# Role of MULE activities in the positive selection of sweet orange cultivars

Bo Wu[1,*], Yiping Cui[1,2,*], Yongping Duan[3], Frederick Gmitter Jr.[4], Feng Luo[1,#],

[1]School of Computing, Clemson University, 100 McAdams Hall, Clemson, SC, USA
[2]Guangdong Provincial Key Laboratory of High Technology for Plant Protection, Plant Protection Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou, China
[3]USDA-ARS, U.S. Horticultural Research Laboratory, 2001 South Rock Road, Fort Pierce, FL, USA.
[4]Department of Horticultural Sciences, Citrus Research and Education Center, University of Florida, IFAS, 700 Experiment Station Road, Lake Alfred, FL, USA

*These authors contributed equally to the manuscript as first authors.

#To whom correspondence should be addressed:

Feng Luo. Tel: +01 864 633 6901. Email: luofeng@clemson.edu;

## Abstract

Sweet orange (SWO) mainly propagates asexually and has a higher bud sport selection frequency than most other citrus species[1]. Many cultivar groups have been subjected to artificial selections for tens to hundreds of years[2]. The reasons for the high bud sport selection frequency and how positive selections impact the mutation activities in SWO are not well understood. Here, we find two highly active Mutator-like transposable element families (CiMULE1 and CiMULE2) with novel insertion loci in most SWO accessions, and their transposition activities in SWO are up to over 100-fold higher than in SWO's two parental species. Almost all SWO cultivars could be distinguished based on their insertion loci, and three to eight tag insertions are found for each cultivar group. Many insertions have putatively altered cultivar traits, including several biallelic pairs. Multiple SWO lineages with high transposition activities of CiMULE1 or CiMULE2 have been selected independently. CiMULE1 and CiMULE2 have played an important role in SWO breeding. And artificial selections led to the significant increase of their activities in SWO within tens to hundreds of years, which in turn accelerated bud sport selection dramatically. CiMULE1 and CiMULE2 may function as accelerators in adaptative evolution in citrus.

## Main

Sweet orange (*Citrus sinensis*, SWO) originated from complex inter-specific hybridization processes between pummelo (*Citrus maxima*) and mandarin (*Citrus reticulata*)[3,4]. Its cultivars are mainly propagated asexually through apomixis or grafting. SWO has one of the largest number of bud sport selections in *Citrus* with diverse horticultural traits[1,4,5], which reached more than 1,300 already in 1936[6]. However, the reason for the high bud sport selection frequency in SWO is not well understood. Moreover, SWO has several cultivar groups subjected to continuous artificial selections for tens to hundreds of years, making them the ideal materials to find the impact of positive selections on mutation activities[7].

Transposable element (TE) is one of the important driving forces of mutation in plants[8]. In a maize lineage, an active mutator system has been reported to increase the mutation rate by 30 times[9]. Wang et al. (2021) detected 877 transposable elements (TE) insertions including 266 Mutator-like transposable elements (MULEs) insertions in 114 SWO accessions[1]. However, no clue was provided on the transposition activity level of SWO contributing to its high bud sport selection frequency. Meanwhile, transposon insertions are the main source (32/40) of the spontaneous structural mutations in three Valencia sweet oranges[4]. Whether TEs have a role in the high bud sport selection frequency remains a question. Here, we study the TE activities in 128 SWO cultivars to understand their contribution to the bud sport diversity and the reciprocal impacts between positive selections and TE activities.

### Two highly active mutator-like elements families in sweet oranges

By comparing 11 published SWO assemblies[1,4], we find two highly active MULE families, CiMULE1 and CiMULE2, with 69 and 55 insertion loci (ILs) in them, respectively (Fig. 1a). 46 of the 69 CiMULE1 ILs and 33 of the 55 CiMULE2 ILs are assembly-specific (Fig. 1a). CiMULE1 and CiMULE2 members both own 9 bp target site duplications (TSDs) at the ILs, and lack intact TIR sequences. CiMULE1 has 22 bp and CiMULE2 has 41 bp inverted repeats starting from the 16th bp at both ends (Fig. 1b).

SWO CiMULE1 has five different structural variants with lengths ranging from 6,344 bp to 14,638 bp (Fig. 2a). A total of 56 intact CiMULE1 members (Supplementary Table 1) in DVS are clustered into 17 clusters by sequence similarity and alignment coverage, which are further assigned into 7 groups (Mu1A-G) according to their terminal haplotypes (Fig. 2a). All CiMULE1 members are non-autonomous and don't encode any proteins. CiMULE1 is also detected in the genomes of 7 other Citrus species (Fig. 2b and Supplementary Table 2). Phylogenetic inference and sequence identity analysis (Fig. 2c) suggest CiMULE1 members have been introgressed among Citrus species (Supplementary Note).

SWO CiMULE2 members share > 99.0% identity and are grouped into 3 clusters due to structural variants (Fig. 2d and Supplementary Table 1). Two CiMULE2 clusters, represented by DVS_Mu2_1 and DVS_Mu2_10, are autonomous and encode an intact Mutator family transposase containing four domains, including a Mutator family transposase domain (PF00872) and a SWIM zinc finger domain (PF04434) (Fig. 2e). The transposase coding region of CiMULE2 has been truncated by a 1,752 bp deletion in the DVS_Mu2_2 cluster. Citrus CiMULE2 members share a minimum nucleotide identity of 98.3%. Together with the phylogenetic inference, it is inferred that CiMULE2 could have been obtained through horizontal gene transfer and then dispersed through introgression in Citrus (Supplementary Note).

We have developed an IL scanning pipeline with high accuracy and recall rates for NGS data (Supplementary Note 1). We use the pipeline to analyze resequencing data of 128 SWO accessions

(Supplementary Table 3) and detect a total of 770 CiMULE1 (Supplementary Table 4) and 2,611 CiMULE2 ILs (Supplementary Table 5). More than half of CiMULE1 (560 / 770) (Supplementary Table 6) and CiMULE2 (1,312 / 2,611) (Supplementary Table 7) ILs have ≥ 0.90 mutant ratios in at least one accession and are referred to as the fixed ILs. We find 544 accession-specific CiMULE1 ILs in 36 SWO accessions, including 344 fixed ILs present in 22 accessions. Meanwhile, 2,179 CiMULE2 ILs are accession-specific in 120 SWO accessions, including 974 fixed ILs detected in 109 accessions. Compared to the fixed ILs, the non-fixed ILs of either CiMULE1 (199 / 210) or CiMULE2 (1,204 / 1,298) are significantly ($p < 0.001$ by Fisher's exact test) enriched in accession specific ILs. These accession-specific non-fixed ILs could be tree- or even branch-specific mutations, and we find 26, 8, and 10 branch-specific CiMULE1 or/and CiMULE2 ILs in different branches of the three trees sequenced respectively (Supplementary Table 8).

**CiMULE1 and CiMULE2 variants in SWO, mandarin, and pummelo**

We analyze the abundance of different CiMULE1 and CiMULE2 groups in 128 SWO, 57 mandarin, and 39 pummelo accessions. Only Mu1A and Mu1E are active in SWO, Mu1A, E, and G are active in pummelo, and seven (Mu1A-G) groups have transposition activity in mandarin. Mu1A accounts for 99.2% (745/751) of the SWO novel CiMULE1 ILs, 78.4% (98/125) mandarin-specific ILs, and 72.6% (45/62) pummelo-specific ILs. In SWO, only 4 of the 5 short variants (not including DVS_Mu1_45) of Mu1A have novel ILs, and the long variant DVS_Mu1_9 has no novel IL detected. Mu1E has the second most abundant ILs in SWO, mandarin, and pummelo, though with different proportions (Extended Data Fig. 1).

