The role of transposon activity in sweet orange breeding

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: +01 864 656 7552. Email: luofeng@clemson.edu

Abstract

Sweet orange (SWO) has multiple asexually propagated lineages under artificial selections for tens to hundreds of years[1] and the most bud sports selected in *Citrus*[2]. The impact of artificial selections on SWO's mutation rates and the reason for its high bud sport frequency is unknown. Here, we find six transposon families with up to 12,706.2-fold higher transposition activity in SWO compared to its two parental species (pummelo and mandarin). These six boosted families have contributed to 98.0% SWO-specific transposon insertions and modified almost all SWO genomes. Based on their insertions, we have discovered three major spread events in SWO cultivation history, dated modern SWOs back to a common ancestor around 500 years ago, and inferred Hunan province of China to be the cradle of SWO. The formation of most SWO cultivar groups involve the activation of at least one boosted transposon family. The activation of CiMULE2 occurred in the common ancestor of modern SWOs, and it has produced tag insertions (with universal and exclusive presence) for all analyzed cultivar groups. SWO lineages with transposon activations and higher transposition activities have been selected repeatedly during SWO's asexual breeding. Noticeably, the transposition activities of active DNA TE families increased several folds within tens of years in a few SWO lineages. Transposons function as accelerators in the asexual breeding of SWO and could have played the same role in adaptive evolution.

Main

The original sweet orange (*Citrus sinensis*, SWO) plant has been inferred to be derived from inter-specific hybridizations between pummelo (*Citrus maxima*) and mandarin (*Citrus reticulata*)[3,4], but when and where have been unknown. Although the Chinese character '橙' (orange) was mentioned as early as 2nd century BC[1], it usually referred to *Citrus* × *aurantium* L. or *Citrus* × *junos* in ancient Chinese literatures rather than *C. sinensis*. Even in *Ben Cao Gang Mu*, a book completed in the year 1578 AD, Li Shi-Zhen described orange fruit as sour but mandarin fruit as sweet and sour, indicating sweet orange was still not the best-known orange then.

SWO has been propagated asexually through apomixis and grafting, and its diversified cultivars have arisen through somatic mutations from the original plant[5]. SWO has one of the largest number of bud sport selections in *Citrus* with diverse horticultural traits[4,6,7], reaching more than 1,300 recorded accessions already in 1936[8]. However, the reason for the high bud sport selection frequency in SWO is not well understood. 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 the mutation activities[9].

Transposable element (TE) is one of the important driving forces of mutation in plants[10]. Previously, TE has shown to be active in SWO. In three Valencia sweet oranges, a majority of spontaneous structural mutations were Mutator-like transposable element (MULE) insertions[4]. Wang et al. (2021) detected 877 TE insertions including 266 MULEs insertions in 114 SWO accessions[6]. However, the role of TEs in SWO asexual breeding remains unclear. Here, we completely survey the TE activities and identify six highly active TE families in SWO. We further examine the contribution of active TE families to the SWO bud sport diversity, as well as the relationship between artificial selections and TE activities in SWO.

Results

Highly active transposable element families in sweet orange

We identify 34 putatively active TE families in SWO through comparing 11 SWO assemblies (Supplementary Figure 1a)[4,6], including MULE, CACTA, Harbinger, and hAT type DNA transposons and LTR (long terminal repeat) type retrotransposons (Supplementary Table 1). To investigate their activities, we develop a TE insertion locus (IL) scanning pipeline with over 95% recall and precision rates for next-generation sequencing (NGS) data (Methods). We then use it to scan SWO, pummelo and mandarin genomes and detect 4,780 SWO-specific, 12,309 mandarin-specific, and 4,900 pummelo-specific ILs in 127 SWO, 57 mandarin and 39 pummelo accessions, respectively. On average, a SWO accession possesses 37.7 species-specific (novel) ILs that are not detected in any pummelo or mandarin.

