tomato accessions carried a functional *MES* and a non-functional *NSGT1* haplotype, ensuring acceptable levels of methyl salicylate in fruits. Yet, future selection of the functional *NSGT1* allele could potentially improve flavor in the modern germplasm.

Author Summary

Tomato fruit release a wide variety of volatiles and these volatiles are responsible for its unique aroma. One of the important volatiles in tomato flavor is methyl salicylate. Methyl salicylate is important for plant defense, but tomato taste panels routinely associate this volatile with low liking as it features the characteristic odor of wintergreen. Several enzymes involved in the biosynthesis of methyl salicylate have been identified in tomato but the genetic diversity at the gene region has not yet been fully explored. In this study, we investigate the genetic variations at the four known gene regions controlling fruit methyl salicylate levels using a diverse collection ranging from red-fruited wild tomatoes to modern and heirloom varieties. We identified extensive genomic structural variations at two of the four loci. The genetic interaction between two loci determines the volatile levels. We also showed dramatic selection pressures ensued at these loci during evolution and domestication of the tomato. Our study greatly improves the understanding of the genetic basis of methyl salicylate production in tomato fruits and gives breeders far better tools to create agrosystem-specific adapted commercial germplasm.

Introduction

Plants synthesize and emit many volatile organic compounds that play key roles in reproduction, communication, and stress responses. The volatiles emitted from flowers allow plants to attract and guide pollinators to ensure fertilization and reproduction [1, 2], whereas aroma from fruits allure frugivores to assist in seed dispersal [3-5] and fend off pre- and post-harvest diseases [5-7]. Volatiles play important roles in intra- and inter-plant communications as these compounds facilitate the signaling about the danger to distant parts of same or to neighboring plants [8, 9].

58 Furthermore, volatiles can act as direct defenses, either by repelling the danger [10, 11], or
59 intoxicating herbivores and pathogens [12, 13]. Volatiles may also serve as signals to attract
60 herbivores' natural enemies such as parasitoids and predators [14-18] and even predatory birds
61 [19].

62 The complex tomato fruit volatile aroma is produced by derivatives from six major
63 biochemical pathways: (i) fatty acid or lipid-derived; (ii) branched chain amino acid (BCAA)-
64 derived such as leucine (Leu) and isoleucine (Ile); (iii) phenylalanine (Phe)-derived including
65 phenylpropanoids and phenolics; (iv) terpenoid-derived; (v) carotenoid-derived; and the (vi)
66 acetate ester volatiles [20-22]. Tomato fruit release a wide variety of volatiles, with emission rates
67 peaking at ripening [23]. These volatiles comprise only 10^{-7} to 10^{-4} of the fresh fruit weight and
68 are responsible for its unique aroma [24]. In general, well-liked tomatoes feature a complex volatile
69 aroma that is high in certain volatiles and low in others, and properly balanced in yet other volatiles
70 [20, 25].

71 Red-fruited tomato is edible and evolved in the tomato clade of the *Solanum* genus in the
72 Solanaceae family. The red-fruited *S. pimpinellifolium* (SP) last shared a common ancestor with
73 the green-fruited species approximately two million years ago [26, 27]. Within the red-fruited
74 clade, SP evolved into wild *S. lycopersicum cerasiforme* (SLC) that then became domesticated
75 into *S. lycopersicum lycopersicum* (SLL) around seven thousand years ago [28, 29]. The huge
76 variation in volatile production in tomato from SP to SLL indicates that its domestication placed
77 strong selection pressures on this trait [25]. One of the important volatiles in tomato flavor is
78 methyl salicylate which is produced in the Phe-derived pathway. Tomato taste panels routinely
79 associate this volatile with low liking as it features the characteristic odor of wintergreen. H₂
80 however, some varieties with relatively high moderate to high levels of methyl salicylate are rated

~~well high~~ for overall liking. -This may be the result of concentrations of -or interactions with other volatiles resulting in overall good flavor -[25, 30].- While the levels of methyl salicylate are generally low in red fruited tomato including heirlooms, some produce intermediate levels of the volatile in ripe fruit, suggesting that selections for yield and disease resistance were at the expense of better flavor, a common concern among consumers [25, 31]. Alternatively, low levels of methyl salicylate in ripe fruit may be beneficial to the plant.

Several enzymes involved in the biosynthesis of methyl salicylate and its glucoside conjugation have been identified in tomato (Fig 1) [32-35]. Methyl salicylate is synthesized from salicylic acid, which is produced from the isochorismate and the Phe-derived pathways [36]. The process of methylation of salicylic acid into methyl salicylate is catalyzed by the enzyme SALICYLIC ACID METHYLTRANSFERASE 1 (*SISAMT1*; *Solyc09g091550*) [32]. Another group of enzymes, METHYLESTERASE 1-4 (*SIMES1-4*; *Solyc02g065240-80*), demethylates methyl salicylate to salicylic acid to maintain homeostasis [35]. The methyl salicylate can also be glycosylated into glycoconjugate compounds for transport and storage in the vacuole [37]. The two major enzyme families involved in the formation and degradation of glycosylated methyl salicylate are: glycosyltransferases and glycosidases. Glycosyltransferases (eg. *SIUGT5*; *Solyc01g095620*) catalyze the formation of diglycoside conjugates [33]. These diglycoside conjugates are hydrolyzed to methyl salicylate by glycosidases. One glycosyl transferase, *Non-Smoky Glycosyltransferase 1* (*NSGT1*; *Solyc09g089585*), encodes a fruit ripening-induced enzyme that adds a glucose moiety to the methyl salicylate diglycoside conjugate resulting in a triglycoside. This triglycoside conjugation is irreversible and prevents the emission of the volatile upon tissue disruption and ripening [21, 34]. Of the enzymes in the methyl salicylate biosynthesis and

103 conjugation pathway, loss of function in either *SIMES1-4* or *NSGT1* results in high methyl
104 salicylate in fruits [34, 35].

105 *NSGT1* is known for the glycosylation of guaiacol, a smoky volatile that, similarly to
106 methyl salicylate, negatively affects consumer liking. Based on genome structural variations (SV),
107 *NSGT1* is represented by five haplotypes that are grouped into functional and non-functional
108 categories based on the presence of a functional enzyme [38]. Similarly, *MES* also exhibits SVs
109 [35], but the diversity at this locus has not yet been fully explored in tomato. The genetic diversity
110 at the other two known genes, *SISAMT1* and *SIUGT5*, has also not been explored in red-fruited
111 tomato. The genetic characterization of these loci can provide evolutionary insights and lead to the
112 identification of haplotypes that might be beneficial for crop improvement.

113 In this study, we analyzed a unique and genetically well-characterized tomato collection of
114 fully wild SP, semi-domesticated SLC and ancestral SLL landraces from South and Central
115 America (Varitome Collection) as well as a subset of cultivated accessions that includes heirloom
116 and modern types (SLL_CUL). Collectively, they are referred to as red-fruited tomato in this study.
117 We investigated the genetic diversity at the known genes controlling methyl salicylate levels in
118 tomato fruits and uncovered extensive genome structural variation (SV) underlying the trait in the
119 red-fruited tomato clade. We also show that different selection pressures ensued at these loci
120 during domestication.

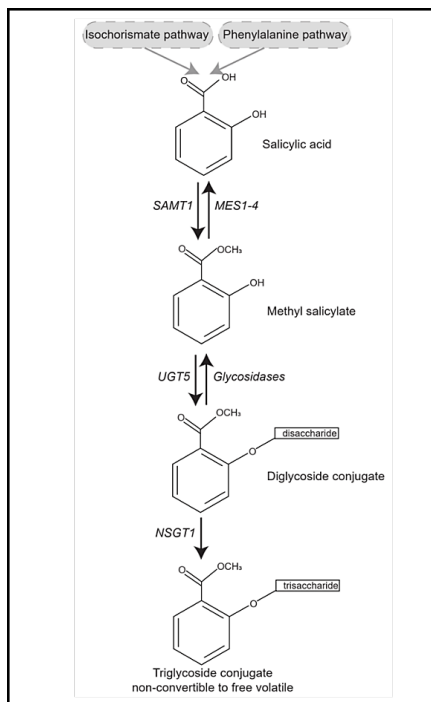


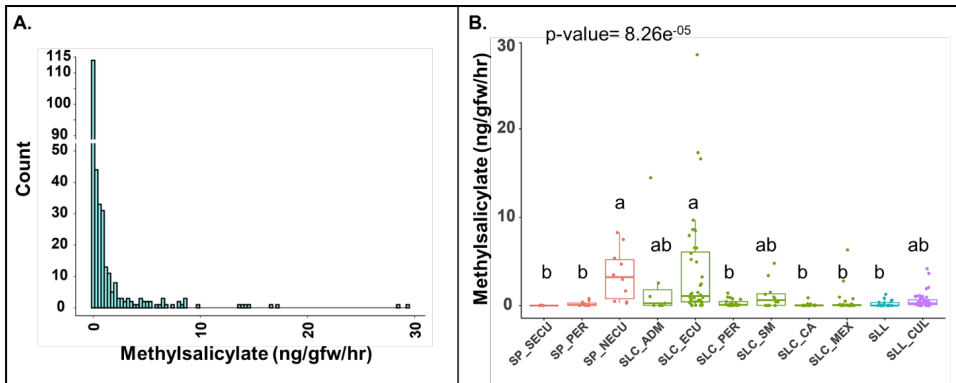
Fig 1. The biosynthesis and glucoside conjugation pathway of methyl salicylate in tomato. The gray boxes represent the two pathways leading to salicylic acid. The arrows represent the direction of the reaction. The enzyme catalyzing the reaction is listed next to the arrow. The genes names are: *SAMT1*: Salicylic Acid Methyltransferase 1; *MES1-4*: Methyltransferase 1-4; *UGT5*: UDP-glycosyltransferase 5; *NSGT1*: Non-smoky Glycosyltransferase 1 [32-35]

128 **Results**

129 **Distribution of methyl salicylate levels in ripe, red-fruited tomato fruits**

130 Quantification of fruit methyl salicylate in 166 accessions from the Varitome Collection and 143
131 cultivated types showed that the levels ranged from 0.003 nanogram per gram fruit weight per hour
132 (ng/gfw/hr) to 29.360 ng/gfw/hr, representing a 9000-fold difference (Fig 2A). Most accessions
133 showed less than 1 ng/gfw/hr of methyl salicylate in the fruits. Significantly higher levels were
134 found in SP from Northern Ecuador and SLC from Ecuador, whereas ancestral landraces, SLL,
135 featured low levels of the volatile. Most of the SLL_CUL produced low levels of methyl salicylate
136 with a few exceptions. We observed variation in other sub-populations, but overall, the fruits
137 emitted low levels of the volatile (Fig 2B).