For CiMULE2 (Extended Data Fig. 2), the DVS_Mu2_2 cluster comprises 10.4% ± 24.8% of the CiMULE2 ILs in SWO, while the remaining belong to DVS_Mu2_1 and DVS_Mu2_10, which are difficult to be distinguished by NGS. DVS_Mu2_2 has similar abundance in mandarin (11.2% ± 15.0%), but higher abundance in pummelo (54.0% ± 26.9%).

**Discrimination and phylogeny of sweet orange cultivars based on CiMULE1 and CiMULE2**

Based on the 1,872 fixed ILs of CiMULE1 (560) and CiMULE2 (1,312), we detect a total of 125 different genotypes among the 128 accessions (Extended Data Fig. 3a). The 125 genotypes are different from each other by an average of 53.1 ± 26.1 ILs (Extended Data Fig. 3b). Only three pairs of sweet orange accessions share the same genotypes, including two Navel oranges (NSL1 and NSL2), two Xuegan accessions (WHXG and WSXG), and two misnamed accessions (WSO and XG, Supplementary Note).

All SWO cultivar groups are separated when the IL-based genotypes are visualized by t-SNE (Fig. 3a). For each SWO cultivar group (with at least 3 accessions), we find 3 to 8 universal and exclusive tag ILs (Fig. 3b and Supplementary Table 9), which may be functionally important in the cultivar group formation. 19 CiMULE1 and 3 CiMULE2 fixed ILs are shared by ≥ 126 sweet orange accessions and referred to as the parental ILs (Fig. 3b and Supplementary Table 9). Further comparison with 57 mandarin and 39 pummelo accessions shows that 5 CiMULE1 parental ILs are from pummelo, 14 CiMULE1 and 2 CiMULE2 parental ILs are mandarin-origin, and the remaining CiMULE2 IL is only detected in SWO. These 22 parental ILs are detected in 122 of the 128 SWO accessions, only 1 parental IL is missing in 5 accessions, and 8 parental ILs are missing in the accession A20, which is inferred as a sexual offspring of SWO (Supplementary Note). The universal existence of the parental ILs and the tag ILs in SWO accessions indicate that the novel ILs are produced with no or rare excisions using an undiscovered manner in the somatic tissues[10].

One CiMULE2 IL (purple bar in Fig. 3b) is denoted as the non-primitive IL since it exists in 124 SWO accessions but is absent in pummelo, mandarin, and four SWO accessions, UKXC, TYC, and two Xuegan (WHXG and WSXG). UKXC possesses all the parental ILs but not a single non-parental IL. WHXG and WSXG possess the parental ILs plus one fixed CiMULE1 IL only present in them. TYC has all the parental ILs and 7 accession specific fixed CiMULE2 ILs. Accordingly, we infer these four accessions to be the closest to the original SWO and denote them as the primitive SWO.

We investigate the phylogenetic relationship among the SWO accessions using their IL-based genotypes. IL-based phylogenetic analysis clusters SWO accessions into cultivar groups with 100% posterior probabilities (Fig. 3c). Several sub-cultivar groups including the Newhall navel orange, Zongcheng (Navel6), Moro blood orange, and Tarocco blood orange are also robustly supported with 100% posterior probabilities. All the non-China origin cultivar groups share two non-China CiMULE2 ILs (black bars in Fig. 3b) and are phylogenetically clustered into a single clade (Fig. 3c). The four Liucheng accessions from south China are clustered into two separate clades, Liu1 (HAL and HAL2) and Liu2 (AL and WHXHC). Liu1 share three other ILs specific in two Egypt accessions and is clustered with them into a robustly supported subclade of the non-China clade, indicating the two Liu1 members may be re-introduced into China. The phylogenetic tree suggests the quick widespread of an SWO lineage has happened twice in SWO cultivation history (Fig. 3c). The first one should have happened in China on an SWO lineage harboring the non-primitive CiMULE2 IL, which is also the common ancestor of modern SWO cultivars/cultivar groups. The second widespread event should have happened shortly after an SWO lineage called “China orange”, which harbored or later obtained the two non-China ILs and became the common ancestor of all non-China origin SWO cultivar groups, was introduced into Europe in the 16$^{th}$ century[1,2].

**Relationship between CiMULE activities and SWO lineage selections**

To compare the activities of CiMULE1 and CiMULE2 among different SWO lineages, we use the weighted fixed IL counts to represent the overall transposition activity in the lineage history, and the weighted non-fixed IL counts to indicate the recent transposition activities in the accessions. A significant positive correlation is observed between fixed and non-fixed (with < 0.9 mutant ratios in all accessions) IL counts for either CiMULE1 or CiMULE2 (Fig. 4a). This implies a general consistency between the historical and recent transposition activities of CiMULE1 and CiMULE2 in SWO cultivars. There is no positive correlation between CiMULE1 and CiMULE2 IL counts (Fig. 4b), indicating they utilize different transposase systems.

We observe significantly higher transposition activities of CiMULE1 in a few SWO accessions and CiMULE2 in most SWO accessions compared to 57 mandarin and 39 pummelo accessions. The primitive SWO accessions have similar weighted fixed or non-fixed CiMULE1 and CiMULE2 IL counts compared to pummelo and mandarin except for TYC, which has enriched (15.0-fold, FDR=1.1E-10) fixed CiMULE2 ILs (Supplementary Table 10). We detect fixed (up to 57.3-fold, FDR < 0.05) and non-fixed (up to 61.25-fold, FDR < 0.05) CiMULE1 IL enrichment in 17 and 8 SWO accessions (Supplementary Table 10). For CiMULE2, all SWO cultivar groups contain accessions with enriched fixed (120 accessions, up to 131.3-fold, FDR < 0.05) and non-fixed (113 accessions, up to 187.19-fold, FDR < 0.05) ILs (Supplementary Table 10). Combined with the phylogenetic tree (Fig. 3c), the first widespread SWO lineage in China most likely already had significantly higher CiMULE2 transposition activity than pummelo and mandarin. For CiMULE1 and CiMULE2, 10/17 and 10/120 of the SWO accessions only have enriched fixed ILs but not non-

fixed ILs (Supplementary Table 10), implying the transposition activities probably have been silenced or reduced in the accessions. 1/8 and 3/113 SWO accessions only harboring enriched non-fixed CiMULE1 and CiMULE2 ILs putatively have gained higher transposition activities recently.

We also find significantly different transposition activities of either CiMULE1 or CiMULE2 among SWO groups (Fig. 4c,d). For CiMULE1, four phylogenetic clades (Fig. 3c) contain accessions with significantly enriched fixed ILs compared to pummelo and mandarin, including ZAOJ, all 7 Blood accessions, both 2 Liu2 accessions, and 7 of 15 Valencia oranges (Supplementary Table 10). These four clades share no common phylogenetic path inferred with CiMULE1 activity increase (indicated by green circles in Fig. 3c), indicating they were selected independently. In Valencia, lineages with high CiMULE1 activities should have been selected at least twice (Fig. 3c).

The transposition activities of CiMULE2 are highly variable among SWO accessions (Fig. 4c,d). Pummelo, mandarin, and primitive SWO are in the statistically homogeneous subset with the lowest fixed IL counts, and the other SWO groups/clades are distributed in four homogeneous subsets with higher mean fixed IL counts (Fig. 4d). Through comparing each SWO accession with the inferred ancestral transposition activities, we detected 19 phylogenetic branches (red circles in Fig. 3c) with significantly enriched fixed CiMULE2 IL counts. The branch of the Jincheng cultivar group has the highest degree of CiMULE2 transposition activity increase observed in SWO. 15 of the 19 branches are internal branches of cultivar groups, including 10 in Navel, 1 in Valencia, 3 in Bingtangcheng, and 1 in Guanggan. Navel orange has a history of no more than 230 years, and its Newhall subgroup that owns two such branches is less than 100 years old, indicating these transposition activity enhancements have happened within hundreds or even tens of years.