IL enrichment tests revealed six boosted TE families in SWO with significantly enhanced transposition activities compared to their activities in mandarin and pummelo, including 5 DNA transposon families (4 MULEs and 1 hAT) and 1 LTR retrotransposon family

(CiGYPSY1). These six TE families have contributed to 98.0% of the species-specific (novel) ILs in SWO, but only 25.4% and 24.8% in mandarin and pummelo, respectively. CiMULE2 alone has produced more than half (54.3%, 2,596) novel TE insertions in SWO, while CiMULE1, CiMULE3, CihAT1, and CiMULE5 have contributed to 751, 511, 468, and 309 novel SWO ILs, respectively. CiMULE2 is estimated to be on average 3,218.6- and 12,706.2-fold more active in SWO than in mandarin and pummelo. The ILs of CiMULE1, CiHAT1, and CiMULE3 are enriched by 59.8- to 1,979.3-fold in SWO compared to the two species. CiMULE5, the most active TE family in mandarin and pummelo (with > 20% novel IL contributions), has 7.0- and 9.6-fold IL enrichment in SWO, respectively. Among the six TE families, CiGYPSY1 has the fewest (48) novel ILs in SWO, and it has approximately 8.5- and 25.7-fold IL enrichment compared to mandarin and pummelo.

Revelation of sweet orange breeding history using TE insertions

We unravel the lineage development and spread history of SWO using the ILs of the six TE families (Figure 1A). Our IL-based phylogenetic tree better supports the SWO (sub)cultivar groups than the somatic single nucleotide mutations[11]. The results show that there have been three major lineage-spread events in the SWO cultivation history (Figure 1B). The first spread event gave rise to the primitive SWOs, including UKXC, TYC, and the Xuegan (WHXG and WSXG) group, and the common ancestor of all the rest analyzed SWO cultivars (referred to as the modern SWO). UKXC, an accession collected from Hunan province, is the closest to the inferred original SWO with only one novel CiGXPSY1 IL difference. TYC and Xuegan originated in the neighboring provinces of Hunan in China, respectively.

In the second spreading event, which is estimated to have occurred between 1470 and 1533 AD, the common ancestor of modern SWO was distributed in China. This inferred ancestor harbored three ILs (called the non-primitive ILs) that are presented in all the modern SWO accessions but not the primitive SWOs, including one from each of CiMULE2, CiHAT1, and CiGXPSY1. This spread event should be quick, and multiple SWO lineages were selected independently afterwards, since the ancestors of China-local cultivar groups (CQ, Dahong, Guanggan, and Bingtangcheng) and the common ancestor of all non-China origin groups share no novel IL other than the three non-primitive ILs. JX2 is the closest to this ancestor with only 4 novel CiMULE5 ILs difference, which is a local cultivar of Jing Xian, Hunan province. Hunan is also the province of origin for Bingtangcheng and Dahong, and geographically adjacent to where CQ and Guanggan originated. Accordingly, Hunan province is most likely the cradle of SWO and the center of its first and second main spread events.

The third spreading event involved a SWO lineage harboring two novel CiMULE2 ILs (called nC-ILs), which led to all the non-China origin SWO groups, including Navel, Valencia, Blood, Hamlin, and Egypt. This event happened soon after the introduction of the first modern sweet orange (Portugal orange) in Portugal in 1635, which was later widely distributed around the world. Consistent with the somatic single nucleotide mutation-based phylogeny[12], our IL-based phylogenetic tree also supports the monophyly

of the modern non-China origin cultivars. Moreover, a few non-China origin lineages have been introduced into China, including several Navel orange lineages and two Liucheng accessions (Liu1) (Figure 1B). Liu1 harbor the two nC-ILs and share several ILs specifically with lineages from Egypt, implying a Mediterranean area origin.

Characterization of sweet orange genomes with Tag TE insertions

Tag-mutations potentially play a role in the phenotype formation of the cultivar groups but have seldom been identified in SWO before[7,11]. Here, we have found group tag-ILs, which are TE insertions presenting only in all accessions of a specific cultivar group, of at least one of the boosted TE families for all analyzed SWO cultivar groups, including Navel, Valencia, Blood, Egypt, Jincheng, Dahong, Guanggan, and Bingtangcheng (Supplementary Table 9). The most active CiMULE2 has contributed 1 (Blood) to 11 (Bingtangcheng) group tag-ILs to all eight analyzed cultivar groups. Group Tag-ILs of CiMULE1, CiMULE3, CiMULE5, and CihAT1 are also identified for 4, 1, 3, and 4 cultivar groups, respectively. CiGYPSY1 has produced group tag-ILs for 7 cultivar groups although it has smallest novel IL count. These tag-ILs could serve as molecular markers for identifying the cultivar group of SWO accessions. All the 127 analyzed sweet orange accessions are discriminated by the CiMULE2 accession tag ILs (TE insertions presenting only in a specific cultivar) except for two Xuegan cultivars. The rest five TE families have also produced accession tag ILs in distinct cultivar groups that could identify specific SWO cultivars (Supplemental Table X).