138
139 **Fig 2.** Methyl salicylate levels in tomato fruits in the red-fruited tomato germplasm. **(A)** Histogram
140 representing the distribution of methyl salicylate in ripe tomato fruits from the Varitome Collection
141 and heirlooms. X-axis represents the methyl salicylate levels in fruits whereas, y-axis denotes the
142 count. **(B)** Methyl salicylate levels in tomato accessions grouped by sub population. SP_SECU:
143 SP from Southern Ecuador, SP_PER: SP from Peru, SP_NECU: SP from Norther Ecuador,



SLC_ADM: SLC admixture, SLC_ECU: SLC from Ecuador, SLC_PER: SLC from Peru, SLC_SM: SLC from San Martin, SLC_CA: SLC from Central America, SLC_MEX: SLC from Mexico, SLL: SLL in Varitome Collection and SLL_CUL: modern and heirloom accessions. The letters in the boxplots indicate the significant differences among different sub population evaluated by Duncan's test ($\alpha < 0.05$). The x-axis represents different sub population whereas, y-axis denotes the methyl salicylate levels in fruits.

150

151 **SNP, INDEL and SV-based GWAS of methyl salicylate levels in ripe tomato fruits**

To investigate the genetic bases of the variation in methyl salicylate levels, we performed a genome wide association study (GWAS) using the whole genome variant information of the Varitome Collection. We identified several GWAS loci associated with fruit methyl salicylate levels (Fig 3, S1 Table) using Single Nucleotide Polymorphisms (SNPs), Insertion-Deletions (INDELs) and Structural Variants (SVs). The identified GWAS loci included *MES* on chromosome 2 (chr 02), *NSGT1* on chr 09, and *SAMT1* also on chr 09. For *MES*, the locus was identified by all three variant types, namely a 4.7 kb deletion at position SL4.0ch02:34,464,833, a SNP at SL4.0ch02:34,895,803, and an INDEL at SL4.0ch02:34,943,905 (Figs 4 A-C). The *MES* locus contains four *Methylesterase* genes (*Solyc02g065240-80*) and the 4.7 kb deletion partially spans the third exon of *Solyc02g065260* (*SlMES3*), whereas the SNP and INDEL mapped 400 kb and 467 kb downstream of the *MES* locus, respectively. The accessions carrying the 4.7 kb deletion produced significantly higher amounts of methyl salicylate compared to the accessions without the deletion (Fig 4A) and the ~~locus-variant~~ explained 28% of total methyl salicylate variation in the Varitome Collection. Similar to the SV at *MES*, accessions having alternate alleles of the associated SNP and INDEL produced significantly higher levels of methyl salicylate in tomato

167 fruits (Figs 4B and C). For *NSGTI*, the locus was identified by a SNP at SL4.0ch09:65,404,388
168 (4.3 kb downstream of *NSGTI*) and an INDEL at SL4.0ch09: 65,401,256 (7.4 kb downstream of
169 *NSGTI*). explaining ~40% of the variation in the population. These overlapping GWAS loci carried
170 three alleles: the reference allele, the alternate (SNP or INDEL) allele, or the 15 kb deletion.
171 Accessions with the deletion at the locus showed significantly higher levels of methyl salicylate
172 compared to accessions without the deletion (Figs 4D and 4E). For *SAMTI*, the locus was detected
173 by a SNP at SL4.0ch09: 67,197,164 which was located 293 kb downstream of the gene. The
174 accessions carrying the reference allele (GG) produced lower levels of methyl salicylate compared
175 with the alternative allele (GA) (Fig 4F). Similar to *SAMTI*, the closest significant variant for
176 *UGT5* on chr 01 was an INDEL (SL4.0ch01: 85,548,458) that mapped to approximately 6.5 Mb
177 downstream the gene. For this variant, all accessions with reference allele produced low methyl
178 salicylate levels in fruits, whereas lines with a nucleotide insertion showed significantly higher
179 levels of the volatile (Kruskal-Wallis test p-value: $6.4e^{-06}$) (Fig 4G).

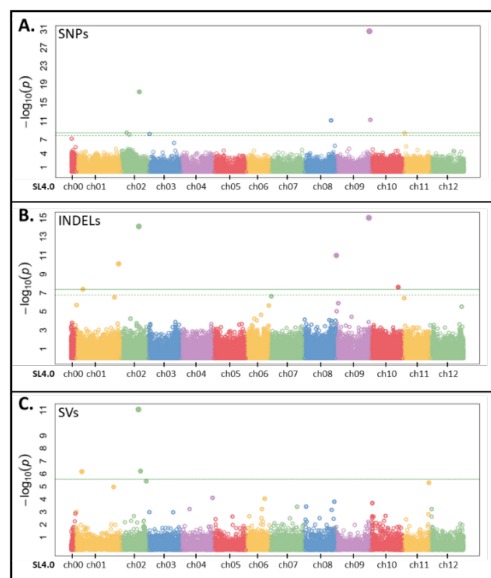


Fig 3. Genome wide association study (GWAS) for methyl salicylate levels in the Varitome Collection. **(A)** GWAS based on SNPs. **(B)** GWAS based on INDELs. **(C)** GWAS based on SVs. The x-axis represents different chromosomes and y-axis denotes the $-\log_{10}(\text{p-value})$ of the variants.

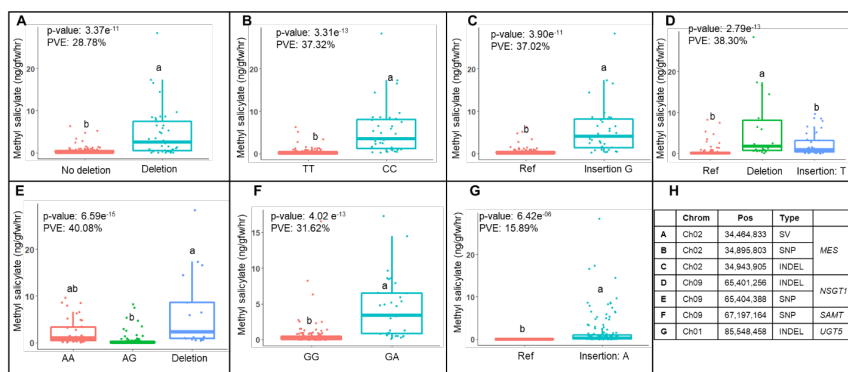


Fig 4. Phenotypic distribution of methyl salicylate levels at the alleles of significant loci at *MES*, *NSGT1*, *SAMT1*, and *UGT5*. Boxplots representing the distribution of methyl salicylate levels

among different alleles of the **(A)** SV, **(B)** SNP, and **(C)** INDEL identified at *MES* locus from GWAS. Boxplots representing the distribution of methyl salicylate levels among different alleles of the **(D)** INDEL and **(E)** SNP identified at *NSGT1* locus from GWAS. **(F)** Boxplots representing the distribution of methyl salicylate levels among different alleles of SNP identified at *SAMT1* locus from GWAS **(G)** Boxplots representing the distribution of methyl salicylate levels among different alleles of the INDEL identified at *UGT5* locus from GWAS. **(H)** The table provides the information of significant variants identified from GWAS along with the known gene loci. The reference alleles are colored in red. The p-value is obtained from Kruskal-Wallis test. The percentage of variation explained (PVE) is calculated using respective variants in the ANOVA model. The letters in the boxplots indicate the significant differences among different alleles evaluated by Duncan's test ($\alpha < 0.05$).

198

199 **Genome landscape of *MES*, *SAMT1* and *UGT5* loci in the red-fruited tomatoes**

Previously, five haplotypes of *NSGT1* were identified based on the SVs at the locus of which haplotype IV and V were associated with higher methyl salicylate levels [38]. To gain more insight about the haplotype diversity of the other genes in this pathway, we investigated the genomic landscape at *MES*, *UGT5* and *SAMT1* loci with the focus on SVs. For *MES*, we compared the genomic sequences of the three highest methyl salicylate accumulating accessions (BGV006852: 28.388 ng/gfw/hr, BGV006931: 17.316 ng/gfw/hr and BGV006148: 14.472 ng/gfw hr) and the three lowest accumulating accessions (BGV007151: 0.0094 ng/gfw hr, BGV008098: 0.0097 ng/gfw/hr and BGV007867: 0.0099 ng/gfw/hr). The high methyl salicylate producing accessions carried several deletions at *MES* and multiple mismatches when compared to the Heinz1706 reference genome. On the other hand, the genomic landscape of the low methyl salicylate producing accessions corresponded more closely to the Heinz1706 reference genome (S1 Fig).

211 Further comparisons based on SVs (S2 Table) showed additional haplotypes in red-fruited tomato
212 to a total of nine (Fig 5, S3 Table). We next investigated each of the predicted genes at *MES* to
213 ascertain functionality with respect to the potential to encode a full-length protein (S2 Fig). The
214 results showed the predicted functional genes at the locus ranged from zero (haplotype 9) to four
215 (haplotype 7). *MES1* was likely functional in every haplotype except 9. *MES3* was truncated in
216 several haplotypes, missing 97 amino acids from its C terminal end except in haplotype 1. It is
217 known that the truncated allele of *MES3* retains at least partial enzyme activity and is therefore
218 considered functional [35]. For *MES2* and *MES4*, several haplotypes appeared to carry a functional
219 allele for one or both genes. *MES4* was either full length or was truncated resulting from a 48 bp
220 deletion in the first exon (S2 Table).