**Correlation between transcription and transposition activities of CiMULE2**

We further investigate the transcription activity of CiMULE2 in 629 SWO, 39 mandarin, and 171 pummelo transcriptomes (Supplementary Table 11). We find 41 distinct CiMULE2 transcripts from the assembly of 80 SWO transcriptomes. The open reading frame of CiMULE2 is on the reverse strand, and the transcription starting loci (TSLs) of 10 CiMULE2 transcripts are located within the 3$^{rd}$ to the 176$^{th}$ base at its 3' terminal, while the remaining 31 transcripts have TSLs beyond the CiMULE2 3'-end by up to 1,053 bp (outer TSLs). These results indicate CiMULE2 members most likely utilize their 3'-downstream regions as promoters. We further scanned the 629 SWO transcriptomes for ILs with expression and outer TSLs. As a result, we find up to 6 such ILs in 203 of the 629 SWO transcriptomes (Supplementary Table 12). One parental IL is transcribed with outer TSL in a majority (8/14) of analyzed SWO cultivar groups (Supplementary Table 13). We also detect 15, 9, 7, 1, and 1 group-specific ILs with outer TSLs in Jincheng (including Xianfeng orange), Navel, Valencia, Egypt, and Hamlin, respectively, including one Jincheng tag IL and one Navel orange tag IL (Supplementary Table 13).

Higher expression levels of CiMULE2 are observed in most SWO cultivars/cultivar groups compared to the primitive SWO (Xuegan), mandarin, and pummelo (Fig. 4e). The expression levels of CiMULE2 also differ among the SWO groups (Fig. 4e). We have observed significant positive correlations between CiMULE2 transcript abundance and both the fixed and non-fixed IL counts in the cultivar groups (Fig. 4f). Since CiMULE2 can utilize external transcription starting loci (TSLs) and promoters, every new IL of CiMULE2 has the potential to increase its transcription and transposition activities, which explains why we have observed such a large variance in its transposition activities among SWO accessions.

**CiMULE1 and CiMULE2 play an import role in sweet orange breeding**

The abundance and widespread existence of CiMULE1 and CiMULE2 indicate they may contribute to the horticultural traits of SWO cultivars. We find that the SWO fixed ILs of CiMULE1 and CiMULE2 are significantly (p=3.9E-36 and 1.1E-43 by Chi-squared tests) enriched in gene-related regions (Extended Data Fig. 4), putatively affecting 478 and 1,205 genes (Supplementary Table 14,15), respectively. Among the gene ontology (GO) terms significantly overrepresented (*p* value < 0.05) in the IL-affected genes, multiple development processes are included, such as shoot system development and development processes involved in reproduction (Extended Data Fig. 5). Multiple genes involved in phytohormone signaling pathways are putatively affected by tag ILs of cultivars groups, including brassinosteroids (BRs), gibberellins (GAs), abscisic acid (ABA), and auxin (Supplementary Note).

The high activity of CiMULE1 and CiMULE2 lead to a few pairs of biallelic ILs in SWO cultivars or cultivar groups. Two fixed accession specific CiMULE2 ILs are found in the intron of DVS4B01875 and in the upstream region (within 2 kb if not specifically indicated) of its allelic gene DVS4A01970 in Moro, which are orthologous to *AtBRC2* from Arabidopsis. *AtBRC2* and *AtBRC1* prevent axillary bud outgrowth in Arabidopsis[11], and disruption of *BRC1* orthologs in citrus converted thorns into branches[12]. Thus, these two ILs putatively contribute to the branchy characteristics of the Moro accession. In Campbell Valencia orange, which is more vigorous, thornier, larger, broader-topped, and slower to come into bearing than ordinary Valencia orange[2], two CiMULE1 ILs are detected in the coding regions of two allelic genes of *CsCYP90C1*, that is involved in BR biosynthesis[13]. Two Navel orange tag CiMULE1 ILs are biallelic insertions only 5 bp apart on chr6A and chr6B, 4,165 bp and 4,025 bp upstream of a pair of *CsABCB19* alleles, respectively. *ABCB19* plays important functions in polar auxin transport[14] and multiple development processes[15,16]. Navel oranges have seedless fruit with a 'navel' at the apex and generally larger fruit sizes than other oranges, which may be related to the disturbed auxin signaling[17,18]. Two Jincheng tag ILs are located in the 5'-UTR and upstream region of a pair of allelic genes (DSWO1A01468 and DSWO1B01454), which are orthologous to *AtDLO1*, *AtDLO2*, and *AtDMR6* that act as immunity suppressors in *Arabidopsis*[19], is putatively related to the good storage stability of Jincheng.

Thirty-five of the SWO cultivar tag ILs putatively affect the function of 42 genes (Supplementary Table 14 and Supplementary Note ). In blood orange, a tag CiMULE1 IL is in the 5'-UTR region of a *CsUFGT* (DSWO2A02401), which plays an important role in the biosynthesis of anthocyanins and is related to flower and fruit colors in plants[20,21]. This IL may have contributed to the anthocyanin accumulation in the fruit of Europe origin blood oranges together with another retrotransposon insertion in the promoter region of the *Ruby* gene[5]. One Bingtangcheng tag IL is in the 5'-UTR of *CsNCED1*, a key gene in ABA biosynthesis and related to fruit coloring and ripening[22]. Another Bingtangcheng tag IL in the 5'-UTR of *CsNHX* has been related to the low acid characteristics of Bingtangcheng[1]. Two Valencia tag ILs are located upstream of *CsGID1* and *CsSERK1*, which play important roles in GA[23] and BR[24] signaling respectively.

**Discussion**

Our results imply the high bud sport selection frequency in SWO could be partially explained by the significantly higher CiMULE1 or/and CiMULE2 transposition activities in most SWO lineages. We have observed biallelic insertions in a few SWO cultivars and cultivar groups, which are most likely the results of selections and unlikely to be observed under low transposition activities,

implying the importance of the high CiMULE1 or/and CiMULE2 activities in the formation of these cultivars. Furthermore, the relatively low gene redundancy in the highly heterozygous SWO genome[4] could also increase the probability of phenotype alteration by single-allele mutations.

The selection and dispersion of the ancestral SWO lineage with high CiMULE2 transposition activity could have accelerated SWO breeding and might be the prerequisite for the development of many modern SWO cultivar groups. Though somatic mutations have been hypothesized to be responsible for the diversified SWO cultivars, very few tag mutations of cultivar groups have been discovered before[1,5]. The universal existence of CiMULE1 or/and CiMULE2 tag ILs in cultivar groups and fixed ILs in cultivars suggest they could have played an important role in SWO breeding. Their ILs can also be used for DNA-based identification of almost all SWO cultivars.

We have observed convergent selections of SWO lineages with increased CiMULE1 or CiMULE2 transposition activities, which are not likely accidental or the results of direct selections, but the byproducts of artificial selections of desired traits, namely, indirect selections[7]. The selection of higher transposition activities in SWO could have been speeded up by two factors, the asexual propagation[7] manner and the highly heterozygous genome[4]. SWO has been under bud sports selections for more than 2,000 years[2], resembling continuous adaptative evolution in nature[25]. The dramatic increase of CiMULE1 and CiMULE2 transposition activity has occurred in a short time (tens to hundreds of years) in SWO, indicating they are sensitive to positive selections and could have been used as mutation accelerators in adaptative evolution. Compared to DNA replication and repair machines' modification, TEs are much more controllable in promoting mutation rates and less harmful in the long term, since many organisms have developed epigenetic mechanisms to silence the transposons[26,27]. We have also observed SWO accessions with putatively lowered transposition activities of CiMULE1 and CiMULE2. Moreover, TE insertions are more likely to cause phenotypical alterations compared to single nucleotide substitutions and small indels.