TE activity in sweet orange breeding

We infer the IL accumulation rates of the six TE families in the original SWO based on the primitive SWOs. The five DNA transposon families have been silent (zero IL accumulation rate) in at least half of the primitive SWO accessions. UKXC, the SWO accession closest to the original SWO, doesn't own any novel IL of the five DNA TE families. Thus, we infer that the five boosted DNA TE families had similarly low transposition activity in the original SWO as in pummelo and mandarin, and acquired significantly enhanced transposition activities afterwards. CiGYPSY1 is the only boosted TE family with novel IL accumulation in all the primitive SWOs. It has maintained similar transposition accumulation rates of 0.0027 ± 0.0019 novel ILs / year in most SWO lineages, implying putatively higher CiGYPSY1 activity in the original SWO plant than the average activity levels in pummelo and mandarin. Meanwhile, the novel CiGYPSY1 ILs are significantly ($p < 0.001$ by Fisher's exact test) enriched in cultivar group tag ILs compared to the five DNA transposon families. This result indicates that the selection of lineages with CiGYPSY1 ILs is another reason for the relatively abundant novel CiGYPSY1 ILs in SWO.

We have observed more than one activation events of each boosted DNA TE family in SWO lineages. CiMULE2 activation has occurred independently in two primitive SWO accessions, the common ancestor of all modern SWOs and TYC. The formation of each modern SWO cultivar group involves the activation of at least one of the rest four

boosted DNA TE families. CiMULE1 and CiHAT1 have been independently activated in the ancestors of the Dahong, Jincheng, Navel, and Blood groups. CiMULE5 activation has occurred in the ancestors of the Navel, Blood, and Egypt groups. CiMULE3 has been activated in the ancestor of the CQ group. We also observed their activations after cultivar group formation. For instance, CiMULE1, CiMULE3, and CihAT1 were not active in the common ancestor of Valencia but became active in independently selected Valencia lineages.

We also find that the IL accumulation rates of the active DNA TE families increased dramatically within tens to hundreds of years in some SWO lineages. The overall IL accumulation rate of the five boosted TE families in modern SWOs vary from 0.008 (JX2) to over 1.5 (some Jincheng accessions). We have observed up to 20-fold difference of CiMULE2 IL accumulation rates among distinct SWO clades (Figure 2A). In Guanggan, Bingtangcheng, Jincheng, Valencia, Navel, and the Egypt group, we have detected independently selected SWO lineages with significantly enhanced CiMULE2 transposition activities compared to the common ancestor of modern SWOs. Sub-clades of the same cultivar groups could develop diverged IL accumulation rates within tens to hundreds of years. The Jincheng cultivar group has a documented history of ~80 years, and the IL accumulation rates of CiMULE2 and CiMULE3 have been increased by up to 3-fold in its three sub-clades. Navel orange sub-clades have up-to 5-fold IL accumulation rate increase in the past ~147 years. Similar cases also occurred on CiMULE1 and CiHAT1 in the Blood group, and on CiMULE2 in the Bingtangcheng, Blood, Valencia, and Egypt groups.

Furthermore, as a genome protection mechanism, transposon silencing has worked differently on the five boosted DNA TE families after their activations. CiMULE2 has a <1% silencing ratio in modern SWO accessions, and CiMULE3 has a zero percent silencing ratio in Jincheng (Figure 2 and Supplementary Figure XX). On the opposite, CiMULE5 is the only TE family being almost completely silenced in the cultivar groups (Navel, Blood and Egypt) after being activated in the group ancestors (Figure 2C). CiMULE1 and CiHAT1 were silenced in a majority of Jincheng, Dahong and Navel oranges, but none of the Blood oranges.

Correlation between CiMULE2 transcription and transposition

CiMULE2 is the only autonomous TE family among the six TE families. We investigate the transcription activity of CiMULE2 in 629 SWO, 39 mandarin, and 171 pummelo transcriptomes to study its relationship with transposition activity (Supplementary Table 11). We detect significantly higher expression levels of CiMULE2 in most SWO cultivars/cultivar groups compared to the primitive Xuegan, mandarin, and pummelo. The expression levels of CiMULE2 are similarly diversified among the SWO groups as its transposition activity levels, and we have observed significant positive correlations between CiMULE2 expression abundance and the IL counts in the cultivar groups. Moreover, we find that CiMULE2 can utilize external transcription starting loci (TSLs) and promoters. Accordingly, every new IL of CiMULE2 has the potential to increase its

expression and transposition activity levels and has resulted in the accelerated transposition activity levels observed among SWO accessions.