221 We next constructed a phylogenetic tree using the genome sequence of *MES* including 10
222 kb that flank the four genes at the locus (Fig 6). Haplotype 1 was found only in a few accessions
223 that did not cluster with any other haplotype (Fig 6, and S4 Table). Haplotype 3 was also distinct
224 and differed from haplotype 1 by a transposon insertion at the C terminal end of *SIMES3* resulting
225 in a truncated version of the gene and a remaining protein segment that was annotated as
226 *Solyc02g065270* (S3 Fig). Haplotypes 2 and 4 were also more ancestral, both showing a deletion
227 in the promoter of *SIMES1* compared to other haplotypes. Haplotype 2 carries the 48 bp deletion
228 in the first exon of *SIMES4*, truncating the protein to 119 amino acids instead of 279 amino acids,
229 whereas haplotype 4 has an intergenic deletion between *SIMES3* and *SIMES4*. Haplotypes 2 and 4
230 were clustered together implying a close evolutionary relationship in addition to a similar SV
231 structure. The most common haplotypes were 6 and 7, found in 122 and 120 accessions,
232 respectively. Haplotype 6 included the reference genome Heinz1706 and M82 and haplotype 7
233 included Moneymaker. Compared to haplotype 5, haplotypes 6 and 7 carried transposon insertions

234 in one of the introns of *SIMES2* and in the intergenic region between *SIMES2* and *SIMES3*.
 235 Haplotype 7 was the only haplotype for which all four genes were predicted functional, whereas
 236 haplotype 6 carried a premature stop codon in *SIMES2*, likely rendering this gene non-functional
 237 (S4 Table). Therefore, haplotype 6 appeared to be derived from haplotype 7. Haplotype 8 clustered
 238 in a separate subclade and lacked *SIMES3* as well as a non-functional *SIMES4* due to a frameshift
 239 (at SL4.0ch02:34,471,277) and three missense (at SL4.0ch02:34,471,287, SL4.0ch02:34,471,546,
 240 and SL4.0ch02:34,475,508) mutations in the latter gene. The last haplotype, haplotype 9 consisted
 241 of several large deletions at the locus resulting in truncation, internal deletions, and elimination of
 242 all four *MES* genes, and was found only in 13 SLL_CUL accessions (e.g., Rio Grande in S4 Fig).
 243 Clearly, the huge variation in haplotype diversity even among the fully wild accessions, suggested
 244 that selection predating domestication shaped the genome structure at this locus.

245 For *SAMT*, the genome sequence showed three tandemly duplicated *Methyltransferase*
 246 genes (*Solyc09g091530-50*). The three genes were found in all accessions and the locus did not
 247 show extensive SV except for a single 255 bp deletion (SL4.0ch09: 66,899,792-66,900,047) in the
 248 promoter of *SISAMT1* (*Solyc09g091550*) that was found in 27 accessions. The *SAMT1*
 249 (*Solyc09g091550*) is known to regulate methyl salicylate levels in tomato fruit [32]. Since we did
 250 not identify any other large structural variants, we proceeded with the haplotype analysis using
 251 SNPs and small INDELs at the 21.3 kb locus (S5 Fig and S5 Table). A total of 530 variants were
 252 identified and were comprised of 399 SNPs and 131 INDELs. Among them, 83% of the variants
 253 were found in regulatory regions. Of the coding region variants, the known *SAMT1* gene
 254 *Solyc09g091550* carried two synonymous and seven missense mutations compared to the Heinz
 255 1706 reference genome for which one or more were found in 34 accessions. This means that
 256 majority of red-fruited accessions in this study encoded the same *SAMT1* protein sequence as

257 Heinz1706. For another *SAMT* gene, *Solyc09g091530*, ten synonymous, nine missense, two
 258 frameshift, and two nonsense mutations were found compared with the Heinz1706 reference
 259 genome. The frameshift and nonsense mutations resulted in null alleles that were found in seven
 260 SP, three SLC and one SLL. Lastly, we found six synonymous, seven missense and one frameshift
 261 mutation in *Solyc09g091540*. The frameshift mutation was found in only three accessions (2 SLC
 262 and 1 SLL) and these also carried the predicted null mutations in *Solyc09g091530*. The haplotype
 263 analysis for *SAMT1* using the 530 variants at the 21.3 kb locus showed four alleles (S5a Fig, S3
 264 Table and S5 Table). Haplotype I was found in 19 SP, 13 SLC and three SLL_CUL and had the
 265 255 bp deletion in the upstream regulatory region of *SISAMT1* as well as six of the 11 accessions
 266 with null mutations in *Solyc09g091530*. Haplotype II was found in eight SP and two SLC,
 267 including two SP with a premature stop codon in *Solyc09g091530*. Haplotype III appeared in 106
 268 SLC, 18 SLL and 37 SLL_CUL accessions, including three accessions (two SLC and one SLL)
 269 with the frameshift mutation in *Solyc09g091530* and *Solyc09g091540*. Haplotype IV was identical
 270 to the Heinz1706 reference genome and found in the majority of SLL_CUL. Haplotype I of
 271 *SISAMT* was the most ancestral haplotype based on the phylogenetic tree constructed from the
 272 consensus genomic sequence of the locus and included most of the SP accessions (S6 Fig). Other
 273 haplotypes, II and III were found to cluster separately under distinct clades in the phylogenetic
 274 tree. The phylogenetic analysis did not include haplotype IV since it was found only in SLL_CUL.

275 At the *UGT5* locus, the only SV was a 403 bp deletion (SL4.0ch01:79,099,508) found in
 276 50 accessions and located 8 kb upstream of the gene. We identified 132 SNPs and 29 small
 277 INDELs at the 5.7 kb *UGT5* locus compared to the Heinz1706 reference genome. Most of the
 278 variants (150) were found in the regulatory regions and only 11 variants (six missense, four
 279 synonymous and one frameshift mutation) were found in the gene. The frameshift mutation was a

likely null allele and was found in 41 accessions. Haplotype analysis using the variants in the 5.7 kb region resulted into four distinct haplotypes (S7A Fig, S3 Table and S6 Table). Haplotype I was found in nine SP accessions and carried one missense and one synonymous mutation in the *SIUGT5* compared to Heinz reference genome. Haplotype II was found in 18 SP, 23 SLC and two SLL_CUL. All the accessions in this haplotype carried two missense mutations and a frameshift mutation in the gene compared to Heinz1706. Furthermore, the accessions in haplotype I and II also carried the 403 bp deletion. Haplotype III included the Heinz1706 reference genome and was found in 31 SLC, 11 SLL, and 59 SLL_CUL. Haplotype IV and was the most common haplotype and was found in all accessions except SP. Accessions belonging to haplotype IV were similar to Heinz1706 within the gene region whereas Heinz1706 upstream regulatory region was more similar to haplotype III accessions. Phylogenetic analysis showed haplotype II as the most ancestral haplotype (S8 Fig). Haplotype I was clustered separately in the tree followed by a clade with both haplotypes III and IV.

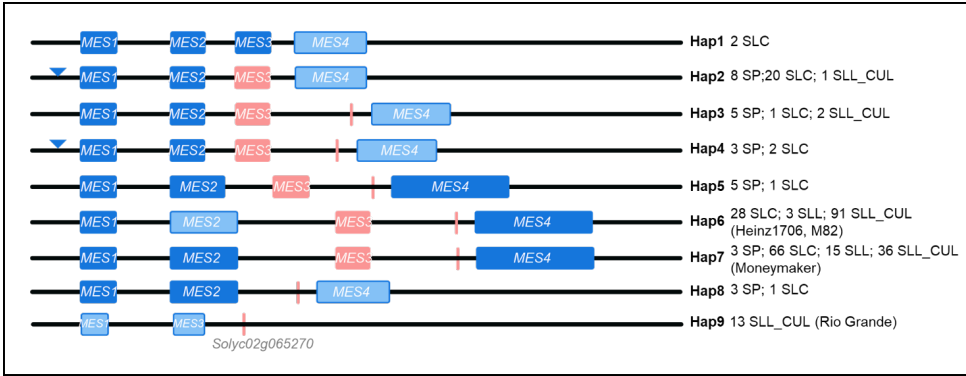


Fig 5. Schematic representation of nine haplotypes at the *MES* locus. Blue inverted triangles represent deletions, dark blue boxes indicate likely functional genes, light blue boxes indicate non-

functional genes whereas the light red boxes represent truncated genes. Next to each haplotype are the accessions belonging to the haplotype.

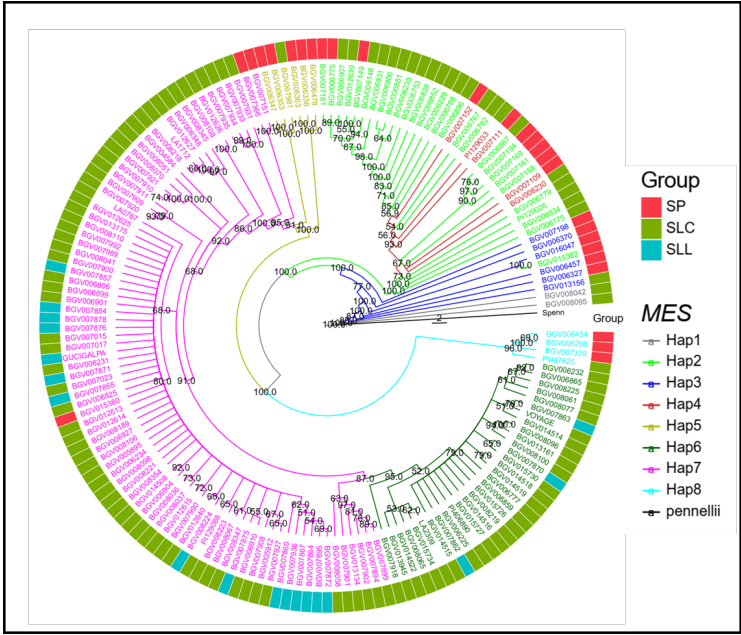


Fig 6. Phylogeny tree constructed using consensus genomic sequence of *MES* locus and flanking regions from the Varitome Collection along with green fruited *S. pennellii* (Spenn). Different colors of accessions and branches represents different *MES* haplotypes. The outer concentric circle represents the grouping of accessions (SP, SLC and SLL). The numbers on the branches represent the bootstrap values. *S. pennellii* is used as outgroup.