## Methods

### CiMULE1 and CiMULE2 in assemblies

We obtained genome assemblies of 11 SWO[1,4], 7 other *Citrus species*, and *Atalantia buxfoliata* (Supplementary Table 2) from the NCBI assembly database. To detect transposable elements with high activity in SWO, we used DVS[4] as the reference genome and mapped the rest ten SWO assemblies to it using Minimap2 v2.17[28]. Then the large (> 50 bp) indels were called by the paftools.js script of Minimap2 v2.17, and the less than 20 kb inserted and deleted sequences were output for TE identification. To discover putative TEs, we clustered the sequences using CD-HIT v4.8.1[29] requiring 90% nucleotide similarity and 90% minimal alignment coverage (of the longest sequence in each cluster). As a result, we found two MULE families with assembly-specific insertions as a signal of being active in more than a quarter (≥ 3) of the SWO assemblies. The two family members were further searched in the 19 assemblies by blastn, requiring both terminal 100 bp regions aligned with E-value < 1E-3 and the entire TE length spanning < 20 kb. We clustered the detected members using CD-HIT v4.8.1 with requiring no less than 99% nucleotide similarity and 99% minimal alignment coverage. We used the DVS CiMULE clusters to represent the SWO CiMULE clusters, since it has the highest completeness among SWO assemblies[4] and all the variant types detected in other SWO assemblies.

### Phylogeny and terminal haplotypes of CiMULE1 and CiMULE2 in *Citrus*

For either CiMULE1 or CiMULE2, DVS cluster representatives and all members from PTR (C. trifoliata), MSYJ (mandarin), and HWB (pummelo) assemblies were subjected to phylogenetic analysis by both Maximum Likelihood (ML) and Bayesian inference (BI). ClustalX v2.1[30] was applied in the multiple sequence alignment. We carried out ML tree construction using IQ-TREE v2.1.3[31] with automatic model selection and 1,000 bootstrap replicates. The BI method was applied using Mrbayes v3.2.7[32] with the optimal model from IQ-TREE model selection, 110,000 generations of Markov chain Monte Carlo (MCMC) sampling, 1/100 tree sampling frequency and 100 burn-in trees. We visualized the phylogenetic tree using iTOL (Interactive Tree Of Life) v5[33]. The haplotype diversity in the 50 bp aligned regions at both terminals of DVS CiMULE1 and CiMULE2 were summarized from the multiple sequence alignments.

### Transposon feature annotation

The conserved DNA motifs/regions were searched in the terminal 200 bp regions of all DVS CiMULE1 and CiMULE2 members using MEME v5.3.0[34]. Peptides translated from all possible open reading frames from the CiMULE cluster representatives were output via the getorf tool from EMBOSS v6.6.0[35]. Conserved protein domains in the peptides were identified by searching (requiring E-value < 1E-3) the Pfam 34.0 database[36] using HMMER v3.3.2[37].

### Simulations of IL scanning

We carried out *in silico* simulations to learn the impact of the different factors (Extended Data Fig. 6a,b) on IL recall and precise locating. Different combinations of factors (Extended Data Fig. 6c,d,e,f), including the mutant ratio, shortest unique K-Mer lengths (SULs), sequencing depth, sequencing library insert size, and sequencing read length, were simulated using python script by 1,000 iterations per combination, and 1,000 random ILs were tested in each iteration. At least two non-duplicate read pairs were required for either recall or locating an IL in the simulation as in real scanning.

## CiMULE IL scanning in NGS data

We developed a pipeline to call CiMULE1 and CiMULE2 ILs from pair-end NGS data using bash and python scripts. Optimizations made to improve the accuracy and recall rates have been explained in Supplementary Note. The procedures in the pipeline mainly include: (1) Identification of >100 bp continuous regions in the pseudo-haplotype chromosome set DVS_B of DVS with identical allelic regions in DVS_A. Then mask these regions using the ambiguous bas N except for their terminal 50 bp regions (Extended Data Fig. 7); (2) Masking all CiMULE1 and CiMULE2 members in DVS, and add standalone copies of DVS_Mu2_1 and 17 CiMULE1 cluster representatives with duplicate terminal 50 bp regions masked except for one (Extended Data Fig. 7); (3) We analyzed the length distributions of the unique K-Mers in the masked reference DVS genome, the standalone CiMULE terminals, and the surrounding sequences of the ILs detected from the 11 SWO assemblies (Extended Data Fig. 8a,b,c); (4) Masking 10 bp regions surrounding the allelic positions of the mono-allelic reference ILs to avoid false positive IL recalls; (5) Mapping the NGS data to the modified DVS reference genome; (6) Detection of two types of read pairs supporting the ILs (only unique alignments are taken into count): A) read pairs including a read split aligned to and B) read pairs without split-mapped reads but dis-concordantly mapped to both the standalone CiMULE sequences and the chromosomes (Extended Data Fig. 6b); (7) The candidate ILs with fewer than two type A read pairs in all detected samples are filtered, and the rest ILs are regarded as high-quality ILs, which are subjected to genotyping in all analyzed accession. The precise coordinates of the high-quality ILs are inferred using the split-mapped reads; (8) For a type B read pair, the possible range of its IL is inferred based on the read pair orientation and the library insert size, and this range is applied to assign the read pair to a high-quality IL according to their overlapping relationship. When the inferred IL range of a type b read pair is overlapped with two high-quality ILs, it would be assigned to the one with split-mapped reads in the sample or not assigned if no split-mapped read of either IL was present; (9) At least two non-duplicate supporting read pairs (either type A or B but must support the same CiMULE family) are required to recall an IL in a sample; (10) The number of read pairs from the mutant-type ($C_m$ in Extended Data Fig. 6b) and the wild-type ($C_w$) cells on each IL are counted in the mutant samples. Then the mutant ratio is calculated depending on whether and how the allelic region (including the 1 kb surrounding region) of the IL is masked: a) when the allelic region is not masked, the mutant ratio: $(C_m/2) / (C_m/2 + C_w)$; b) when the allelic region is fully masked: $C_m / (C_m/2 + C_w)$; c) when the allelic region is partially masked, the read pairs mapped overlapping the allelic region will be re-mapped to DVS_A, and the count of the effective wild-type read pairs (Extended Data Fig. 6b) will be added to $C_w$, then the mutant ratio is calculated as $C_m / (C_m/2 + C_w)$; (11) For CiMULE1, the terminal haplotype recombination of an IL is inferred according to the CiMULE1 members its reads uniquely mapped to. The ILs are assigned to one of the 17 clusters if the cluster has a unique terminal haplotype combination. Otherwise it will be only assigned to a MU1A-G group accordingly. In the scripts, BEDTools v2.29.2[38] is applied in genome masking and genomic interval overlapping analysis, BWA v0.7.17[39] is applied in read mapping, and Samtools v1.10[40] is used to read the alignment files.

## Resequencing read abundance analysis of CiMULE1 and CiMULE2

For CiMULE1 or CiMULE2, the normalized sequencing read abundance (FPKM, fragments per kilobase region per million mapped read pairs) was calculated based on the reads mapped to the homologous regions shared by all its members. We than carried out Pearson's correlation test between the total IL counts (fixed IL counts plus mutant ratio weighted non-fixed IL counts) and FPKM for CiMULE1 and CiMULE2, respectively (Extended Data Fig. 9).

The ILs of the different CiMULE2 variants could not be distinguished via the terminal haplotypes using NGS data. We first calculated the FPKM of the 1,752 bp deleted segment ($FPKM_d$) in DVS_Mu2_2 and the shared regions ($FPKM_s$) among the three structural variants. The total normalized read abundance of the two long clusters was inferred $FPKM_d$, and the abundance of DVS_Mu2_2 was calculated as ($FPKM_s$ - $FPKM_d$).