Enrichment of transposon insertions impacting plant development

Impact of TE insertions on horticultural traits

The universal existence of the tag ILs and the enrichment of transposon activation events during cultivar group formation implied that the six boosted TE families have played essential roles in SWO breeding. Their ILs have putatively affected a total of 2,955 genes in SWO (Supplementary Table 14,15). These affected genes are significantly enriched in plant development processes and phytohormone signaling pathways (Figure 3), which are closely related with citrus horticultural traits.

Thirty-five of the SWO cultivar group tag ILs putatively affect the function of 42 genes (Supplementary Table 14). In blood orange, a tag CiMULE1 IL is in the 5'-UTR region of the *CsUFGT* (DSWO2A02401), which plays an important role in the biosynthesis of anthocyanins and is related to flower and fruit colors in plants[13,14]. 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[7]. One Bingtangcheng tag IL is in the 5'-UTR of *CsNCED1*, a key gene in ABA biosynthesis and related to fruit coloring and ripening[15]. Another Bingtangcheng tag IL in the 5'-UTR of *CsNHX*, which has been related to the low acid characteristics of Bingtangcheng[6]. Two Valencia tag ILs are located at the upstream of *CsGID1* and *CsSERK1*, which play important roles in gibberellins[16] and brassinosteroids[17] signaling respectively. Several other cultivar group tag ILs are located upstream of or inside genes related to phytohormones, including brassinosteroids, gibberellins, abscisic acid, and auxin.

We have observed several pairs of biallelic ILs in SWO cultivars or cultivar groups. Two 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*[18], and disruption of *BRC1* orthologs in citrus converted thorns into branches[19]. 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[1], two CiMULE1 ILs are detected in the coding regions of two allelic genes of *CsCYP90C1*, which is involved in brassinosteroids biosynthesis[20]. 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[21] and multiple development processes[22,23]. 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[24,25]. 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*[26], is putatively related to the good storage stability of Jincheng.

Discussion

Although somatic mutations have been hypothesized to be responsible for the diversified SWO cultivars, very few tag mutations of cultivar groups have been discovered before[6,7]. Our research reveals that TE insertions are present as tag mutations for all SWO cultivar groups and significantly enriched in affecting horticultural traits. Our results show that the high bud sport selection frequency of SWO could be basically explained by the activation of boosted TE families in SWO. The overall IL accumulation rates in modern SWO are significantly higher than those in the original SWO. This high transposition activity level has obviously accelerated SWO breeding and plays an essential role in modern SWO industry. Most significantly, the activation of CiMULE2 most likely has been the prerequisite for the development of many modern SWO cultivar groups.

This study indicates that transposons may play an important role in adaptive evolution. It has long been hypothesized that positive selections are prone to select lineages with higher mutation rates indirectly in asexual organisms, which has only been proven on microorganisms. SWO has been under bud sports selections for over 500 years[1], resembling continuous adaptive evolution in nature[27]. The increased transposition activities in SWO lineages are inferred to be the byproducts of repeated artificial selections of desired traits[9]. The dramatic increase of the IL accumulation rates in SWO has occurred in a short time (tens to hundreds of years), suggesting the involved TE families are very sensitive to the selections and have actually functioned as mutation accelerators under the artificial selections. Compared to other mechanisms altering mutation rates such as mutations on the DNA replication and repair machines, 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[28,29], which is observed in SWO.

Methods

CiMULE1 and CiMULE2 in assemblies

We obtained genome assemblies of 11 SWO[4,6], 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[30]. 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[31] 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[32] was applied in the multiple sequence alignment. We carried out ML tree construction using IQ-TREE v2.1.3[33] with automatic model selection and 1,000 bootstrap replicates. The BI method was applied using Mrbayes v3.2.7[34] 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[35]. 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[36]. Peptides translated from all possible open reading frames from the CiMULE cluster representatives were output via the getorf tool from EMBOSS v6.6.0[37]. Conserved protein domains in the peptides were identified by searching (requiring E-value < 1E-3) the Pfam 34.0 database[38] using HMMER v3.3.2[39].

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 1. 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 disconcordantly 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[40] is applied in genome masking and genomic interval overlapping analysis, BWA v0.7.17[41] is applied in read mapping, and Samtools v1.10[42] 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. https://en.wikipedia.org/wiki/Analysis_of_variance

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 3nd 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 3rd 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[40]. 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[43]. 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.