305 **Association of the haplotypes at *MES*, *NSGT1*, *SAMT1* and *UGT5* loci with methyl**
306 **salicylate levels**

307 We evaluated the association of the haplotypes at these the loci with methyl salicylate levels,
308 starting with *MES* and *NSGT1*. Among the nine *MES* haplotypes, haplotype 2 produced higher
309 levels of methyl salicylate than the other haplotypes suggesting a loss-of-function allele (Fig 7A).
310 Haplotype 4 and haplotype 9 also produced relatively high levels of methyl salicylate (Haplotype
311 4: PI129033: 16.61 ng/gfw/hr and haplotype 9: TS-209: 29.36 ng/gfw/hr), however these
312 haplotypes represented only a few accessions (n=5 in haplotype 4 and n=13 in haplotype 9). Most
313 accessions with haplotypes 1, 3, 5, 6, 7, or 8 produced low levels of methyl salicylate, with
314 exceptions, suggesting the presence of functional alleles that converted the volatile into salicylic
315 acid. The variation at *NSGT1* is represented by five haplotypes [38]. These haplotypes are
316 classified in two groups based on expected functionality: the functional *NSGT1* haplotypes I, II
317 and III and the non-functional *NSGT1* haplotypes IV and V. The non-functional *NSGT1* haplotypes
318 produced significantly higher levels of methyl salicylate compared to the functional *NSGT1*
319 haplotype (Fig 7B). The combined effect of the *NSGT1* and *MES* haplotypes showed a strong
320 interaction between the loci. *MES* haplotype 2 and a non-functional *NSGT1* haplotype IV or V
321 resulted in the highest methyl salicylate levels (Fig 7C). Previous results from a biparental cross
322 also supported the notion that *MES* and *NSGT1* strongly interact with respect to the methyl
323 salicylate levels [35]. Other haplotype combinations carried too few accessions to reliably
324 ascertain association with volatile levels, although a trend between haplotype 4 and *NSGT1* V
325 suggested that this *MES* haplotype was non-functional.

326 The four *SAMT1* haplotypes were not statistically different in methyl salicylate
327 accumulation (S5B Fig). Even so, *SAMT1* haplotype II seemed to produce higher levels of methyl

328 salicylate but was represented by too few accessions to significantly associate volatile levels with
329 the haplotype. The 255 bp deletion showed no association with the volatile levels and thus was
330 unlikely causative (S5C Fig). Interestingly, the 11 accessions carrying a null mutation in
331 *Solyc09g091530* were found to produce low levels of methyl salicylate contrary to expectations
332 (S5D Fig). Similar to *SAMT1*, there was no significant difference among the *UGT5* haplotypes
333 (S7B Fig). *UGT5* haplotype I showed a trend of higher methyl salicylate levels but was found in
334 only nine accessions which is too few for a meaningful analysis. We also did not find a significant
335 association of the 403 bp deletion with the methyl salicylate levels in fruits (S7C Fig). Overall, we
336 could not show ~~functional~~-association of a causative variant at *SAMT1* and *UGT5* ~~haplotypes~~
337 with methyl salicylate levels. This may be, ~~in part~~ due in part to the large effect of *MES* and
338 *NSGT1* on volatile levels, and ~~in part due to~~ the relatively low number of accessions that carried
339 the same haplotype at the two major loci to allow a thorough characterization of the two minor
340 loci.

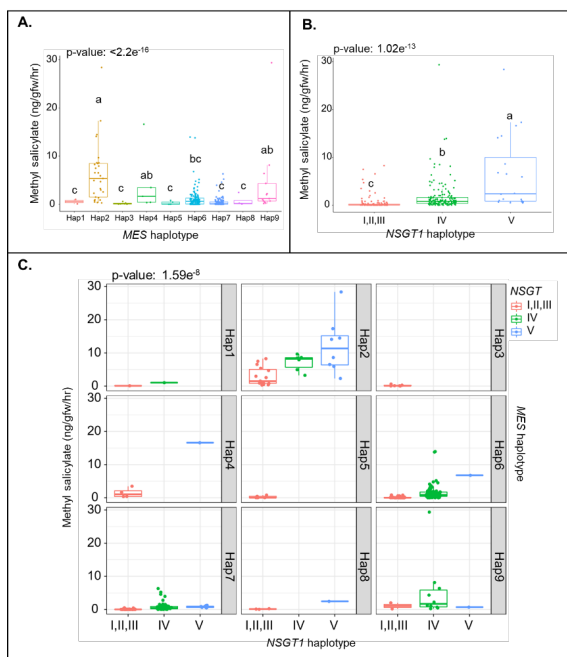


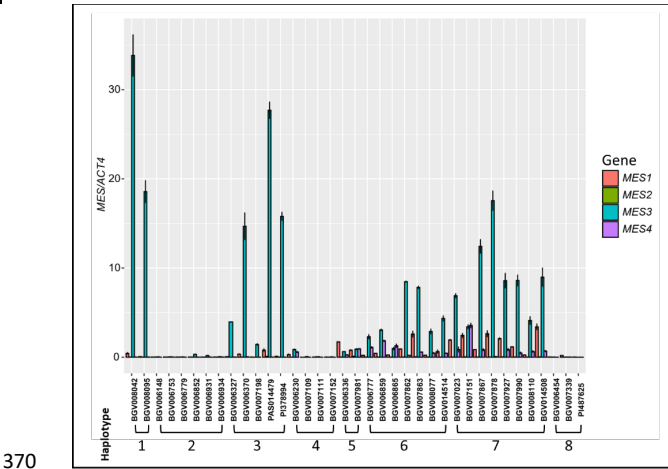
Fig 7. Distribution of methyl salicylate in the Varitome Collection and cultivated tomato. **(A)** *MES* haplotypes, **(B)** *NSGT1* haplotypes, and **(C)** across different *MES* and *NSGT1* haplotypes. The letters in the boxplots indicate the significant differences among different haplotypes evaluated by Duncan's test ($\alpha < 0.05$).

Expression of *SIMES1-4* in *MES* locus

All haplotypes at the *MES* locus, except haplotype 9, carried putative functional copies of one to several *MES* genes. Especially *SIMES1*, in all accessions except haplotype 9, appeared to be full length and functional (Fig 5). Since the function of *SIMES1* is known [35], we sought to investigate the putative role of other members at the locus in methyl salicylate production in ripe tomato fruit. We investigated their expression by Real-Time Quantitative Reverse Transcription Polymerase

Chain Reaction (qRT-PCR) (Fig 8). We selected 37 Varitome accessions that represented eight *MES* haplotypes (Haplotype 1-8). In accessions belonging to haplotype 1 and 3, *SIMES3* was well expressed in ripe tomato fruit, whereas the expression of the other *MES* genes was low or undetectable. Accessions belonging to these haplotypes produced very low amounts of methyl salicylate, ranging from 0.03 to 1.03 ng/gfw/hr (Fig 7A), implying a functional allele. In haplotype 2, 4, and 8, the expression of all four *MES* genes was low to non-detectable. Especially haplotypes 2 and 4 were associated with high accumulation of methyl salicylate likely due to the lack of expression of any of the *MES* genes. The two accessions representing haplotype 5 showed slightly higher expression of *SIMES1*, *SIMES3* and *SIMES4*, compared to haplotype 2, 4, and 8. In general for haplotype 6 and 7, *SIMES3* was highly expressed followed by *SIMES1* and last *SIMES4*. The expression of *SIMES2* was undetectable in nearly all accessions suggesting that this gene was not important for methyl salicylate accumulation in the fruit. As expected, the expression patterns of *MES* were negatively correlated with the fruit methyl salicylate levels. Haplotypes producing high level of methyl salicylate: haplotype 2 and 4, showed low to non-detectable expression of *MES* genes. On the other hand, haplotypes that produced low levels of methyl salicylate, haplotypes 1,

368 3, 6 and 7, *SIMES1* and/or *SIMES3* were highly expressed in ripe fruits. A summary of *MES*
369 haplotypes, gene structure, and expression is shown in Table 1.



371 **Fig 8.** Relative gene expression of *MES1-4* in different Varitome accessions across the *MES*
372 haplotypes. The bar diagram represents the mean expression normalized by *ACT4* expression with
373 $\pm$ standard deviation. The number below the accessions represents the *MES* haplotype.

375 **Table 1.** Summary of *MES* haplotypes and gene expression. “+” represents the gene is highly
376 expressed and “-” represents either low or no gene expression. *For haplotype 9, expression levels
377 were predicted based on genome structure of the haplotype.

Haplotype		<i>MES1</i>	<i>MES2</i>	<i>MES3</i>	<i>MES4</i>	Average MeSA (ng/gmfw/hr)
1	Expression	+	-	+	-	0.562
	mutation				48 bp del in first exon	
2	Expression	-	-	-	-	6.383
	mutation	363 bp del in promoter		SNP at first exon	48 bp del in first exon	
3	Expression	+	-	+	-	0.191
	mutation				48 bp del in first exon	
4	Expression	-	-	-	-	4.506

	<u>mutation</u>	<u>363 bp del in promoter</u>		<u>SNP at first exon</u>	<u>48 bp del in first exon</u>	
<u>5</u>	<u>Expression</u>	<u>+</u>	<u>=</u>	<u>+</u>	<u>+</u>	<u>0.273</u>
	<u>mutation</u>					
<u>6</u>	<u>Expression</u>	<u>+</u>	<u>=</u>	<u>+</u>	<u>+</u>	<u>1.135</u>
	<u>mutation</u>					
<u>7</u>	<u>Expression</u>	<u>+</u>	<u>=</u>	<u>+</u>	<u>+</u>	<u>0.479</u>
	<u>mutation</u>					
<u>8</u>	<u>Expression</u>	<u>=</u>	<u>=</u>	<u>=</u>	<u>=</u>	<u>0.720</u>
	<u>mutation</u>			<u>deletion</u>	<u>48 bp del in first exon</u>	
<u>9*</u>	<u>Expression</u>	<u>=</u>	<u>=</u>	<u>=</u>	<u>=</u>	<u>4.378</u>
	<u>mutation</u>	<u>deletion</u>	<u>deletion</u>	<u>deletion</u>	<u>deletion</u>	