**IL-based phylogenetic analysis**

The genotypes of the analyzed accessions on each IL were detected as 1 (possessing) and 0 (not possessing). For SWO cultivar groups including 3 or more accessions, we looked for the tag ILs which are required to be present in > 90% accessions and absent in all accessions from other cultivar groups. In phylogenetic analysis, only fixed ILs present in at least two SWO accessions were applied. We carried out phylogenetic inference on 127 SWO accessions (not including the offspring A20) and two pummelo accessions as outgroup using the BI method by MrBayes v3.2.7, with model assuming equal mutation rates across all ILs, 110,000 generations of MCMC sampling, 1/100 tree sampling frequency and 100 burn-in trees. Branches with less than 50% posterior probabilities were deleted in the phylogenetic tree. The two SWO lineage widespread events were inferred based on the two inner nodes pointing to the multiple China origin and non-China origin cultivar groups, respectively.

**Weighting fixed and non-fixed IL counts**

The counts of fixed ILs were weighted inversely to the number of accessions sharing them among the 127 SWO, 57 mandarin, and 39 pummelo accessions (Extended Data Fig. 10a). As a result, the accession specific fixed ILs would own the highest weights, while the ILs shared by the largest number of accessions would have the lowest. When comparing the fixed IL counts among SWO cultivar groups, an assumption was made that the mean divergence time of the sampled accessions in different cultivar groups were not significantly different. Considering the large difference of fixed ILs in different cultivar groups, the impact of the different divergent time should be limited on our comparison. Moreover, the analyzed mandarin and pummelo accessions were all from sexual hybridizations and have diverged much earlier than the SWO cultivars, making the inference that higher CiMULE transposition activities have been selected in SWO even more robust.

The number of branches and leaves used in sequencing could have an impact on the number of non-fixed ILs observed in the NGS data. To minimize such sampling difference, the counts of non-fixed ILs were weighted using their mutant ratio in the NGS data (Extended Data Fig. 10b).

**Inference of phylogenetic branches with enhanced transposition activity**

We carried out one-way ANOVA (analysis of variance) on weighted counts of fixed and non-fixed CiMULE1 and CiMULE2 ILs among pummelo, mandarin, and SWO groups/clades. The analyzed SWO groups/clades include the primitive SWO group, four Valencia accessions (VAL1) that diverged the earliest in Valencia, and SWO phylogenetic clades including ≥3 accessions and with ≥ 80% posterior probabilities. The distributions of the weighted IL counts in pummelo and mandarin (groups with the largest sample sizes) are right skewed, thus we performed $\log_2$ transformations of the four types of weighted IL counts before one-way ANOVA. The null hypothesis was rejected at $p < 0.001$ level in all four ANOVA tests. We then applied the Student-Newman-Keuls post-hoc tests to detect different homogeneous subsets which include groups with non-significantly different means at the $\alpha = 0.05$ level implemented in IBM® SPSS®

Statistics v26 (IBM, Armonk, USA). To identify SWO accessions with significantly enriched IL counts, we calculated the z-scores and FDRs (right tailed) for each SWO accession based on the population including both pummelo and mandarin accessions on the four types of $\log_2$ transformed weighted IL counts.

We inferred the phylogenetic branches with enhanced transposition activity through comparing the corresponding SWO accessions/clades with the background (ancestral) fixed IL distribution. For CiMULE1, most SWO groups/clades were in the same homogeneous subset with pummelo and mandarin, thus we applied the pummelo and mandarin accessions as the background distribution. Then we detected branches with enhanced CiMULE1 transposition activities as those pointing to SWO accession(s) with significantly (FDR < 0.05) enriched fixed CiMULE1 ILs. For CiMULE2, pummelo, mandarin, and primitive SWO are in the homogeneous subset with the lowest fixed IL counts, and the other SWO groups/clades are distributed in four homogeneous subsets with higher means. We calculated the distances between the four SWO homogeneous subsets and each non-primitive SWO groups from the first and second SWO lineage widespread events using their mean $\log_2$ transformed weighted fixed IL counts (Supplementary Table 16). The 3$^{nd}$ homogeneous subset (indicated by the yellow back group in the left panel of Fig. 4D) has the minimum mean distance from the SWO groups and used to represent the ancestral CiMULE2 transposition activity level (null distribution). Then we calculated the z-scores and FDRs (right tailed) for each SWO accession by comparing them to the 3$^{rd}$ homogeneous subset to detect accessions and branches with significantly (FDR < 0.05) enhanced CiMULE2 transposition activities.

**Transcription analysis of CiMULE2**

We downloaded 740 SWO, 171 pummelo, and 39 mandarin RNA-seq data based on poly(A) + selected RNA from NCBI. We mapped the RNA-seq data to the DVS genome with masked CiMULE2 members in the chromosomes and a standalone DVS_Mu2_1. Transcriptomes including ≥ 20% rRNA reads were regarded as low-quality and excluded from further analysis, leaving 629 SWO, 169 pummelo, and 39 mandarin transcriptomes useable (Supplementary Table 11). To minimize the methodological difference among the dataset, the total effective read count in each transcriptome was calculated as the number of mapped reads minus the rRNA reads. Then the normalized read abundance (FPKM) of CiMULE2 was calculated using the total number of reads uniquely mapped to DVS_Mu2_1 in the reference.

We assembled 80 SWO transcriptome sequencing data[4], and analyzed the transcription staring loci of the transcripts overlapping with DVS CiMULE2 members. To detect CiMULE2 ILs with outer TSL transcription in the 629 SWO transcriptomes, we scanned for the uniquely mapped read pairs overlapping both the CiMULE2 3' terminal 300 bp region and the 3' terminal downstream 500 bp regions at the ILs found in the previous section, and a minimum of 2 read pairs were required to identify the expression of CiMULE2 from an IL in a transcriptome.

**Enrichment analysis of the ILs**

We identified the overlaps between ILs and genic/intergenic regions using BEDTools v2.29.2[38]. The DVS genome contains 46.5% genic (including exonic and intronic) regions and 18.2% gene upstream (2 kb) regions, indicating a random insertion is in the gene-related regions by 64.7%. Chi-squared tests were carried out in the enrichment test of ILs in gene-related regions.

We performed functional enrichment of the IL-affected genes using PANTHER v16[41]. The whole-

genome genes in the not masked regions of DVS were used as the reference set, and the putative IL-affected genes were the test set. Fisher's exact test was applied in the statistical analysis.

## Reference


1 Wang, L. *et al.* Somatic variations led to the selection of acidic and acidless orange cultivars. *Nat. Plants*, doi:10.1038/s41477-021-00941-x (2021).

2 Webber, H. J., Batchelor, L. D. & Reuther, W. *The Citrus Industry*. 1-39 (Univ. California Press, 1967).

3 Wu, G. A. *et al.* Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* **32**, 656–662, doi:10.1038/nbt.2906 (2014).

4 Wu, B. *et al.* A chromosome-level phased Citrus sinensis genome facilitates understanding Huanglongbing tolerance mechanisms at the allelic level in an irradiation induced mutant. *bioRxiv*, 2022.2002.2005.479263, doi:10.1101/2022.02.05.479263 (2022).

5 Butelli, E. *et al.* Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* **24**, 1242–1255, doi:10.1105/tpc.111.095232 (2012).

6 Shamel, A. D. & Pomeroy, C. S. Bud mutations in horticultural crops. *J. Hered.* **27**, 487–494, doi:10.1093/oxfordjournals.jhered.a104171 (1936).

7 Raynes, Y., Wylie, C. S., Sniegowski, P. D. & Weinreich, D. M. Sign of selection on mutation rate modifiers depends on population size. *Proc. Natl. Acad. Sci. U.S.A* **115**, 3422–3427, doi:10.1073/pnas.1715996115 (2018).

