nature plants

ARTICLES

https://doi.org/10.1038/s41477-021-00941-x

Check for updates

Somatic variations led to the selection of acidic and acidless orange cultivars

Lun Wang[1,2,8], Yue Huang[1,2,8], ZiAng Liu[1,2,8], Jiaxian He[1,2], Xiaolin Jiang[1], Fa He[1], Zhihao Lu[1,2], Shuizhi Yang[3], Peng Chen[3], Huiwen Yu[1], Bin Zeng[3], Lingjun Ke[1], Zongzhou Xie[1], Robert M. Larkin[1], Dong Jiang[4], Ray Ming[5], Edward S. Buckler[6,7], Xiuxin Deng[1,2] and Qiang Xu[1,2]✉

Somatic variations are a major source of genetic diversification in asexual plants, and underpin clonal evolution and the breeding of asexual crops. Sweet orange is a model species for studying somatic variation because it reproduces asexually through apomixis and is propagated asexually through grafting. To dissect the genomic basis of somatic variation, we de novo assembled a reference genome of sweet orange with an average of three gaps per chromosome and a N50 contig of 24.2 Mb, as well as six diploid genomes of somatic mutants of sweet oranges. We then sequenced 114 somatic mutants with an average genome coverage of 41×. Categorization of the somatic variations yielded insights into the single-nucleotide somatic mutations, structural variations and transposable element (TE) transpositions. We detected 877 TE insertions, and found TE insertions in the transporter or its regulatory genes associated with variation in fruit acidity. Comparative genomic analysis of sweet oranges from three diversity centres supported a dispersal from South China to the Mediterranean region and to the Americas. This study provides a global view on the somatic variations, the diversification and dispersal history of sweet orange and a set of candidate genes that will be useful for improving fruit taste and flavour.

Somatic variations are a phenomenon shared by most perennials and asexual crops and are a major source of genetic diversification, a driving force of clonal evolution and useful for breeding[1–3]. Somatic mutants occur at high frequencies in perennials such as fruit tree crops. Indeed, 80 years ago, 1,337 and 863 somatic mutants were recorded in sweet orange and apple, respectively[4]. Somatic mutations in citrus led to a broad spectrum of phenotypes, including changes in fruit shape, colour, acidity, maturation season, developmental changes related to sterility, flowering time and tree architecture[5].

Somatic variations in asexual perennial plants are not as well studied as those associated with human cancer, which range from single nucleotide polymorphisms (SNPs) to complex genome structural variations (SVs) originating from chromosomal rearrangements[6–8]. In asexual perennials, a fraction of the somatic variations is associated with transposable elements (TEs). For example, transposon-induced somatic variations are associated with parthenocarpy in apple[9], activation of anthocyanin biosynthesis[10,11] and abnormal development of the inflorescence in grape[12]. Additionally, genome analyses of grape linked the signature of a large chromosome-replacement pattern derived from a chromoanagenesis-like event to a berry colour phenotype[13]. Whole-genome sequencing of a long-lived oak tree, a large mushroom and a seagrass yielded 17, 111 and 1,216 (pairwise compared) somatic SNP mutations, respectively, in different parts within the individual[14–17]. These data provide evidence for a low frequency of spontaneous somatic mutations in long-lived plants and clonally related subpopulations of agarics and marine plants. However, the genomic basis of comprehensive somatic variation, especially for large transposon activation in fruit crops, remains largely unclear. Rapidly developing genomic approaches provide an opportunity to investigate somatic variations at the whole-genome level[18–20].

Citrus is grown in more than 114 countries[21]. The best-known citrus species are sweet orange (*Citrus sinensis*), mandarin (*Citrus reticulata*), pummelo (*Citrus grandis*), grapefruit (*Citrus paradise*) and lemon (*Citrus limon*). Although the domestication of sweet orange is thought to have occurred in China, the evidence supporting this hypothesis is ambiguous. Orange was first mentioned in the Chinese prose poem "Shanglin Fu" by Ssu-ma Hsiang-ju approximately 179 to 117 BC[22,23]. Sweet orange was grown for many centuries in China and had apparently reached an advanced stage of cultivation before it became well known to Europeans[24]. In fact, during the Han Dynasty (202 BC to 220 AD), a special government official managed the affairs of the citrus industry and collected tributes to the emperor.