Formatted: Centered

Formatted: Centered

Formatted: Centered

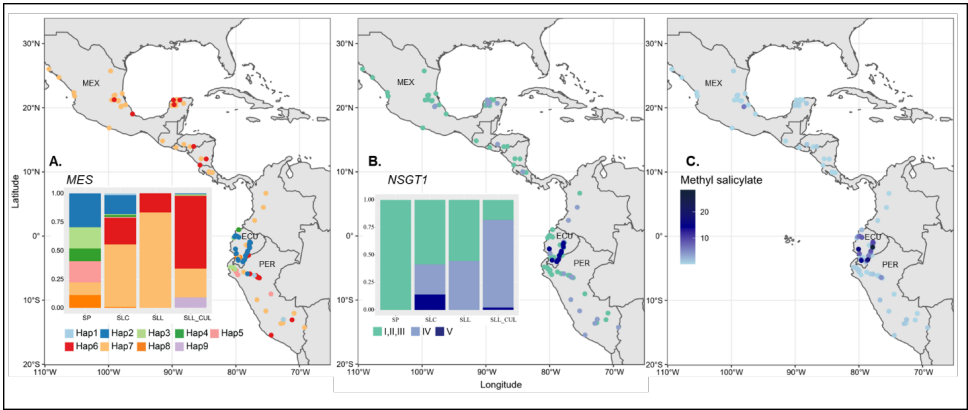
Formatted: Centered

Formatted: Centered

Evolution of the *MES* and *NSGTI* loci during tomato domestication and selection

SP and SLC accessions that were collected in Ecuador showed higher levels of methyl salicylate whereas SLL and most of SLL_CUL featured low levels of the volatile (Fig 2B). This suggested selection by ~~both~~ the terrestrial ecosystem and/or domestication on volatile production, presumably through the *MES* and *NSGTI* loci. To investigate how these two loci were selected during the evolution of tomato, we evaluated the haplotype distribution in the Varitome Collection and cultivated accessions (Fig 9). Most of the SLL_CUL, the majority of SLL, and many SLC carried the *MES* haplotype 6 and 7, suggesting strong selection for these alleles during domestication (Fig 9A) as these haplotypes were found to produce low levels of methyl salicylate (Fig 7A). The distribution map also showed that haplotypes 6 and 7 spread across several regions from Ecuador, Peru, to Colombia and Mexico. Not surprisingly, the most common allele in Ecuador was *MES* haplotype 2, associated with high levels of methyl salicylate. For *NSGTI*, we found the nonfunctional haplotypes IV and V to be the most common haplotypes in SLL_CUL. The proportion of accessions belonging to *NSGTI* haplotype IV and V was found to increase from SP (0%) to SLC (41.32%), SLL (44.44%) and SLL_CUL (81.81%) (Fig 9B). Based on the

394 distribution map, haplotype I, II, III had spread across several regions in South and Central
395 America whereas haplotype V was found only in Ecuador and northern Peru. Together, these data
396 suggested that functional *NSGT1* (Haplotype I, II, III) were under selection in SP but not in
397 cultivated tomato, whereas the opposite was true for *MES*. For the two loci combined, it appeared
398 that a strong selection for higher levels of methyl salicylate were particularly important to certain
399 SP and SLC accessions from Ecuador and Peru (Fig 9C), with a non-functional haplotype for both
400 *MES* and *NSGT1* (Figs 9A and 9B). The high levels of methyl salicylate might have contributed
401 to the survival in the specific ecosystem that is common to that region.



403 **Fig 9.** Proportion bar stack plot and distribution map of accessions in Varitome Collection (A)
404 *MES* haplotypes (B) *NSGT1* haplotypes. (C) Distribution map of methyl salicylate levels in the
405 accessions.

407 **Discussion**

408 Methyl salicylate is an important plant volatile that functions as a signaling molecule within and
409 between plants. Its role in defense responses in leaves is relatively well understood. It serves as an

herbivore and pathogen-induced volatile that attract parasitoids and predators of the herbivores. Plants emit methyl salicylate when attacked by herbivores, sending cues to parasitoids and other predators [11, 39-41]. Furthermore, methyl salicylate acts as a signaling molecule to activate the genes that are important in regulating the plants' response to biotic and abiotic stress [35, 42-44]. Methyl salicylate also functions as an inter-plant signal that activates disease resistance and the expression of defense-related genes in surrounding plants [39, 43]. Even though its role in defense responses is well established, the function of methyl salicylate in ripe fruit is not well understood. We show that the levels of this volatile vary dramatically in red-fruited tomato fruits and are particularly high in accessions originating from Ecuador. This suggests that methyl salicylate in ripe fruits may be beneficial to plants and that certain levels of this volatile in tomato are acceptable to consumers. It is likely that high levels of methyl salicylate in ripe fruits could serve as either an attractant to frugivores or deterrent to insect pests that often function as vectors of pathogenic microbes and viruses. No studies have been conducted on the effect of methyl salicylate levels on fruit health, but external applications and pretreatment of this volatile have been reported to reduce post-harvest losses in tomatoes. Methyl salicylate application reduces chilling injury [45] alleviates post-harvest fungal decay [46, 47], and suppresses loss of several desirable aromatic compounds [48]. Hence, even though the volatile is associated with low consumer liking, methyl salicylate could be important in reducing the impact of biotic and abiotic stressors on the plant and plant products such as the fruit.

MES, along with *NSGT1* is a major contributor to fruit methyl salicylate levels in the tomato germplasm. Moreover, the interaction between *MES* and *NSGT1* explained most of the variation in fruit volatile levels in a biparental mapping population [35], supporting the findings from this study. The combination of *MES* haplotype 2 and a non-functional *NSGT1* haplotype IV or V

433 resulted in the highest methyl salicylate levels indicating that the interaction of these two loci leads
 434 to higher levels of this volatile. Other genes in this pathway, *SISAMT1* and *SIUGT5*, appeared as
 435 minor players in this germplasm. The high or low methyl salicylate level haplogroups of *SAMT1*
 436 and *UGT5* could be explained by their genotype at *MES* and *NSGT1* loci. *SAMT1* haplotype II
 437 seemed to produce high levels of methyl salicylate. All accessions with *SAMT1* haplotype II
 438 carried a non-functional *MES* haplotype (haplotype 2 or 4) which might explain the high volatile
 439 levels even though the production of methyl salicylate is dependent on a functional *SAMT1*. In
 440 addition, the 11 accessions carrying a null mutation in *Solyc09g091530*, whose function in methyl
 441 salicylate production is unknown, carried the functional *NSGT1* haplotype which could explain the
 442 low volatile levels in the fruit. It is likely that there are other fruit-specific glycosyltransferases
 443 and methyltransferases with activity on methyl salicylate and salicylic acid respectively,
 444 downplaying individual effect of the cloned *SISAMT1* and *SIUGT5* [32, 33, 49]. Additionally, too
 445 few accessions remained after considering the effect of the major two loci (*MES* and *NSGT1*) to
 446 conclude which was the most likely active or inactive allele at the minor loci.

447 The *MES* and *NSGT1* loci have been shaped predominantly by SVs. It is known that SVs
 448 greatly affect gene expression and protein functions resulting in variations in phenotype [38, 50].
 449 *MES* carries several deletions and insertions, spanning the four *Methyltransferase* genes either
 450 partially or completely deleting genes from the locus. These SVs were likely responsible in part
 451 for the difference in *MES* gene expression observed among the haplotypes (Table 1). *SIMES1*
 452 appeared to be the only functional gene in all haplotypes, except for haplotype 9. However, the
 453 expression of this gene was very low or undetected in haplotypes carrying the 363 bp deletion in
 454 the promoter of *SIMES1* but was higher for haplotype 8 which did not have the deletion. *SIMES3*
 455 was the highest expressed gene at the *MES* locus especially in haplotypes 1, 3, 5, 6 and 7. We did

not find a correlation between gene expression and transposon insertion or truncation of *SIMES3*. Instead, we identified a SNP (SL4.0ch02: 34,463,357; C to T) in the first exon of *SIMES3* found only in accessions in haplotypes 2 and 4. This SNP is predicted to cause an amino acid change from leucine to phenylalanine at position 83, which is one of the salicylic acid binding sites of the protein [51] and possibly consequential. This finding suggests that *SIMES3* is likely functional in addition to *SIMES1*. For *SIMES4*, expression levels were associated with the truncation of the protein. Haplotypes 1, 2, 3, 4, and 8 carrying a truncated *SIMES4* had low or undetectable expression, whereas the haplotypes 5, 6, and 7 carried full length *SIMES4* that showed high expression.

Several events in selection and domestication of cultivated tomatoes have affected the fruit methyl salicylate levels and *MES* controlling the production of this volatile. Recent evolutionary models proposed by Razifard et al. [28] and Blanca et al. [29] ~~support can be compared to~~ the *MES* haplotype distribution. SP from Peru, considered as the most diverse and ancestral, includes accessions with several *MES* haplotypes (Fig 10). During its spread north, the tomato plants would have encountered biotic and abiotic conditions associated with higher latitudes, such as changes in temperature, water availability, or disease, which likely shaped the natural selection for *MES* haplotypes. SP from southern Ecuador carried the functional *MES* haplotype (7), while SP from northern Ecuador carried the non-functional haplotypes (2 and 4). Non-functional *MES* haplotype and high levels of methyl salicylate may have contributed to their survival in adverse environments. It is also possible that birds or animal dispersers, having preference for minty flavor, contributed to selection and migration, but little is known about this. However, selections imposed by the new environment necessitates adaptation that might have favored the selection of a specific *MES* haplotype over the other. According to Blanca et al. [29], SP spread into Mesoamerica where

479 it evolved into SLC and the ancestral SLC migrated back to South America where it hybridized
480 with native SP populations. Following this model, the most functional *MES* haplotype 7, with
481 potentially four functional genes, was selected as Peruvian and Ecuadorian SP evolved into SLC
482 in Mesoamerica. *MES* haplotype 6 evolved from haplotype 7, after a premature stop-codon
483 mutation in *MES2*. Upon migration back to South America, admixture with Ecuadorian SP resulted
484 in SLC from Ecuador carrying the most diverse *MES* haplotypes. SLC Peru evolved from
485 admixture with Peruvian SP resulting in at least four *MES* haplotypes, and together with
486 Ecuadorian SLC spread northwards to evolve into ancestral SLL carrying only haplotype 6 and 7.
487 On the other hand, Razifard et al. [28] proposed that SP from northern Ecuador evolved into SLC
488 in South America carrying several domestication alleles which then migrated to Mesoamerica
489 where it evolved into SLL. Based on this model, the functional *MES* haplotype 7 and its derivative
490 haplotype 6 are found in Peruvian and Ecuadorian SP and were selected early as tomato migrated
491 northwards to evolve into SLL. The evolution in *MES* locus seems to agree more with the model
492 proposed by Razifard et al. [28] as the linear gene flow model from Peruvian and Ecuadorian SP
493 to Mexican SLL explained the decrease in diversity at the locus. Regardless of the model, large
494 genetic diversity at the *MES* locus was maintained in South America whereas migration
495 northwards into Mesoamerican SLC and eventually SLL resulted in the selection of only two *MES*
496 haplotypes 6 and 7. Thus, the evolution at the *MES* locus suggests a complex migratory and
497 evolutionary path of tomato.