8 Foster, T. M. & Aranzana, M. J. Attention sports fans! The far-reaching contributions of bud sport mutants to horticulture and plant biology. *Horticulture research* **5**, 44, doi:10.1038/s41438-018-0062-x (2018).

9 Robertson, D. S. Characterization of a mutator system in maize. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **51**, 21–28, doi:10.1016/0027-5107(78)90004-0 (1978).

10 Liu, K. & Wessler, S. R. Transposition of Mutator-like transposable elements (MULEs) resembles hAT and Transib elements and V(D)J recombination. *Nucleic Acids Res.* **45**, 6644–6655, doi:10.1093/nar/gkx357 (2017).

11 Aguilar-Martínez, J. A., Poza-Carrión, C. & Cubas, P. Arabidopsis BRANCHED1 acts as an integrator of branching signals within axillary buds. *Plant Cell* **19**, 458–472, doi:10.1105/tpc.106.048934 (2007).

12 Zhang, F. *et al.* Reprogramming of Stem Cell Activity to Convert Thorns into Branches. *Current biology : CB* **30**, 2951-2961.e2955, doi:10.1016/j.cub.2020.05.068 (2020).

13 Ohnishi, T. *et al.* C-23 hydroxylation by Arabidopsis CYP90C1 and CYP90D1 reveals a novel shortcut in brassinosteroid biosynthesis. *Plant Cell* **18**, 3275–3288, doi:10.1105/tpc.106.045443 (2006).

14 Titapiwatanakun, B. *et al.* ABCB19/PGP19 stabilises PIN1 in membrane microdomains in Arabidopsis. *The Plant journal : for cell and molecular biology* **57**, 27–44, doi:10.1111/j.1365-313X.2008.03668.x (2009).

15 Cecchetti, V. *et al.* ABCB1 and ABCB19 auxin transporters have synergistic effects on early and late Arabidopsis anther development. *Journal of integrative plant biology* **57**, 1089–1098, doi:10.1111/jipb.12332 (2015).

16 Okamoto, K. *et al.* An ABC transporter B family protein, ABCB19, is required for cytoplasmic streaming and gravitropism of the inflorescence stems. *Plant signaling & behavior* **11**, e1010947, doi:10.1080/15592324.2015.1010947 (2016).

17 Agustí, M. *et al.* The synthetic auxin 3,5,6-TPA stimulates carbohydrate accumulation and growth in citrus fruit. *Plant Growth Regulation* **36**, 141-147, doi:10.1023/A:1015077508675 (2002).
18 Bermejo, A. *et al.* Auxin and Gibberellin Interact in Citrus Fruit Set. *J. Plant Growth Regul.* **37**, 491–501, doi:10.1007/s00344-017-9748-9 (2018).
19 Zeilmaker, T. *et al.* DOWNY MILDEW RESISTANT 6 and DMR6-LIKE OXYGENASE 1 are partially redundant but distinct suppressors of immunity in Arabidopsis. *Plant J.* **81**, 210-222, doi:10.1111/tpj.12719 (2015).
20 Matus, J. T. *et al.* A group of grapevine MYBA transcription factors located in chromosome 14 control anthocyanin synthesis in vegetative organs with different specificities compared with the berry color locus. *Plant J.* **91**, 220-236, doi:10.1111/tpj.13558 (2017).
21 Sun, W. *et al.* Biochemical and Molecular Characterization of a Flavonoid 3-O-glycosyltransferase Responsible for Anthocyanins and Flavonols Biosynthesis in Freesia hybrida. *Front. Plant Sci.* **7** (2016).
22 Alquezar, B., Rodrigo, M. J., Lado, J. & Zacarías, L. A comparative physiological and transcriptional study of carotenoid biosynthesis in white and red grapefruit (Citrus paradisi Macf.). *Tree Genet. Genom.* **9**, 1257-1269, doi:10.1007/s11295-013-0635-7 (2013).
23 Griffiths, J. *et al.* Genetic characterization and functional analysis of the GID1 gibberellin receptors in Arabidopsis. *Plant Cell* **18**, 3399–3414, doi:10.1105/tpc.106.047415 (2006).
24 Gou, X. *et al.* Genetic evidence for an indispensable role of somatic embryogenesis receptor kinases in brassinosteroid signaling. *PLoS Genet.* **8**, e1002452, doi:10.1371/journal.pgen.1002452 (2012).
25 Gregory, T. R. Artificial Selection and Domestication: Modern Lessons from Darwin's Enduring Analogy. *Evolution: Education and Outreach* **2**, 5–27, doi:10.1007/s12052-008-0114-z (2009).
26 Seczynska, M., Bloor, S., Cuesta, S. M. & Lehner, P. J. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature*, doi:10.1038/s41586-021-04228-1 (2021).
27 Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature reviews. Genetics* **8**, 272–285, doi:10.1038/nrg2072 (2007).
28 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, doi:10.1093/bioinformatics/bty191 (2018).
29 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659, doi:10.1093/bioinformatics/btl158 (2006).
30 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)* **23**, 2947–2948, doi:10.1093/bioinformatics/btm404 (2007).
31 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534, doi:10.1093/molbev/msaa015 (2020).
32 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542, doi:10.1093/sysbio/sys029 (2012).
33 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293-W296, doi:10.1093/nar/gkab301 (2021).
34 Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43**, W39-49, doi:10.1093/nar/gkv416 (2015).
35 Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open

Software Suite. *Trends Genet.* **16**, 276–277, doi:10.1016/s0168-9525(00)02024-2 (2000).

36 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412-D419, doi:10.1093/nar/gkaa913 (2021).

37 Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121, doi:10.1093/nar/gkt263 (2013).

38 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, doi:10.1093/bioinformatics/btq033 (2010).

39 Li, H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. (2013).

40 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

41 Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394-D403, doi:10.1093/nar/gkaa1106 (2021).

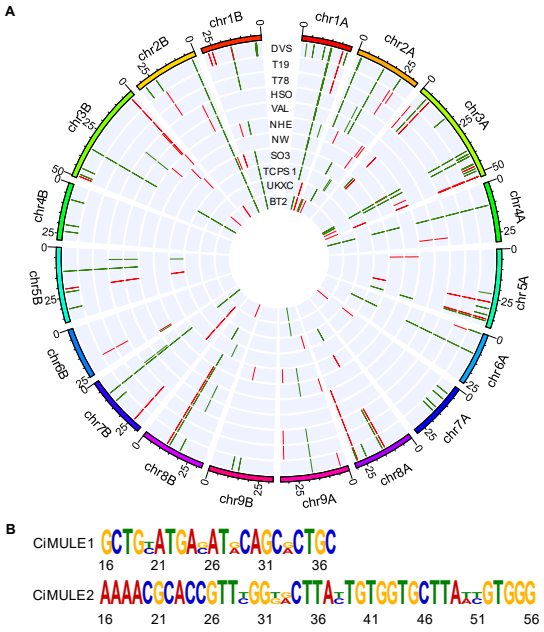


**Fig. 1. CiMULE1 and CiMULE2 members in sweet orange assemblies and their terminal inverted repeats. a**, Insertion loci of CiMULE1 and CiMULE2 in 11 sweet orange assemblies. The ideogram of the phased sweet orange reference genome (chr1-9A, B) is shown outmost, and the unit of the tick marks is million base pairs (Mb). Assemblies from our previous study[3]: DVS, a phased ordinary Valencia sweet orange assembly (the reference); T19 and T78, assemblies of two irradiation-induced Valencia sweet orange mutants. Assemblies from the study of Wang et al. (2021)[4]: HSO, di-haploid Valencia sweet orange genome; VAL and SO3, diploid Valencia sweet orange assemblies; NHE and NW, navel oranges; TCPS1, UKXC, and BT2, cultivars from South China. The green and red bars denote the ILs of CiMULE1 and CiMULE2, respectively. **b**, Sequence logos of the inverted segmental repeats starting from the 16$^{th}$ bp at both ends of CiMULE1 and CiMULE2.