Citric acid is a marker metabolite in the citric acid cycle, which is central to energy metabolism in all organisms that perform aerobic respiration. Citric acid accounts for nearly 90% of the total organic acids that accumulate in most citrus fruits[25]. Citric acid is a major factor contributing to the flavour of citrus fruits and thus affects fruit beverage quality. High-acid citrus fruits are often important ingredients in dishes and juice. Owing to somatic variations, distinct genotypes of sweet orange that produce high-acid, low-acid and acidless fruits have arisen throughout its dispersal history.

Results

De novo assembly of a reference genome and six diploid genomes for sweet orange. A combination of long-range sequencing via

[1]Key Laboratory of Horticultural Plant Biology (Ministry of Education), Huazhong Agricultural University, Wuhan, P. R. China. [2]Hubei Hongshan Laboratory, Wuhan, China. [3]Horticulture Institute, Hunan Academy of Agricultural Sciences, Changsha, P. R. China. [4]Citrus Research Institute, Southwest University, Chongqing, P. R. China. [5]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [6]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY, USA. [7]Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA. [8]These authors contributed equally: Lun Wang, Yue Huang, ZiAng Liu. ✉e-mail: xuqiang@mail.hzau.edu.cn

Table 1 | Statistics for the genome assembly of double-haploid sweet orange

	Ultralong reads and Hi-C sequences (average = 30.1 kb) v.4.0	Long-read sequencing (average = 8.5 kb) v.3.0	Illumina short-read sequencing (300–500 bp) v.2.0
Size of the assembled scaffold (bp)	336,647,285	338,392,684	320,520,682
Largest contig (bp)	35,869,611	7,530,179	323,337
Number of contigs	165	1,002	16,890
Chromosome size	304,216,219	270,105,696	238,997,219
Number of gaps on the chromosome	27	316	8,317
N50 contig (bp)	24,160,866	1,803,079	49,898[a]
N90 contig (bp)	3,932,138	122,160	11,219
Largest scaffold (bp)	50,232,142	13,334,880	8,158,231
Scaffold N50 (bp)	32,314,164	3,107,487	1,687,863
Scaffold N90 (bp)	26,160,000	320,628	105,173
Number of gene models	29,875/49,567	29,301/45,016	29,445/44,387
Mean transcript length (bp)	2,008	2,161	1,817
Mean coding sequence length (bp)	1,200	1,307	1,255
Percentage of TEs	41.12	48.10	20.49

[a]N50 values of the genome assembly were calculated using sequences longer than 300 bp.

the nanopore ultralong sequencing platform and PacBio long-read sequencing platform with error correction using Illumina short reads data was applied to de novo assemble a reference genome from a double-haploid sweet orange line. PacBio single-molecule long reads with 73.5× coverage were generated. Additionally, 214× Illumina sequencing reads of various insert sizes from published datasets[20] were used to construct scaffolds, fill gaps and correct assembly errors. Our assembly filled 6,117 out of 12,578 gaps in the previous genome, occupying 9,584,715 bp of total length. In a further step, a total of 108.2× coverage of ultralong reads (average 30.1-kb read length) was generated on the nanopore platform to improve chromosome assembly contiguity. In addition, Hi-C data with 117.5× genome coverage were used to extend the chromosome. Finally, a reference genome of sweet orange was assembled, yielding an average of three gaps per chromosome. The N50 contig was 24.16 Mb (Table 1). We identified a total of 20.63% more repetitive elements in this intact genome relative to the draft genome of sweet orange[20] because of the advantages provided by the ultralong read sequencing strategy (Supplementary Table 1). Based on genome annotation, we predicted 29,875 protein-coding genes and 49,567 transcripts. The gene transcripts had an average length of 2,008 bp, an average coding sequence length of 1,200 bp and an average exon length of 352 bp.

Assembling and phasing of a heterozygous diploid genome of 'Valencia' sweet orange was performed by 10X Genomics linked reads with 179.3× coverage, PacBio reads with 32.8× coverage and Illumina reads with 58.5× coverage. Compared with three basic *Citrus* species (mandarin, pummelo and citron), commercially cultivated sweet oranges are highly heterozygous, with similar levels of heterozygosity to other known hybrids such as lemon and grapefruit (Supplementary Fig. 1). By mapping short and single-molecule long reads to the reference genome, a total of 3,909,563 SNPs and 12,520 SVs were identified, respectively (Supplementary Fig. 2). The SVs consisted of 7,253 deletions and 5,267 insertions. The average length of insertions and deletions (InDels) was approximately 900 bp, with a range from 30 bp to 10 kb. The SNPs and PacBio mapping results were utilized to perform haplotype phasing. As a result, 325 haplotype blocks were obtained, and the haplotype blocks spanned 298,037,005 bp of the genome with a N50 of 3.7 Mb.