498 The genetic variation for the known genes associated with methyl salicylate biosynthesis
499 and conjugation in the Varitome Collection was large, especially at the *MES* locus. Moreover, even
500 within each distinct haplotype, we observed wide phenotypic variation (Fig 7), suggesting that
501 additional genetic factors control the biosynthesis process. GWAS using the Varitome Collection

identified several additional QTL associated with fruit methyl salicylate levels in addition to known genes (Fig 3 and S2 Table). These QTL could harbor novel genes, either affecting the homeostasis of methyl salicylate including additional *Glycosyltransferases* and *Glucosidases*. The Varitome Collection thus, is a valuable tool to discover additional genes associated with volatile production through a genetic mapping approach. Identification of the novel genes would allow us to gain a deeper understanding of the regulation of methyl salicylate in the fruits.

The results presented herein contribute to understanding the genetics of the biosynthesis and conjugation of methyl salicylate in ripe tomato fruits. The beneficial haplotypes at the *MES* and *NSGT1* loci could be introgressed into cultivated germplasm to obtain optimal flavor depending on the agroecosystem. [Introgression of beneficial haplotypes of NSGT1 has the added advantage of decreasing levels of the another undesirable volatile, guaiacol, thereby levels and improving flavor.](#) Thus, the genomic information from this study can be used to develop molecular tools to perform marker-assisted selection for the desirable haplotype at each locus. Intermediate levels of methyl salicylate in modern and heirloom cultivars suggest that certain amounts of this volatile are acceptable or were inadvertently selected because of a strong breeding focus on yield and disease resistance. If the latter, breeding that is focused on the optimal allele selection that favors flavor is now an attainable goal. Thus, the genetic diversity knowledge can be used to fine-tune the methyl salicylate levels to arrive at an acceptable amount of methyl salicylate for overall plant health and resistance as well as consumer satisfaction to balance the desired taste with other post-harvest traits such as shelf life.

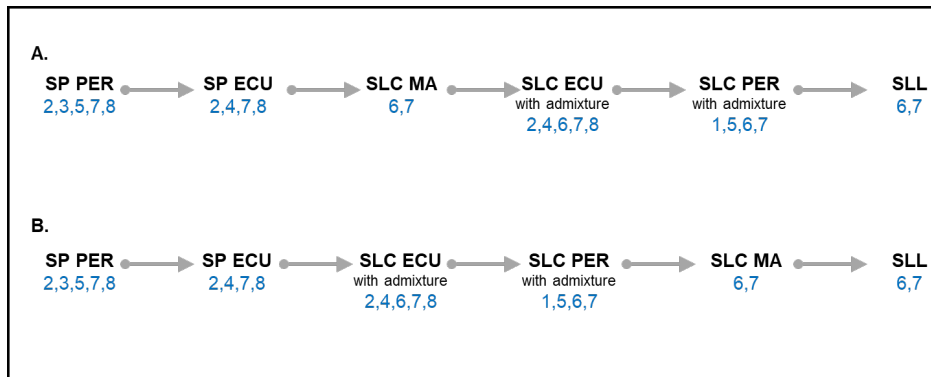


Fig 10. Linear illustration of evolutionary model proposed by **A.** Blanca et al. [29] and **B.** Razifard et al. [28]. The numbers in blue represent the *MES* haplotypes. The gray solid arrow represents the direction of evolution and dotted arrow represents proposed hybridization. The abbreviations are PER: Peru, ECU: Ecuador, and MA: Mesoamerica. SLC from PER and ECU includes some accessions which are admixture of SP and SLC.

Methods

Plant materials and phenotyping

A panel of 166 tomato accessions, the Varitome Collection, consisting of 27 wild SP, 121 semi domesticated SLC and 18 ancestral landrace SLL, as well as 143 heirloom and modern varieties (SLL_CUL) were grown and phenotyped for volatile accumulation as previously described [25, 28]. Based on collection site and whole genome information, the species that comprise these accessions were further divided into three SP subpopulations (SP_SECU from Southern Ecuador, SP_NECU from Norther Ecuador and SP_PER from Peru); six SLC subpopulations (SLC_ECU from Ecuador, SLC_PER from Peru, SLC_SM from San Martin region of Peru, SLC_CA from

538 Central, Northern South and Southern North America, SLC_MEX from Mexico, and SLC_ADM
539 which are admixture SLC accessions); and two SLL subpopulations namely SLL comprised of
540 ancestral landraces and SLL_CUL comprised of heirloom and modern accessions. The red ripe
541 fruits were harvested, ~~chopped~~cut, and loaded in glass tubes. ~~A and the~~air was passed over samples
542 for an hour ~~to collect~~and the volatiles were collected on a ~~a~~ SuperQ Resin column. Volatiles were
543 then eluted off the column with methylene chloride and ran on GC-MS for analysis as previously
544 described [25]. The volatile measurements were average of volatile levels from at least two and up
545 to five harvests.

546 **Statistical analysis**

547 All statistical analysis were conducted using R [52] unless mentioned otherwise. The figures were
548 generated using “ggplot2” package [53] in R. The multiple mean comparison was performed using
549 the “duncan.test” function from “agricolae” package in R [54].

550 **Variant calling and Genome-wide Association Study**

551 Raw ILLUMINA read files were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>; SRA:
552 SRP150040, SRP045767, SRP094624, and PRJNA353161). We augmented the original GWAS
553 by Razifard et al. [28] by aligning the reads to the most recent tomato genome SL4.0 build as
554 opposed to SL2.50. We also interrogated the three main variant types, namely SNPs, INDELs and
555 SVs for the GWAS compared to SNPs only [35]. Raw read evaluation, filtering, alignment, and
556 variant calling were performed as previously described [55]. The phenotype deviating from
557 normality (p value from Shapiro test <0.01) was normalized using quantile normalization in R
558 (52). Associations between genotype and phenotype were calculated using the FarmCPU model in
559 GAPIT (version 3) [56] following the established protocols [35]. Association analysis was

560 conducted using 21,893,681 SNPs, 2,735,297 INDELs and 27,477 SVs. The association results
561 were plotted using the “ggplot2” package in R [53].

562 **Haplotype analyses**

563 To identify the haplotypes at the known genes regulating methyl salicylate levels, we first
564 determined the presence of SVs at or near each by comparing the ~~high-quality~~high-quality genome
565 sequences of 28 Varitome accessions
566 (ftp://ftp.solgenomics.net/genomes/tomato100/March_02_2020_sv_landscape/variants/). The
567 genomes of these accessions were sequenced using Oxford Nanopore Technology and Illumina
568 short read sequencing and, therefore, were well suited for the identification of SV at each of the
569 loci. For the loci with evidence of SV, *MES* and *NSGT1*, SVs were identified in Integrative
570 Genome Viewer (IGV) [57] on the pattern of short reads in the tomato accessions, and haplotypes
571 were defined largely based on SVs. The size of the *MES* locus ranged from 15 kb to 35 kb, whereas
572 the size of the *NSGT1* locus was as defined previously [38]. For *SISAMT1*, the locus comprised
573 three tandem duplicated genes without large SVs. To define the haplotypes at the *SAMT1* and
574 *UGT5* loci, SNPs and small INDELs within the coding region, 3 kb upstream and 1 kb downstream
575 of the termination of transcription were extracted using VCFtools [58]. The method to generate
576 the haplotypes based on SNPs and INDELs was described previously [55]. Variants (SNPs and
577 INDELs) were annotated using SnpEff [59] using a local built database of SL4.0 tomato reference
578 genome. The FGENESH [60] tool was used to verify and predict the gene models. The haplotype
579 distribution map was created using R package “rnaturalearth” ([https://CRAN.R-](https://CRAN.R-project.org/package=rnaturalearth)
580 [project.org/package=rnaturalearth](https://CRAN.R-project.org/package=rnaturalearth)) and “ggspatial” ([https://CRAN.R-](https://CRAN.R-project.org/package=ggspatial)
581 [project.org/package=ggspatial](https://CRAN.R-project.org/package=ggspatial)).

582 **Phylogenetic tree construction of the *Methylesterases* at *MES* locus**

583 A phylogenetic tree was built using Neighbor-Joining method with the genomic sequences of
584 genes at the *MES* locus and 10 kb flanking the locus. The consensus nucleotide sequences were
585 generated using IGV and, exported in Geneious Prime v2021.2.2 (<https://www.geneious.com/>).
586 The sequences were then aligned using the Clustal alignment option. The tree was constructed
587 using Neighbor-Joining method and Tamura-Nei genetic distance model using *S. pennellii*
588 (sequence obtained from SGN; <https://solgenomics.net/>) as an outgroup. Bootstrap support for the
589 tree was obtained using 100 bootstrap replicates. The tree was then plotted using “ggtree” package
590 in R [61].

591 **qRT-PCR**

592 qRT-PCR was performed following a protocol that was used previously [35]. Total RNA was
593 extracted from red ripe fruits of 37 accessions and used for the qRT-PCR, including some that
594 were collected before [35]. *ACTIN4* (*Solyc04g011500*) was used as the control to standardize
595 relative expression levels. The relative expression was calculated using the formula of ΔC_q method
596 using a reference gene: $R = 2^{C_q(\text{reference}) - C_q(\text{target})}$, where C_q is quantification cycle, reference
597 is *ACT4*, and targets are *MES1*, *MES2*, *MES3*, or *MES4*. The expression analysis figure was
598 generated using “ggplot2” package in R [53]. The sequences of primers used in the study are
599 provided in S7 Table.

600 **Data access**

601 All the raw sequencing data used in this study are publicly available in NCBI
602 (<https://www.ncbi.nlm.nih.gov/>; SRA: SRP150040, SRP045767, SRP094624, and
603 PRJNA353161).

604 **Acknowledgements**

605 We thank the Georgia Advanced Computing Resource Center, UGA, for providing the
606 computation resources. M.S. is the recipient of the John Ingle Innovation in Plant Breeding Award
607 and Roger Boerma Plant Breeding Excellence Scholarship Award. This work was funded by NSF
608 IOS 2151032 (EvdK and DT) and NSF IOS 1732253 (EvdK).