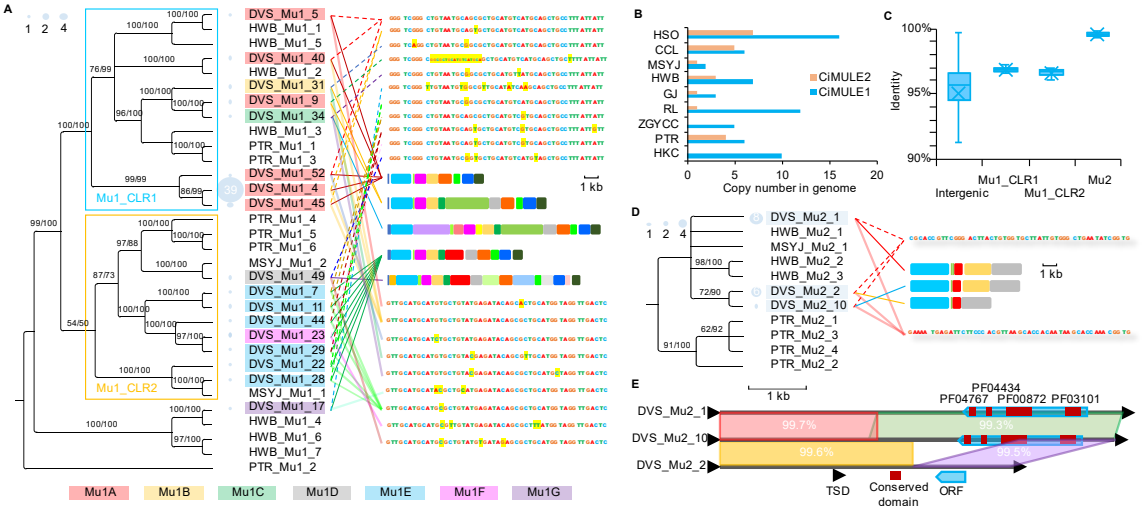


**Fig. 2: Phylogenetic relationship and characteristics of *Citrus* CiMULE1 and CiMULE2 families. a** and **d**, The midpoint rooted phylogenetic trees (left), structural variants (middle right), and 50 bp terminal haplotypes (top and bottom right) of CiMULE1 and CiMULE2, respectively. CiMULE1 and CiMULE2 members from genomes of PTR (*Citrus trifoliata*), MSYJ (*Citrus reticulata*), and HWB (*Citrus maxima*), and 17 CiMULE1 and 3 CiMULE2 cluster representatives from the diploid Valencia sweet orange (DVS) genome were applied in the phylogenetic analyses. The bubbles left to the node names show the sizes of the corresponding clusters in DVS. Only branches supported by > 50 percent bootstrap (the number before slashes) tests using the Maximum likelihood method and with > 50 percent probabilities (after slashes) by Bayesian inference are shown in the phylogenetic trees. For the structural variant diagrams of either mutator family, the bins of the same colors denote homologous regions. The two CiMULE1 clusters (Mu1_CLR1 and Mu2_CLR2) are indicated by the blue and yellow frames. The CiMULE1 members from DVS are assigned into seven subgroups (Mu1A-G, highlighted with distinct background colors) sharing no terminal haplotype for insertion locus scanning. In the terminal haplotypes, minor alleles of the variants are highlighted with yellow backgrounds. **b**, Copy numbers of CiMULE1 and CiMULE2 members in the genomes of 8 *Citrus* species and *Atalantia buxfoliata* (Supplementary Table T7). HSO, di-haploid *Citrus* × *sinensis*; CCL, *Citrus* × *clementina*; GJ, *Citrus japonica*; RL, *Citrus medica*; ZGYCC, *Citrus ichangensis*; HKC, *Atalantia buxfoliata*. **c**, Boxplots showing the nucleotide identity distributions among PTR and DVS CiMULE1 and CiMULE2 members and other intergenic orthologous regions. The cross symbols denote the mean values, and the outliers are shown as dots. **e**, Alignment and conserved protein domains of CiMULE2 cluster representatives from DVS. Homologous regions between the members are connected with colorful parallelograms with nucleotide identities shown inside. TSD, target site duplication; ORF, open reading frame. Conserved protein domains (from the Pfam 34.0 database): PF04767, Pox_F17, DNA-binding phosphoprotein; PF04434, SWIM zinc finger; PF00872, Mutator family transposase; PF03101, FAR1 DNA-binding domain.