To investigate large somatic variations, we assembled five more diploid genomes of sweet oranges. In total, PacBio or nanopore long reads of 29.2–117.1× genome coverage were generated for five somatic mutants (Supplementary Table 2). More than 90% of the contigs were anchored and oriented to chromosomes based on the di-haploid genome. The completeness of all five assembled genomes exceeded 90% when evaluated using BUSCO.

Diversity in fruit acidity among somatic mutants and a prototype of sweet orange. Somatic mutants are the most important resource for citrus breeding. Among the 215 sweet orange varieties that are cultivated worldwide, 80% of the varieties with clear origins arose from somatic mutants (Supplementary Table 3). On a broader scale, among the 736 citrus varieties, 60% with clear origins arose from somatic mutants. In citrus, somatic mutants originate from mutations in buds. The bud mutants were then selected by humans and vegetatively propagated through grafting (Extended Data Fig. 1). Frequent somatic mutants have been reported in three diversity centres of sweet orange: South China, the Mediterranean and the Americas. For example, somatic mutants from the Mediterranean region produce acidless orange (that is, 'Succari' sweet orange), blood orange and late-maturing orange (that is, Valencia sweet orange), and those from China include low-acid varieties (that is, 'BingTangCheng', designated BTC hereafter). The genome sequences of these somatic mutants are almost identical and markedly distinct from those of the sexual hybrids of sweet oranges (Supplementary Fig. 3).

We found native populations of sweet oranges in a region surrounded by the Nanling Mountains, Yuan River and Wuling Mountains (hereafter referred to as the NYW region) in South China (Fig. 1a). This region has a long history of sweet orange cultivation and has produced abundant indigenous cultivars with a broad spectrum of phenotypic variation (Supplementary Fig. 4). These sweet oranges vary in their fruit acidity, flesh and peel colours, maturation season, seedlessness, navel fruit shape, leaf shape and tolerance to micronutrient deficiencies[26] (Fig. 1b). In particular, there is wide variation in the degree of acidity, ranging from no acid (0.52 ± 0.04 mg ml^{-1}) to extremely high levels of acidity (21.97 ± 3.07 mg ml^{-1}) (Fig. 1c). Indeed, the levels of acidity in high-acid sweet oranges are similar to the levels in wild mandarins and 46-fold higher than in acidless sweet orange (Fig. 1c). The extremely high-acid sweet oranges from the NYW region are regarded as proto-sweet oranges because their fruits are relatively