References

1. Raguso RA. Wake up and smell the roses: the ecology and evolution of floral scent. Annual review of ecology, evolution, and systematics. 2008;549-69. doi:10.1146/annurev.ecolsys.38.091206.095601
2. Schiestl FP. Ecology and evolution of floral volatile-mediated information transfer in plants. New Phytologist. 2015;206(2):571-7. doi:10.1111/nph.13243
3. Pichersky E, Gershenzon J. The formation and function of plant volatiles: perfumes for pollinator attraction and defense. Current opinion in plant biology. 2002;5(3):237-43. doi:10.1016/s1369-5266(02)00251-0
4. Borges RM, Ranganathan Y, Krishnan A, Ghara M, Pramanik G. When should fig fruit produce volatiles? Pattern in a ripening process. Acta Oecologica. 2011;37(6):611-8. doi:10.1016/j.actao.2011.06.003
5. Rodríguez A, Alquézar B, Peña L. Fruit aromas in mature fleshy fruits as signals of readiness for predation and seed dispersal. New Phytologist. 2013;197(1):36-48. doi:10.1111/j.1469-8137.2012.04382.x
6. Fallik E, Archbold DD, Hamilton-Kemp TR, Clements AM, Collins RW, Barth MM. (E)-2-Hexenal can stimulate Botrytis cinerea growth in vitro and on strawberries in vivo during storage. Journal of the American Society for Horticultural Science. 1998;123(5):875-81. doi:10.21273/JASHS.123.5.875
7. Rodríguez A, San Andrés V, Cervera M, Redondo A, Alquézar B, Shimada T, et al. Terpene down-regulation in orange reveals the role of fruit aromas in mediating interactions with insect herbivores and pathogens. Plant Physiology. 2011;156(2):793-802. doi:10.1104/pp.111.176545

- 632 8. Karban R, Shiojiri K, Huntzinger M, McCall AC. Damage-induced resistance in
633 sagebrush: volatiles are key to intra-and interplant communication. *Ecology*. 2006;87(4):922-30.
634 doi:10.1890/0012-9658(2006)87[922:drisva]2.0.co;2
- 635 9. Qualley A, Dudareva N. *Plant volatiles*. eLS. 2010.
636 doi:10.1002/9780470015902.a0000910.pub2
- 637 10. De Moraes CM, Mescher MC, Tumlinson JH. Caterpillar-induced nocturnal plant
638 volatiles repel conspecific females. *Nature*. 2001;410(6828):577-80. doi:10.1038/35069058
- 639 11. Kessler A, Baldwin IT. Defensive function of herbivore-induced plant volatile emissions
640 in nature. *Science*. 2001;291(5511):2141-4. doi:10.1126/science.291.5511.2141
- 641 12. Andersen RA, Hamilton-Kemp TR, Hildebrand DF, McCracken Jr CT, Collins RW,
642 Fleming PD. Structure-antifungal activity relationships among volatile C6 and C9 aliphatic
643 aldehydes, ketones, and alcohols. *Journal of Agricultural and Food Chemistry*. 1994;42(7):1563-
644 8. doi:10.1021/jf00043a033
- 645 13. Vancanneyt G, Sanz C, Farmaki T, Paneque M, Ortego F, Castañera P, et al.
646 Hydroperoxide lyase depletion in transgenic potato plants leads to an increase in aphid
647 performance. *Proceedings of the national academy of sciences*. 2001;98(14):8139-44.
648 doi:10.1073/pnas.141079498
- 649 14. Aartsma Y, Bianchi FJ, van der Werf W, Poelman EH, Dicke M. Herbivore-induced
650 plant volatiles and tritrophic interactions across spatial scales. *New Phytologist*.
651 2017;216(4):1054-63. doi:10.1111/nph.14475
- 652 15. Bouwmeester H, Schuurink RC, Bleeker PM, Schiestl F. The role of volatiles in plant
653 communication. *The Plant Journal*. 2019;100(5):892-907. doi:10.1111/tpj.14496
- 654 16. Gasmi L, Martínez-Solís M, Frattini A, Ye M, Collado MC, Turlings TC, et al. Can
655 herbivore-induced volatiles protect plants by increasing the herbivores' susceptibility to natural

656 pathogens? *Applied and environmental microbiology*. 2019;85(1):e01468-18.
 657 doi:10.1128/AEM.01468-18
 658 17. Couty A, Kaiser L, And DH, Pham-Delegue MH. The attractiveness of different odour
 659 sources from the fruit–host complex on *Leptopilina boulardi*, a larval parasitoid of frugivorous
 660 *Drosophila* spp. *Physiological Entomology*. 1999;24(1):76-82. doi:10.1046/j.1365-
 661 3032.1999.00116.x
 662 18. Melo JWS, Lima DB, Pallini A, Oliveira JEM, Gondim MG. Olfactory response of
 663 predatory mites to vegetative and reproductive parts of coconut palm infested by *Aceria*
 664 *guerreronis*. *Experimental and Applied Acarology*. 2011;55(2):191-202. doi:10.1007/s10493-011-
 665 9465-1
 666 19. Mäntylä E, Alessio GA, Blande JD, Heijari J, Holopainen JK, Laaksonen T, et al. From
 667 plants to birds: higher avian predation rates in trees responding to insect herbivory. *PLoS One*.
 668 2008;3(7):e2832. doi:10.1371/journal.pone.0002832
 669 20. Tieman D, Bliss P, McIntyre LM, Blandon-Ubeda A, Bies D, Odabasi AZ, et al. The
 670 chemical interactions underlying tomato flavor preferences. *Current Biology*. 2012;22(11):1035-
 671 9. doi:10.1016/j.cub.2012.04.016
 672 21. Rambla JL, Tikunov YM, Monforte AJ, Bovy AG, Granell A. The expanded tomato fruit
 673 volatile landscape. *Journal of Experimental Botany*. 2014;65(16):4613-23. doi:10.1093/jxb/eru128
 674 22. Martina M, Tikunov Y, Portis E, Bovy AG. The genetic basis of tomato aroma. *Genes*.
 675 2021;12(2):226. doi:10.3390/genes12020226
 676 23. Dudareva N, Pichersky E, Gershenzon J. Biochemistry of plant volatiles. *Plant*
 677 *physiology*. 2004;135(4):1893-902. doi:10.1104/pp.104.049981
 678 24. Hui YH, Chen F, Noll LM, Guiné RP, Martín-Belloso O, Mínguez-Mosquera MI, et al.
 679 *Handbook of fruit and vegetable flavors: Wiley Online Library*; 2010.

680 25. Tieman D, Zhu G, Resende Jr MF, Lin T, Nguyen C, Bies D, et al. A chemical genetic
681 roadmap to improved tomato flavor. *Science*. 2017;355(6323):391-4. doi:10.1126/science.aal1556

682 26. Särkinen T, Bohs L, Olmstead RG, Knapp S. A phylogenetic framework for evolutionary
683 study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC evolutionary biology*.
684 2013;13(1):1-15. doi:10.1186/1471-2148-13-214

685 27. Pease JB, Haak DC, Hahn MW, Moyle LC. Phylogenomics reveals three sources of
686 adaptive variation during a rapid radiation. *PLoS biology*. 2016;14(2):e1002379.
687 doi:10.1371/journal.pbio.1002379

688 28. Razifard H, Ramos A, Della Valle AL, Bodary C, Goetz E, Manser EJ, et al. Genomic
689 evidence for complex domestication history of the cultivated tomato in Latin America.
690 *Molecular biology and evolution*. 2020;37(4):1118-32. doi:10.1093/molbev/msz297

691 29. Blanca J, Sanchez-Matarredona D, Ziarolo P, Montero-Pau J, van der Knaap E, Díez
692 MJ, et al. Haplotype analyses reveal novel insights into tomato history and domestication driven
693 by long-distance migrations and latitudinal adaptations. *Horticulture Research*. 2022;9.
694 doi:10.1093/hr/uhac030

695 30. Klee HJ, Tieman DM. The genetics of fruit flavour preferences. *Nature Reviews*
696 *Genetics*. 2018;1. doi:10.1038/s41576-018-0002-5

697 31. Bruhn CM, Feldman N, GARLITZ C, Harwood J, Ivans E, Marshall M, et al. Consumer
698 perceptions of quality: apricots, cantaloupes, peaches, pears, strawberries, and tomatoes. *Journal*
699 *of food quality*. 1991;14(3):187-95. doi:10.1111/j.1745-4557.1991.tb00060.x

700 32. Tieman D, Zeigler M, Schmelz E, Taylor MG, Rushing S, Jones JB, et al. Functional
701 analysis of a tomato salicylic acid methyl transferase and its role in synthesis of the flavor
702 volatile methyl salicylate. *The Plant Journal*. 2010;62(1):113-23. doi:10.1111/j.1365-
703 313X.2010.04128.x

33. Louveau T, Leita C, Green S, Hamiaux C, Van der Rest B, Dechy-Cabaret O, et al. Predicting the substrate specificity of a glycosyltransferase implicated in the production of phenolic volatiles in tomato fruit. *The FEBS journal*. 2011;278(2):390-400. doi:10.1111/j.1742-4658.2010.07962.x
34. Tikunov YM, Molthoff J, de Vos RC, Beekwilder J, van Houwelingen A, van der Hooft JJ, et al. Non-smoky glycosyltransferase1 prevents the release of smoky aroma from tomato fruit. *The Plant Cell*. 2013;25(8):3067-78. doi:10.1105/tpc.113.114231
35. Frick EM, Sapkota M, Pereira Garcia L, Wang Y, Hermanns A, Giovannoni JJ, et al. A family of methyl esterases converts methyl salicylate to salicylic acid in ripening tomato fruit. *Plant Physiology*. 2023; doi:10.1093/plphys/kiac509
36. Rivas-San Vicente M, Plasencia J. Salicylic acid beyond defence: its role in plant growth and development. *Journal of experimental botany*. 2011;62(10):3321-38. doi:10.1093/jxb/err031
37. Baldwin IT. Plant volatiles. *Current Biology*. 2010;20(9):R392-R7. doi:10.1016/j.cub.2010.02.052
38. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*. 2020;182(1):145-61. e23. doi:10.1016/j.cell.2020.05.021
39. Shulaev V, Silverman P, Raskin I. Airborne signalling by methyl salicylate in plant pathogen resistance. *Nature*. 1997;385(6618):718-21. doi:10.1038/385718A0
40. De Boer JG, Dicke M. The role of methyl salicylate in prey searching behavior of the predatory mite *Phytoseiulus persimilis*. *Journal of chemical ecology*. 2004;30(2):255-71. doi:10.1023/b:joec.0000017976.60630.8c