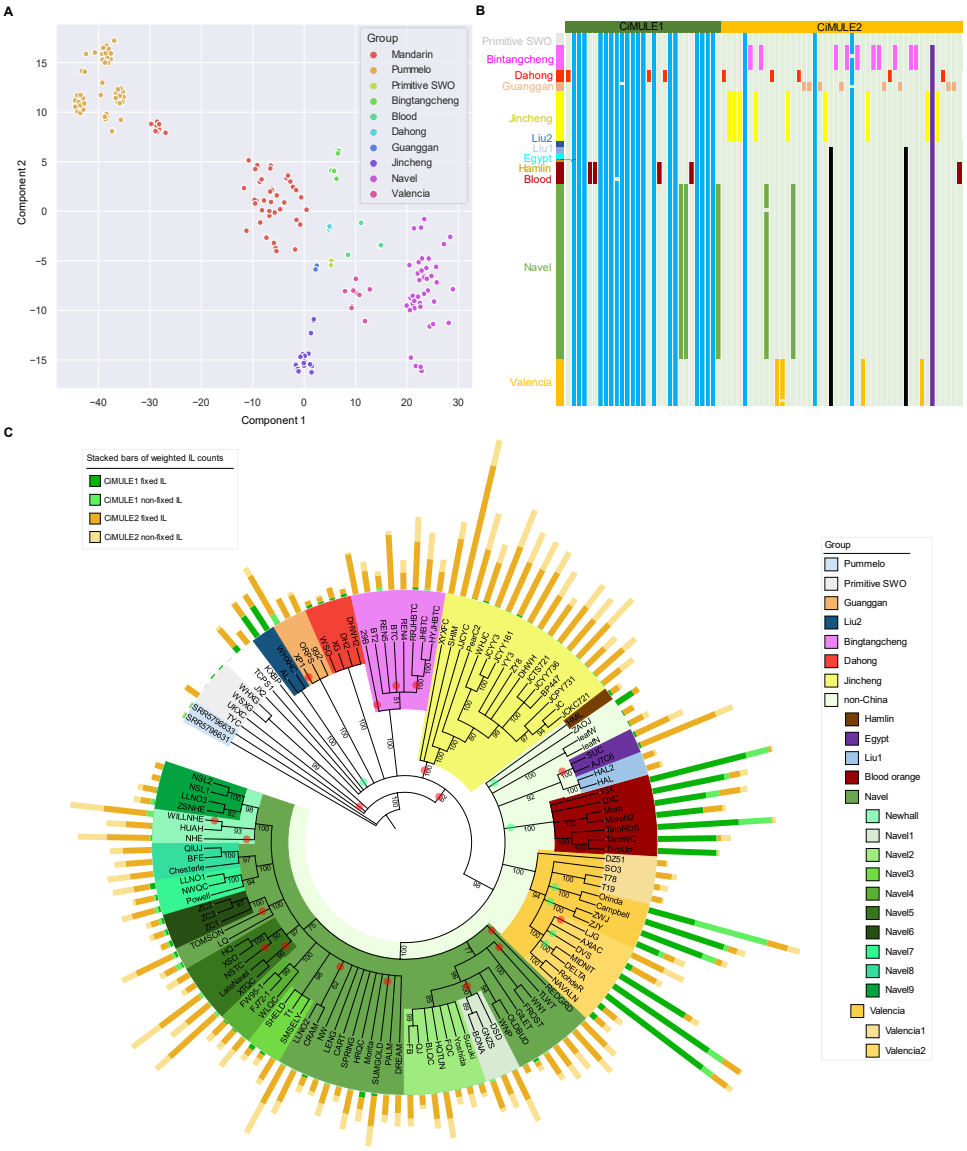


**Fig. 3: CiMULE insertion locus (IL)-based sweet orange (SWO) cultivar discrimination, phylogenetic inference, and transposition activity comparison. a**, t-distributed stochastic neighbor embedding (t-SNE) visualization of CiMULE IL-based sweet orange, mandarin, and pummelo genotypes. 115 SWO accessions from groups containing ⩾ 3 accessions, 49 mandarin accessions, and 62 pummelo accessions were subjected to analysis. **b**, Genotypes of the SWO accessions on parental and tag ILs. Each row denotes a sweet orange accession, and each column indicates an IL in the barcode region. Light green denotes the absence of the IL in the corresponding accession, while all other colors denote the presence. At least three accessions were required for tag ILs to be identified in a cultivar group, and the tag ILs are colored the same as the cultivar groups in the graph. The blue bars denote the parental ILs. The two black bars are the two ILs shared by all non-China origin accessions. The purple bar indicates the non-primitive IL absent in the four primitive SWO accessions. **c**, IL-based phylogenetic tree and the weighted IL counts in each accession. 553 fixed ILs present in at least two sweet orange accessions were applied in phylogenetic tree construction by Bayesian inference. The displayed tree is a consensus tree of 1,000 trees from 100,000 iterations of sampling. Branches with < 50% posterior probabilities have been deleted, and the numbers on the branches denote the posterior probabilities (%). The weighted fixed and non-fixed CiMULE1 and CiMULE2 IL counts have been drawn as stacked bars for each accession. Robustly supported (≥ 80% posterior probabilities) groups or clades are highlighted with background colors as listed in the legend except for primitive SWO and Valencia1. The green and red circles indicate the branches with inferred increased CiMULE1 and CiMULE2 transposition activities, respectively.

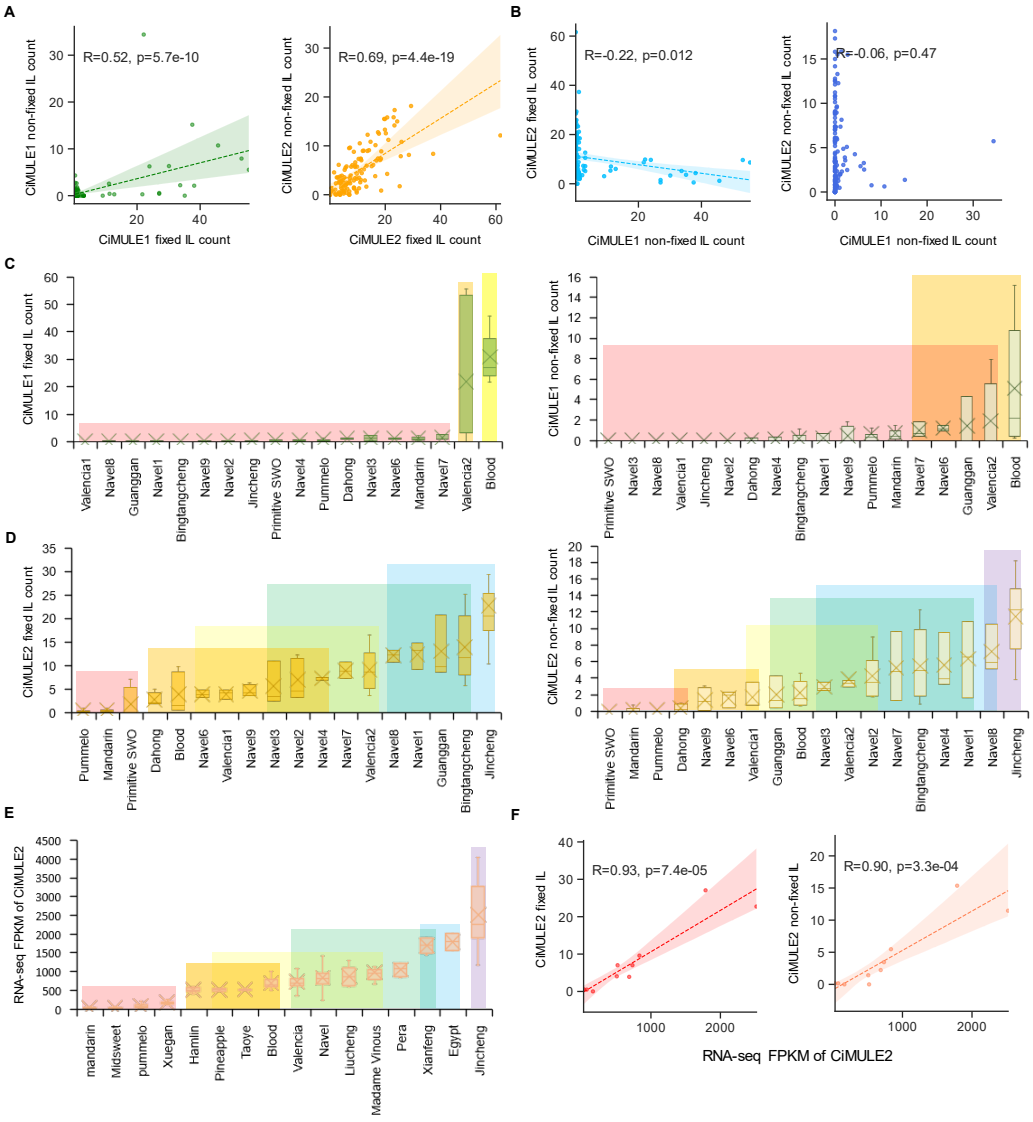


**Fig. 4: Distributions of CiMULE insertion locus (IL) counts and CiMULE2 expression abundance.** The weighted fixed and non-fixed IL counts are used in the graphs. In the dot plots (**a**, **b**, and **f**), the linear regression lines (dashed) and the 95% confidence intervals (shadows) are depicted if significant ($p < 0.05$) correlation is observed. In the box plots, the differently colored background squares indicate distinct homogeneous subsets at the $\alpha = 0.05$ level from post hoc tests of one-way analysis of variance (ANOVA). **a**, Correlation between fixed and non-fixed IL counts in 127 sweet orange accessions. Left panel, CiMULE1 ILs; right panel, CiMULE2 ILs. **b**, Correlation between CiMULE1 and CiMULE2 IL counts in the sweet orange accessions. Left panel, fixed ILs; right panel, non-fixed ILs. **c**, Box plot showing the distribution of fixed (left) and non-fixed (right) CiMULE1 IL counts in different phylogeny clades. The clades are the same as those described in Fig. 3C except for Pummelo (including 49 pummelo accessions) and Mandarin (including 62 mandarin accessions). The mean IL counts are indicated by the forks. **d**, Distribution of fixed (left) and non-fixed (right) CiMULE2 IL counts in the phylogeny clades. **e**, Normalized RNA-seq abundance of CiMULE2 in mandarin, pummelo, and multiple sweet orange cultivar groups. FPKM, fragments per kilobase of transcript per million mapped reads. Two primitive SWO accessions, WHXG and WSXG, belong to the Xuegan group, and another primitive SWO accession (TYC) belongs to the Taoye group. **f**, Correlations between the mean RNA-seq abundance and the mean fixed (left) and non-fixed (right) CiMULE2 IL counts in mandarin, pummelo, and eight SWO cultivar groups. The eight cultivar groups are Xuegan, Hamlin, Taoye, Blood, Valencia, Navel, Egypt, and Jincheng.