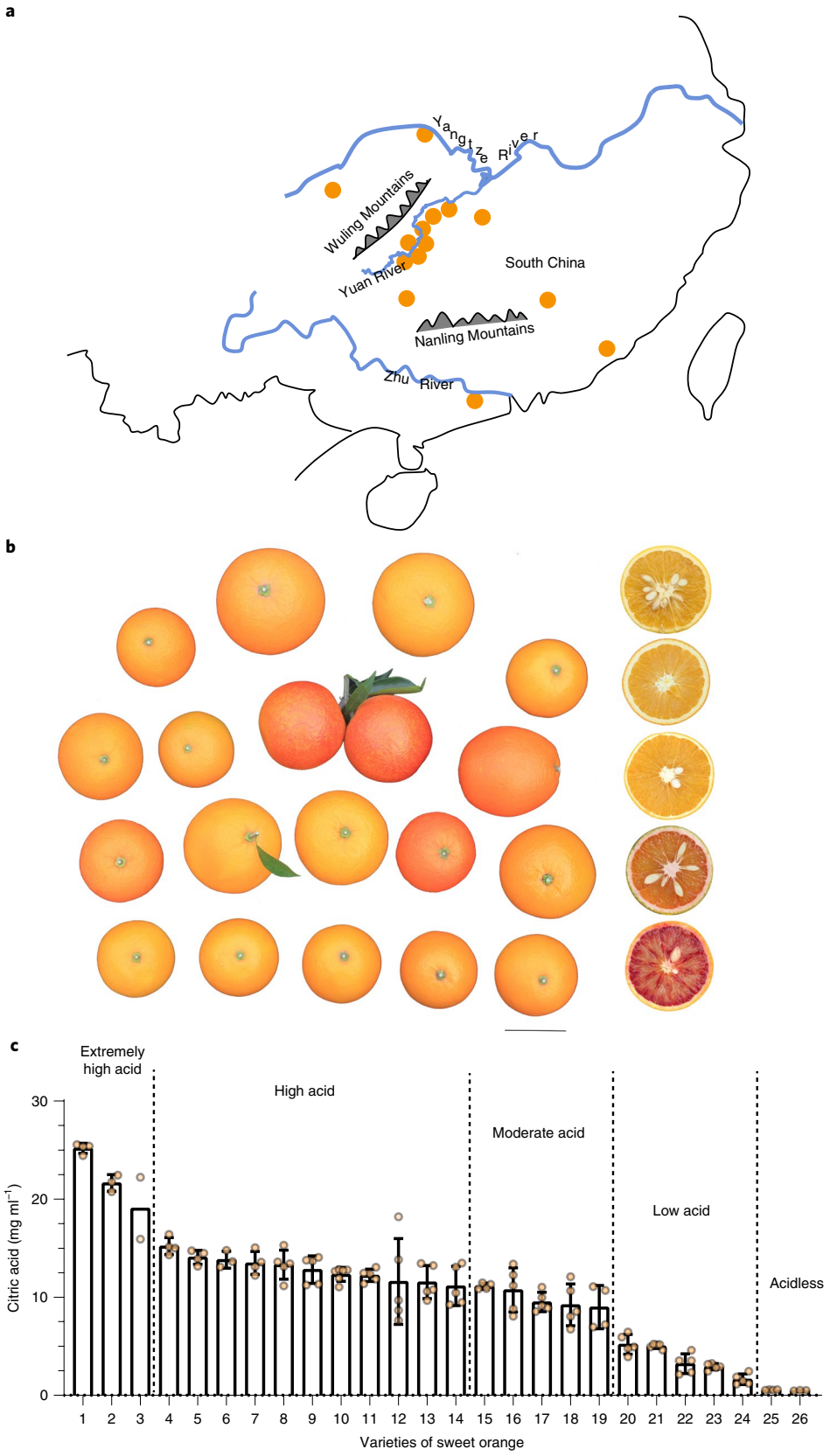


Fig. 1 | Distribution of sweet oranges sampled in South China and phenotypes of sweet orange somatic mutants. a, Geographical distribution of sweet oranges (orange circles) in South China. The NYW region refers to the region containing the Nanling Mountains, the Yuan River and the Wuling Mountains. Extremely high-acid sweet oranges, a prototype of sweet oranges, were collected around the Yuan River. **b**, Phenotypic variation of different sweet orange mutants. Scale bar, 5 cm. **c**, Citric acid content of fruits from somatic mutants of sweet orange. Values are shown as the mean ± s.e.m. Biological independent samples of each varieties were two to six. Statistical data of the biological independent samples are provided in the source data.

small and show nascent status compared with the other 114 varieties of sweet orange. In citrus, high acidity is a marker trait of undomesticated or wild germplasm[27].

Somatic variations in the sweet orange genome. Sweet oranges are monophyletic with genetic diversification arising from spontaneous somatic mutations because they reproduce asexually via apomixis[28] and are clonally propagated by grafting. To evaluate the somatic variations among sweet oranges at the whole-genome level, we sequenced 114 somatic mutants of sweet orange (37 from South China and 77 genotypes that are cultivated worldwide) with an average genome coverage depth of 41× (Supplementary Table 4). Principal component analysis (PCA) and genome structure and phylogenetic analyses of these sweet oranges and other citrus showed that the sweet oranges were nearly identical (Fig. 2), sharing 99.99% sequence identity based on the estimation of SNPs in consensus regions of the six genomes of sweet orange cultivars. These data confirm their genetic background and indicate that they are somatic mutants.

We identified 8,628 somatic SNPs in the 114 sweet orange mutants (Supplementary Fig. 5 and Fig. 2). We independently verified 21 out of the 25 SNPs (that is, validation rate of 84%) between high-acid and low-acid cultivars (DH2 and BTC) by Sanger sequencing of the pertinent target sites (Supplementary Table 5) and 18 out of the 23 somatic SNPs (validation rate of 78%) between a high-acid orange (DH2) and two acidless oranges (HAL and Succari) by sequencing the gene-coding regions (Supplementary Table 6). Among these 8,628 somatic SNPs, 6,614 (accounting for 76.66%) are heterozygous mutations (that is, conversion from the homozygous to the heterozygous states). The homozygous mutations (accounting for 23.34%) were a lower proportion of somatic mutations than heterozygous mutations (Fig. 2). The somatic mutation rate in the sweet orange population comprising 114 accessions was estimated to be 1.12×10^{-7} per base per accession after normalizing to the genome size. A total of 751 of these SNPs potentially affect gene-coding properties. Indeed, 698 of these SNPs were distributed in the coding regions of 669 genes, including 215 synonymous SNPs and 472 nonsynonymous SNPs. The other 53 SNPs were predicted to alter splice sites, start codons or stop codons (Supplementary Table 7).