1 Webber, H. J., Batchelor, L. D. & Reuther, W. *The Citrus Industry*. 1-39 (Univ. California Press, 1967).
2 Wang, L. *et al.* Somatic variations led to the selection of acidic and acidless orange cultivars. *Nat. Plants* (2021). https://doi.org:10.1038/s41477-021-00941-x
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 (2014). https://doi.org:10.1038/nbt.2906
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 (2022). https://doi.org:10.1101/2022.02.05.479263
5 Wu, G. A. *et al.* Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology* **32**, 656–662 (2014). https://doi.org:10.1038/nbt.2906
6 Wang, L. *et al.* Somatic variations led to the selection of acidic and acidless orange cultivars. *Nat. Plants* (2021). https://doi.org:10.1038/s41477-021-00941-x
7 Butelli, E. *et al.* Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* **24**, 1242–1255 (2012). https://doi.org:10.1105/tpc.111.095232
8 Shamel, A. D. & Pomeroy, C. S. Bud mutations in horticultural crops. *Journal of Heredity* **27**, 487–494 (1936). https://doi.org:10.1093/oxfordjournals.jhered.a104171
9 Raynes, Y., Wylie, C. S., Sniegowski, P. D. & Weinreich, D. M. Sign of selection on mutation rate modifiers depends on population size. *Proceedings of the National Academy of Sciences* **115**, 3422–3427 (2018). https://doi.org:10.1073/pnas.1715996115
10 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 (2018). https://doi.org:10.1038/s41438-018-0062-x
11 Wang, L. *et al.* Somatic variations led to the selection of acidic and acidless orange cultivars. *Nature Plants* **7**, 954–965 (2021). https://doi.org:10.1038/s41477-021-00941-x
12 Webber, H. J., Batchelor, L. D. & Reuther, W. *History and development of the citrus industry*. (Univ. California Press, 1967).
13 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. *The Plant Journal* **91**, 220-236 (2017). https://doi.org:10.1111/tpj.13558
14 Sun, W. *et al.* Biochemical and Molecular Characterization of a Flavonoid 3-O-glycosyltransferase Responsible for Anthocyanins and Flavonols Biosynthesis in Freesia hybrida. *Frontiers in Plant Science* **7** (2016).
15 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 Genetics & Genomes* **9**, 1257-1269 (2013). https://doi.org:10.1007/s11295-013-0635-7
16 Griffiths, J. *et al.* Genetic characterization and functional analysis of the GID1 gibberellin receptors in Arabidopsis. *The Plant cell* **18**, 3399–3414 (2006). https://doi.org:10.1105/tpc.106.047415

17 Gou, X. *et al.* Genetic evidence for an indispensable role of somatic embryogenesis receptor kinases in brassinosteroid signaling. *PLoS genetics* **8**, e1002452 (2012). https://doi.org:10.1371/journal.pgen.1002452
18 Aguilar-Martínez, J. A., Poza-Carrión, C. & Cubas, P. Arabidopsis BRANCHED1 acts as an integrator of branching signals within axillary buds. *The Plant cell* **19**, 458–472 (2007). https://doi.org:10.1105/tpc.106.048934
19 Zhang, F. *et al.* Reprogramming of Stem Cell Activity to Convert Thorns into Branches. *Current biology : CB* **30**, 2951-2961.e2955 (2020). https://doi.org:10.1016/j.cub.2020.05.068
20 Ohnishi, T. *et al.* C-23 hydroxylation by Arabidopsis CYP90C1 and CYP90D1 reveals a novel shortcut in brassinosteroid biosynthesis. *The Plant cell* **18**, 3275–3288 (2006). https://doi.org:10.1105/tpc.106.045443
21 Titapiwatanakun, B. *et al.* ABCB19/PGP19 stabilises PIN1 in membrane microdomains in Arabidopsis. *The Plant journal : for cell and molecular biology* **57**, 27–44 (2009). https://doi.org:10.1111/j.1365-313X.2008.03668.x
22 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 (2015). https://doi.org:10.1111/jipb.12332
23 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 (2016). https://doi.org:10.1080/15592324.2015.1010947
24 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 (2002). https://doi.org:10.1023/A:1015077508675
25 Bermejo, A. *et al.* Auxin and Gibberellin Interact in Citrus Fruit Set. *Journal of Plant Growth Regulation* **37**, 491–501 (2018). https://doi.org:10.1007/s00344-017-9748-9
26 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 (2015). https://doi.org:10.1111/tpj.12719
27 Gregory, T. R. Artificial Selection and Domestication: Modern Lessons from Darwin's Enduring Analogy. *Evolution: Education and Outreach* **2**, 5–27 (2009). https://doi.org:10.1007/s12052-008-0114-z
28 Seczynska, M., Bloor, S., Cuesta, S. M. & Lehner, P. J. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature* (2021). https://doi.org:10.1038/s41586-021-04228-1
29 Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature reviews. Genetics* **8**, 272–285 (2007). https://doi.org:10.1038/nrg2072
30 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). https://doi.org:10.1093/bioinformatics/bty191
31 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006). https://doi.org:10.1093/bioinformatics/btl158
32 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)* **23**, 2947–2948 (2007). https://doi.org:10.1093/bioinformatics/btm404
33 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular biology and evolution* **37**, 1530–1534 (2020). https://doi.org:10.1093/molbev/msaa015