41. Zhu J, Park K-C. Methyl salicylate, a soybean aphid-induced plant volatile attractive to the predator *Coccinella septempunctata*. *Journal of chemical ecology*. 2005;31(8):1733-46. doi:10.1007/s10886-005-5923-8
42. Hardie J, Isaacs R, Pickett JA, Wadhams LJ, Woodcock CM. Methyl salicylate and (-)-(1R, 5S)-myrtenal are plant-derived repellents for black bean aphid, *Aphis fabae* Scop.(Homoptera: Aphididae). *Journal of chemical ecology*. 1994;20(11):2847-55. doi:10.1007/BF02098393
43. Park S-W, Kaimoyo E, Kumar D, Mosher S, Klessig DF. Methyl salicylate is a critical mobile signal for plant systemic acquired resistance. *Science*. 2007;318(5847):113-6. doi:10.1126/science.1147113
44. Gao H, Guo M, Song J, Ma Y, Xu Z. Signals in systemic acquired resistance of plants against microbial pathogens. *Molecular Biology Reports*. 2021;48(4):3747-59. doi:10.1007/s11033-021-06344-7
45. Ding C-K, Wang CY, Gross KC, Smith DL. Reduction of chilling injury and transcript accumulation of heat shock proteins in tomato fruit by methyl jasmonate and methyl salicylate. *Plant Science*. 2001;161(6):1153-9. doi:10.1016/S0168-9452(01)00521-0
46. Zhang X, Min D, Li F, Ji N, Meng D, Li L. Synergistic effects of L-arginine and methyl salicylate on alleviating postharvest disease caused by *Botrytis cinerea* in tomato fruit. *Journal of agricultural and food chemistry*. 2017;65(24):4890-6. doi: 10.1021/acs.jafc.7b00395
47. Min D, Li F, Zhang X, Shu P, Cui X, Dong L, et al. Effect of methyl salicylate in combination with 1-methylcyclopropene on postharvest quality and decay caused by *Botrytis cinerea* in tomato fruit. *Journal of the Science of Food and Agriculture*. 2018;98(10):3815-22. doi:10.1002/jsfa.8895

749 48. Zeng X, Wang L, Fu Y, Zuo J, Li Y, Zhao J, et al. Effects of methyl salicylate pre-
750 treatment on the volatile profiles and key gene expressions in tomatoes stored at low
751 temperature. *Frontiers in Nutrition*. 2022;9. doi:10.3389/fnut.2022.1018534

752 49. Kamiyoshihara Y, Tieman DM, Klee HJ. Analyses of Plant UDP-Dependent
753 Glycosyltransferases to Identify Their Volatile Substrates Using Recombinant Proteins. *Plant*
754 *Signal Transduction: Springer*; 2016. p. 199-207. doi:10.1007/978-1-4939-3115-6_16

755 50. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural
756 variation on human gene expression. *Nature genetics*. 2017;49(5):692-9. doi:10.1038/ng.3834

757 51. Zubieta C, Ross JR, Koscheski P, Yang Y, Pichersky E, Noel JP. Structural basis for
758 substrate recognition in the salicylic acid carboxyl methyltransferase family. *The Plant Cell*.
759 2003;15(8):1704-16. doi:10.1105/tpc.014548

760 52. R Core Team. R: A language and environment for statistical computing. Vienna, Austria:
761 R Foundation for Statistical Computing; 2021. <https://www.R-project.org/>

762 53. Wickham H. *ggplot2*. Wiley Interdisciplinary Reviews: Computational Statistics.
763 2011;3(2):180-5. <https://ggplot2.tidyverse.org>.

764 54. de Mendiburu F, de Mendiburu MF. Package ‘agricolae’. R Package, version. 2019;1(3).
765 <https://CRAN.R-project.org/package=agricolae>

766 55. Pereira L, Sapkota M, Alonge M, Zheng Y, Zhang Y, Razifard H, et al. Natural genetic
767 diversity in tomato flavor genes. *Frontiers in plant science*. 2021;12:914.
768 doi:10.3389/fpls.2021.642828

769 56. Wang J, Zhang Z. GAPIT Version 3: boosting power and accuracy for genomic
770 association and prediction. *Genomics, proteomics & bioinformatics*. 2021;19(4):629-40.
771 doi:10.1016/j.gpb.2021.08.005

- 772 57. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al.
773 Integrative genomics viewer. *Nature biotechnology*. 2011;29(1):24-6. doi:10.1038/nbt.1754
- 774 58. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant
775 call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-8. doi:10.1093/bioinformatics/btr330
- 776 59. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for
777 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the
778 genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80-92.
779 doi:10.4161/fly.19695
- 780 60. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic
781 genes, pseudogenes and promoters. *Genome biology*. 2006;7(1):1-12. doi:10.1186/gb-2006-7-s1-
782 s10
- 783 61. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and
784 annotation of phylogenetic trees with their covariates and other associated data. *Methods in*
785 *Ecology and Evolution*. 2017;8(1):28-36. doi:10.1111/2041-210X.12628

786

787 **Supporting information caption**

788 **S1 Fig.** Genomic alignment of Varitome accessions to SL4.0 genome version at *MES* locus. (A)
789 Alignment of three highest methyl salicylate producing accessions against the SL4.0 genome build
790 at the *MES* locus. (B) Alignment of three least methyl salicylate producing accessions against the
791 SL4.0 genome build at the *MES* locus. Yellow boxes represent the gene models.

792 **S2 Fig.** ClustwalW alignment of predicted protein sequence of *SIMES1-4* of all *MES* haplotypes.
793 Alignments of (A) *SIMES1* (B) *SIMES2* (C) *SIMES3* and (D) *SIMES4*

794 **S3 Fig.** Comparison of *MES* haplotype 2 and 5 and representation of a transposon insertions
795 between *SIMES2* and *SIMES3*; and within *SIMES3*. The gene model below is a screenshot from
796 SL4.0 genome version of the exact region which is the insertion showing a transposon insertion
797 between *SIMES2* and *SIMES3*.

798 **S4 Figure.** Alignment of Rio Grande against the SL4.0 genome build at the *MES* locus. Alignment
799 for *SIMES1*, *SIMES2*, *SIMES3* and *SIMES4*. Yellow boxes represent the gene models.

800 **S5 Figure.** Haplotype analysis of *SAMT1* locus. (A) Heatmap representing the genotypes of
801 accessions (rows) for the polymorphisms identified (columns). Reference genotypes are
802 represented in blue, alternate in red, heterozygous in yellow and missing data in white. The black
803 rectangular box represents the position of the genes in the locus. (B) Distribution of methyl
804 salicylate in red fruits in different accessions among different SAMT haplotypes. (C) Distribution
805 of methyl salicylate in red fruits in different accessions with and without deletion in the *SAMT1*
806 locus. (D) Distribution of methyl salicylate between accessions with and without null mutations
807 of *Solyc09g091530*.

808 **S6 Fig.** Phylogeny tree constructed using consensus DNA sequence of *SAMT1* locus and 10 kb
809 flanking region of *SAMT* locus from Varitome collection. Different colors of accessions and
810 branches represents different MES haplotypes. The outer two concentric circles represent the
811 grouping of accessions (SP, SLC and SLL) and corresponding methyl salicylate levels (MeSA) in
812 the accessions respectively. The numbers on the branches represent the bootstrap values.

813 **S7 Fig.** Haplotype analysis of *UGT5* locus. (A) Heatmap representing the genotypes of accessions
814 (rows) for the polymorphisms identified (columns). Reference genotypes are represented in blue,
815 alternate in red, heterozygous in yellow and missing data in white. The black rectangular box

816 represents the position of the gene in the locus. (B) Distribution of methyl salicylate in red fruits
 817 in different accessions among different *UGT5* haplotypes. (C) Distribution of methyl salicylate in
 818 red fruits in different accessions with and without deletion in the promoter of *SIUGT5*.

819 **S8 Fig.** Phylogeny tree constructed using consensus DNA sequence of *UGT5* locus and 10 kb
 820 flanking region of *UGT5* locus from Varitome collection. Different colors of accessions and
 821 branches represents different MES haplotypes. The outer two concentric circles represent the
 822 grouping of accessions (SP, SLC and SLL) and corresponding methyl salicylate levels (McSA) in
 823 the accessions respectively. The numbers on the branches represent the bootstrap values.

824 **S1 Table.** List of significant loci associated with methyl salicylate levels and their respective
 825 variants from SNP, INDEL and SV GWAS, their position, p-value, LOD values and cloned genes.

826 **S2 Table.** List of SVs identified in MES locus and used for defining MES haplotypes

827 **S3 Table.** Methyl salicylate levels (ng/gfw/hr) in Varitome collection and heirlooms and their
 828 respective *MES* and *NSGTI* haplotype

829 **S4 Table.** Annotation of all the variants (SNPs and INDELs) in the *MES* locus in Varitome
 830 collection. The *MES* haplotype information is provided in the column besides the accession names.
 831 The annotation information is provided along each variant. 0|0 represents reference allele, 1|1
 832 represents alternate allele, and 1|0 represents heterozygous allele. The yellow highlighted variants
 833 are the variants discussed in the text.

834 **S5 Table.** Annotation of all the variants (SNPs and INDELs) in the *SAMTI* locus in Varitome
 835 collection. The *SAMTI* haplotype information is provide in the column besides the accession
 836 names. The annotation information is provided along each variant. 0|0 represents reference allele,
 837 1|1 represents alternate allele, and 1|0 represents heterozygous allele. The yellow highlighted
 838 variants are the variants discussed in the text.

839 **S6 Table.** Annotation of all the variants (SNPs and INDELs) in the *UGT5* locus in Varitome
840 collection. The *UGT5* haplotype information is provide in the column besides the accession names.
841 The annotation information is provided along each variant. 0|0 represents reference allele, 1|1
842 represents alternate allele, and 1|0 represents heterozygous allele. The yellow highlighted variants
843 are the variants discussed in the text.

844 **S7 Table.** qRT PCR primers for the expression analysis of *SIMES1-4* and *ACT4*.