A total of 2,818 somatic InDels were identified in the 114 sweet orange mutants. Most of the InDels were located in intergenic regions, and 36 InDels were predicted to cause large effects, including codon changes and frame shifts (Supplementary Table 7). Small InDels (length of <5 bp) accounted for 62% of the total InDels (Supplementary Fig. 6).

A total of 2,321 somatic SVs, including 104 large deletions, 31 duplications and 2,186 large insertions, were detected in the 114 sweet orange mutants (Fig. 3). Independent PCR amplification experiments validated 12 out of the 14 (validation rate of 85.71%) large insertions shared by seven low-acid mutants (BTC group) (Supplementary Tables 8 and 9, Extended Data Fig. 2 and Supplementary Fig. 7). The PCR-based validation rate of the deletions between different mutants was 100% (Supplementary Table 10 and Supplementary Figs. 8–10). Long-read sequencing data were also used to evaluate the accuracy of SVs predicted by the short reads. The average validation rate was 87.67% (Supplementary Table 8). We analysed the proportion of deleted fragment components and found that deletions more frequently originated from the mandarin haplotype in the blood orange, Valencia and Jincheng samples. Deletion in the pummelo haplotype was more frequent in the navel orange and low-acid group than in the other sweet orange groups (Supplementary Fig. 11). We also observed 33 genomic regions with extra-large deletions and duplications longer than 500 kb (Extended Data Fig. 3 and Supplementary Figs. 12–23). In addition, another 42 extra-large somatic variations (including 4 deletions, 17 insertions and 21 inversions over 500 kb) were identified by comparing the genome assembly of five mutants (TCPS1, BT2, UKXC, NW and NHE) to the reference genome (Supplementary Table 11).

A cluster analysis divided all of the sweet orange somatic mutants into seven groups (Fig. 2c). One group of high-acid and extremely high-acid types originated in South China. Four groups of moderate-acid types, including the Jincheng group, originated in South China, the blood orange and Valencia groups probably originated in the Mediterranean area, and the navel orange originated in the Americas. One low-acid group (BTC) originated in South China. One acidless group had a mixed origin. In addition, we constructed a phylogenetic network of representative individuals, which revealed reticulate relationships between these groups (Supplementary Fig. 24).

Somatic mutations in sweet orange. To estimate the somatic mutation rate in sweet orange, three 30-year-old trees from Yichang and Chongqing in South China were used to identify somatic mutations in their different branches. We sequenced four or five branches from each of the trees with a 35× genomic depth for each sample. Application of the somatic SNP calling pipeline resulted in 29, 32 and 25 somatic SNPs for the three trees (Supplementary Table 12). Based on these data, the somatic mutation rate was estimated to be 3.03×10^{-10} per base per year.

Deleterious mutations accumulate during asexual propagation[29,30]. Mutations that change amino acid sequences were predicted using the sorting intolerant from tolerant (SIFT) program, which classifies mutations as tolerant (SIFT score > 0.05) or deleterious (SIFT score ≤ 0.05). Nearly 5.47% of the 8,628 somatic SNPs in the 114 sweet oranges were nonsynonymous substitutions, and 44.92% (212) of these nonsynonymous substitutions were putative deleterious substitutions (SIFT ≤ 0.05). We also combined the criteria of SIFT (≤ 0.05) and PhastCons scores (>0.15) based on genomic evolution (Supplementary Fig. 25) and obtained a more conservative set of 130 deleterious mutations (Supplementary Table 7). We analysed the group-specific deleterious somatic mutations and found that the deleterious mutations consisted of 1–6% SNPs in the seven groups (Supplementary Fig. 26). Interestingly, the acidless group had more deleterious mutations than the other sweet orange groups. Blood orange, a moderate-acid sweet orange group, had the fewest deleterious mutations among the sweet orange groups.

Somatic TE polymorphisms. By using the high-depth sequencing data with an average of 41× genome coverage for each of the 114 somatic mutants, we developed a population-based strategy (Methods and Supplementary Fig. 27) to identify large fragment insertions (LFIs) along the genome, which resulted in a validation rate of more than 85% for both the long-read sequencing data and the independent PCR experiments, as mentioned above.