34 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* **61**, 539–542 (2012). https://doi.org:10.1093/sysbio/sys029

35 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293-W296 (2021). https://doi.org:10.1093/nar/gkab301

36 Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Research* **43**, W39-49 (2015). https://doi.org:10.1093/nar/gkv416

37 Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276–277 (2000). https://doi.org:10.1016/s0168-9525(00)02024-2

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

39 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 Research* **41**, e121 (2013). https://doi.org:10.1093/nar/gkt263

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

41 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint*, arXiv:1303.3997v1302 (2013).

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

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

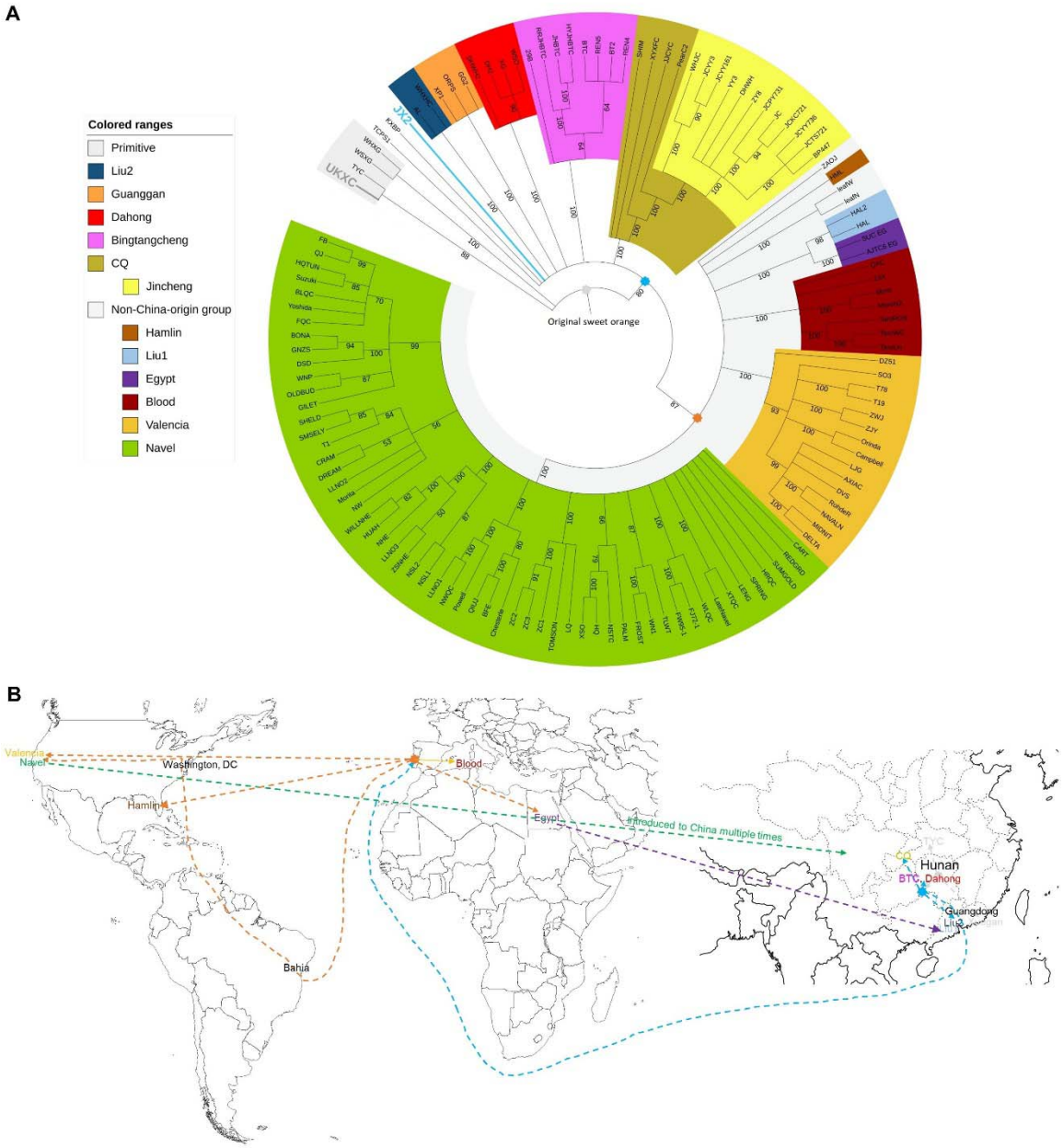

Figure 1. Sweet orange lineage development and spread history. A. Transposon insertion loci-based sweet orange breeding history. In the cladogram, the numeric values on the branches denote the corresponding posterior probabilities (%) from Bayesian inference, and branches with <50% posterior probabilities have been deleted. Cultivar groups have been shown with distinct background colors as denoted in the legend. The silver, light blue, and orange polygons represent the inferred ancestral lineages distributed in the three main spread events, respectively. B. The spread of sweet orange and rise of sweet orange cultivar groups. Southern China has been enlarged in the partial world map to better show the spread of sweet orange lineages inside China. The straight dashed lines point from the ancestral lineage location to the afterward cultivar origin location with unclear transmission route information. The curved dashed line depicts the approximate transmission route before the cultivar group origin according to the documents.