We realigned the mate pairs of discordant reads to the repeat database for the reference genome to identify large insertion points. Approximately 40.1% of these insertions (877) were predicted to encompass a specific type of TE. Among the transposons, 52.3% were long terminal repeat retrotransposons (LTR), with Gypsy, Copia and other minor types contributing 26.7%, 22.9% and 2.7%, respectively (Fig. 4a). Remarkably, 30.5% of the transposons were Mutator-like element (MuLE/MuDR)-type DNA transposons. This ratio was higher than expected for the entire genome of sweet orange ($P \leq 1.27 \times 10^{-07}$, Fisher's exact test). A total of 23.7% of the TEs were inserted in genic regions, and 60.3% were inserted in intronic regions (Supplementary Table 13).

The frequency of TE transposition differed among the seven groups shown in Fig. 4b. The transposition frequency was the highest in the Jincheng group, followed by the blood orange and Valencia orange groups (Fig. 4b). By comparing somatic MuLE/MuDR, LINE, Copia and Gypsy transpositions in all the somatic TE insertions in the seven sweet orange groups, we found that MuLE/

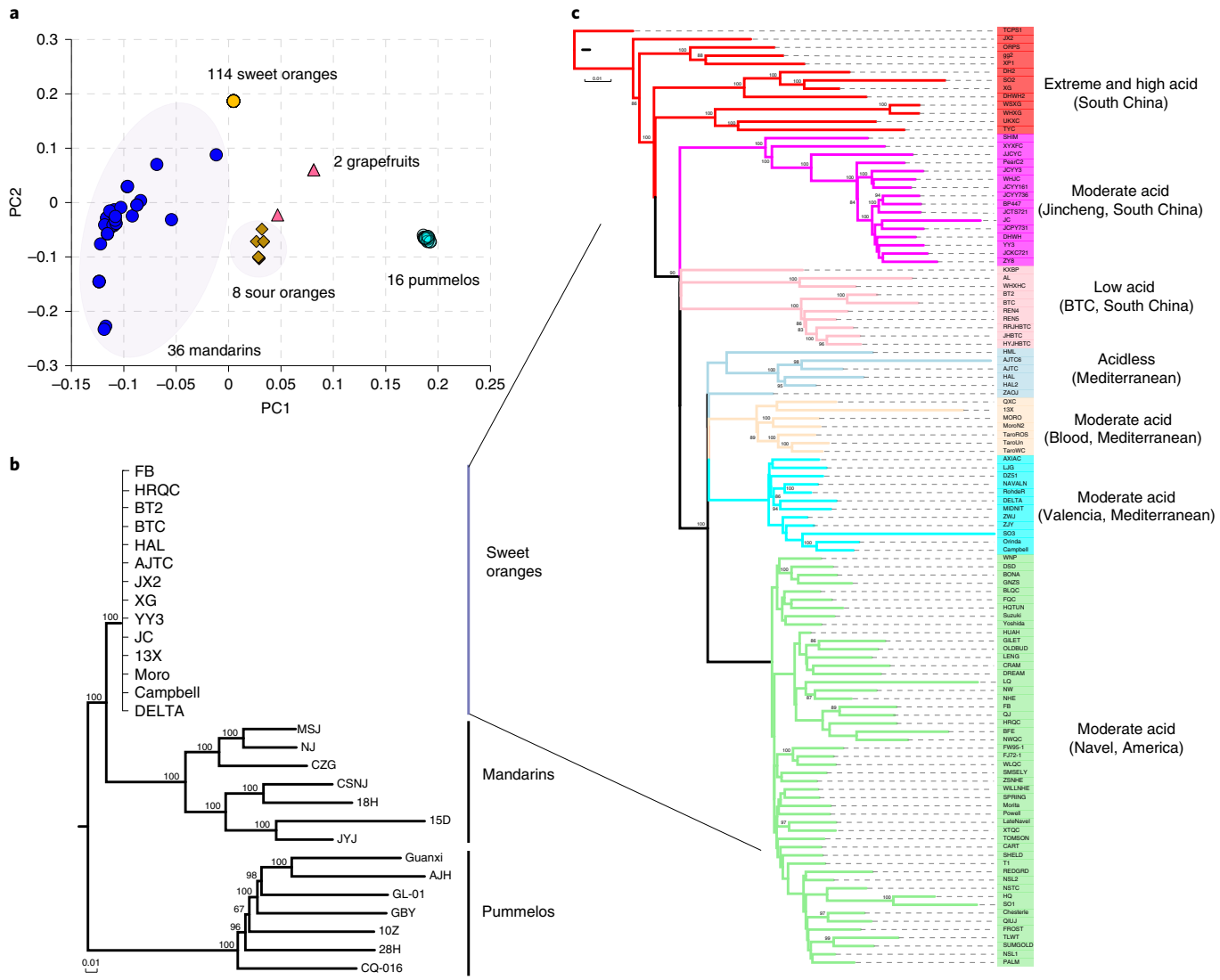


Fig. 2 | PCA and phylogenetic relationships of sweet oranges and other sexual/asexual citrus species. a, PCA of 114 sweet oranges, 36 mandarins, 8 sour oranges, 2 grapefruits and 16 pummelos. PCA of sweet oranges, mandarins, pummelos and citrus hybrids indicated that 114 sweet oranges showed nearly identical genetic backgrounds in contrast to the diversified mandarins, sour oranges and pummelos. **b**, Phylogenetic tree of 14 representative sweet oranges, 7 pummelos and 7 mandarins based on genome-wide SNP data. The tree confirmed that the sweet oranges had identical backgrounds, while the mandarins and pummelos were highly diversified. **c**, Phylogenetic relationships among 114 sweet oranges based on all 8,628 somatic SNPs. All of the sweet orange accession names are listed in Supplementary Table 4. The trees were produced using the ML method implemented with the raxmlHPC software. Bootstrap values over 60 are indicated at the tree nodes.

MUDR transpositions were the most active in the Jincheng and acidless groups and least active in the high-acid sweet oranges. LTR retrotransposons accounted for a large number of the somatic transpositions in the other five groups. A neighbour-joining tree of the TE data was consistent with the cluster tree based on somatic SNPs (Supplementary Fig. 28).

Diversification of sweet orange and somatic variations associated with fruit acidity. Based on sequence variations and levels of fruit acidity, we classified the sweet oranges into four types: extremely high and high-acid type; moderate-acid type (Valencia, navel, Jincheng and blood orange groups); low-acid type (BTC); and acidless type (HAL and Succari). We found a total of 1,248 group-specific SNPs, 31 allelic deletions or duplications and 154 transpositions, which were predominant in the individuals within one group and distinguished from other groups (Supplementary Table 14). To identify somatic variations associated with fruit acidity, we compared large somatic variations and SNPs from the moderate-acid oranges (blood orange as the representative), low-acid oranges (BTC as the representative) and acidless oranges (HAL and Succari) to the high-acid sweet oranges. We detected 123 somatic variations specific to moderate-acid oranges, 239 somatic variations specific to low-acid oranges and 106 somatic variations specific to acidless oranges (Supplementary Table 15).

A large insertion found in the intron of the *CsRAE1* gene in all of the blood oranges indicates that this mutant allele is associated with fruit acidity (Fig. 5a,b and Extended Data Fig. 4). *RAE1* encodes a F-box protein that modulates a zinc-finger transcription factor that regulates the transcription of genes that encode ALMT transporters[31,32]. Based on the sequence, the insertion was predicted to be a Pack-MuLE transposon. We used independent PCR experiments to validate the insertion of this TE (Supplementary