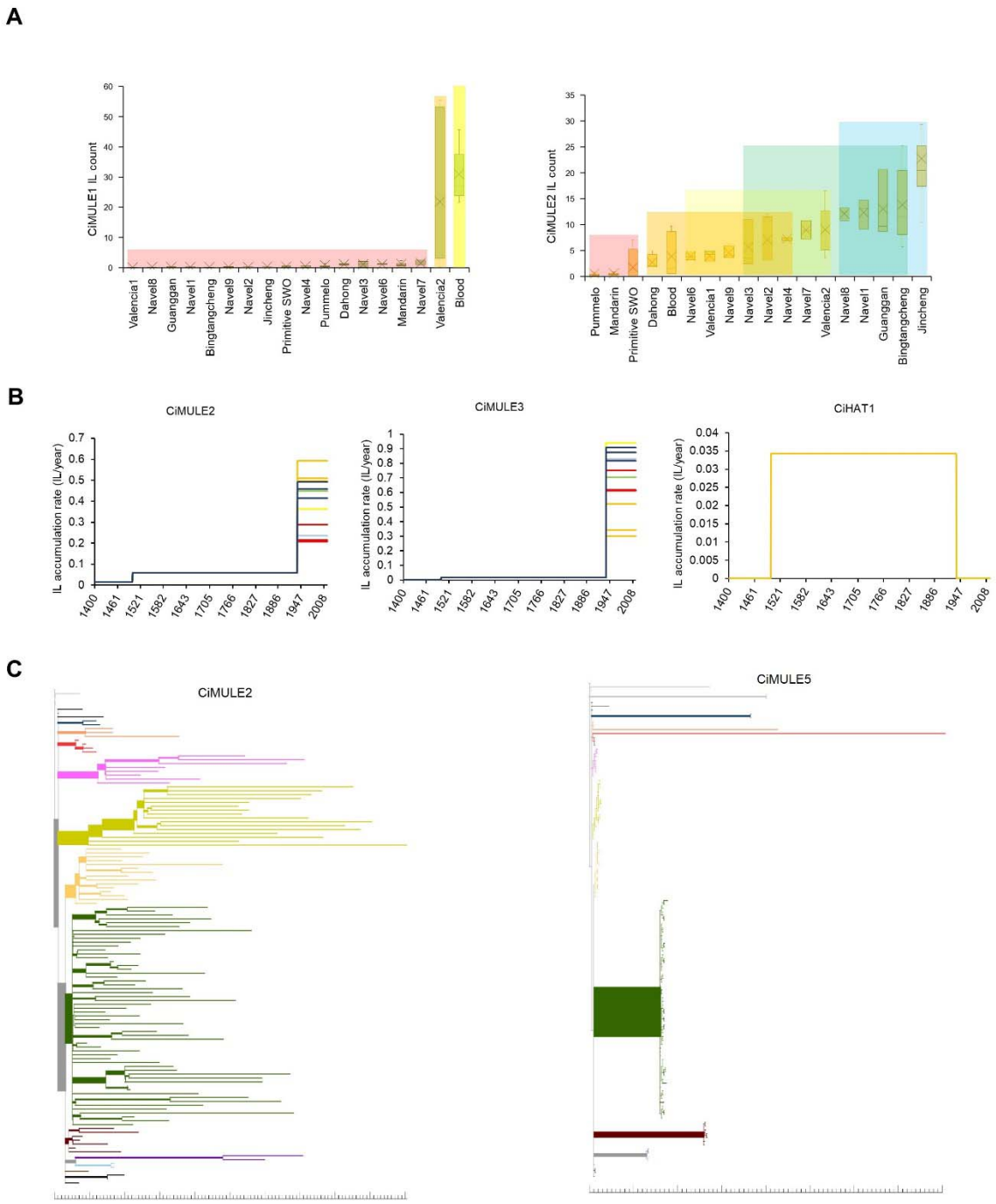

Figure 2. Heterogenous transposon insertion accumulation among sweet orange lineages. A. Distribution of species-specific insertion locus (IL) counts of CiMULE1 and CiMULE2 in mandarin, pummelo, and distinct sweet orange phylogenetic clades. Different background colors denote statistically (by one-way ANOVA) different homogenous subsets. B. IL accumulation rates of the Jincheng sub-clades and their ancestor. Distinct phylogenetic clades have been depicted using lines with different colors. The TE families have been shown on the top of the graphs. C. IL accumulation in the phylogenetic branches. The tree topology from Figure 1A is used, and the branch lengths (each minor scale represents 1 novel IL in the scale bars) are in

proportion to the accumulated novel IL counts on the phylogenetic path. The line colors are used to show the cultivar group information as the legend in Figure 1A.

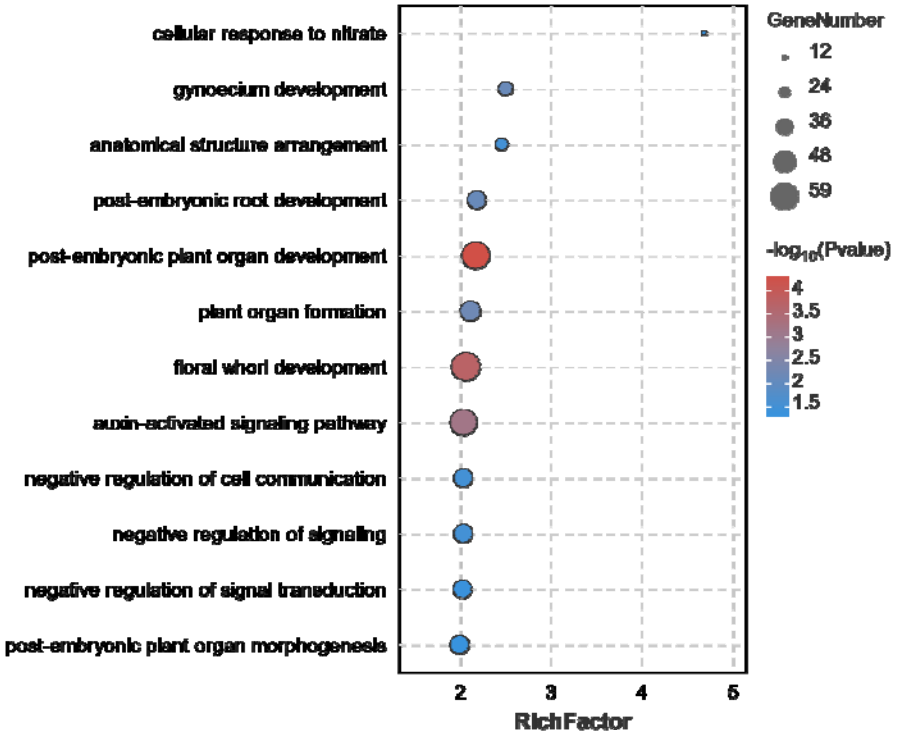

Figure 3. Gene ontology enrichment of transposon affected genes.