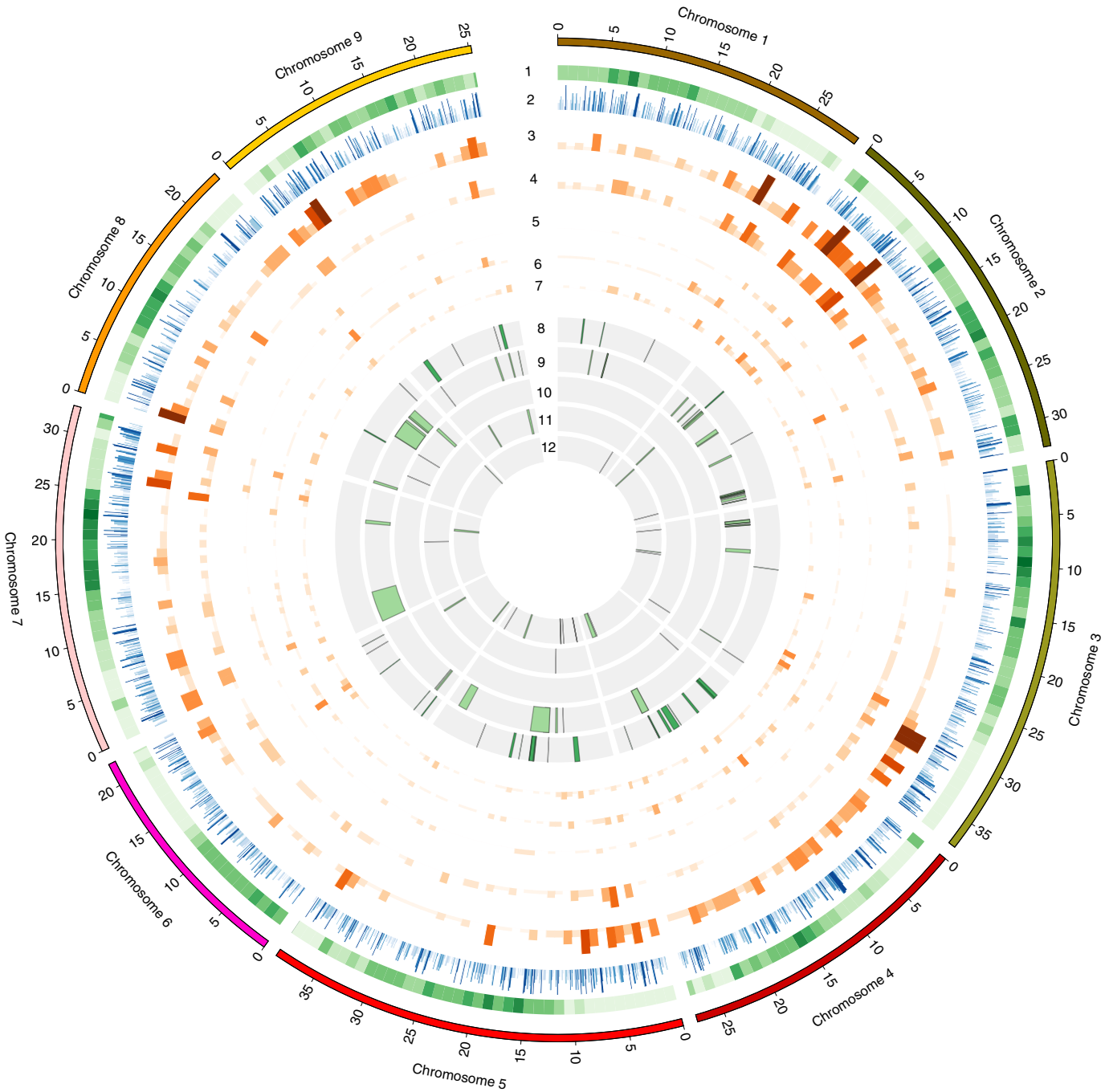


Fig. 3 | Genomic location of somatic variations in 114 sweet oranges. (1), TE density per 1 Mb in the reference genome of sweet orange. (2), Total somatic SNV density (per 100 kb). (3–7), Somatic transposition density (per 1 Mb) in the Jincheng, navel, BTC, Valencia and blood orange groups. (8–12), Deleted and duplicated regions in the Jincheng, navel, BTC, Valencia and blood orange groups. The genome coordinates are from v.3.0 of di-haploid sweet orange.

Table 9). The expression of the *RAE1* gene was upregulated in blood orange (moderate-acid type) but not in 'Dahong' (high-acid type) (Extended Data Fig. 4b). Quantification of gene expression in eight stages of development in the navel orange 'Newhall' and in a late-maturing navel orange indicated that the *RAE1* gene is overall upregulated and the peak of expression occurs at 240 days after flowering (DAF) and 270 DAF when fruit acidity is decreasing (Extended Data Fig. 4d,e). Furthermore, *RAE1* was transiently overexpressed in kumquat (*Fortunella margarita*). In the fruits that expressed *RAE1* at an average of 19.5-fold higher levels relative to the control (Extended Data Fig. 4f), pH values were significantly increased and citric acid levels were significantly decreased relative to the control (Extended Data Fig. 4g,h).

Based on our finding that a 6.9-kb transposon was inserted upstream of a Na^+/H^+ transporter[33,34] (*CsNHX*) in all of the low-acid mutants (BTC), we speculated that this insertion allele is associated with low acidity (Fig. 5a,b and Extended Data Fig. 5). Knocking out the *NHX* gene in *Arabidopsis* leads to changes in vacuolar pH and intracellular ion homeostasis[35]. Moreover, a member of this gene family was reported to be associated with proton gradient in maize roots[36]. We used PCR-based experiments to independently verify the insertion of the transposon. We cloned the full length of the

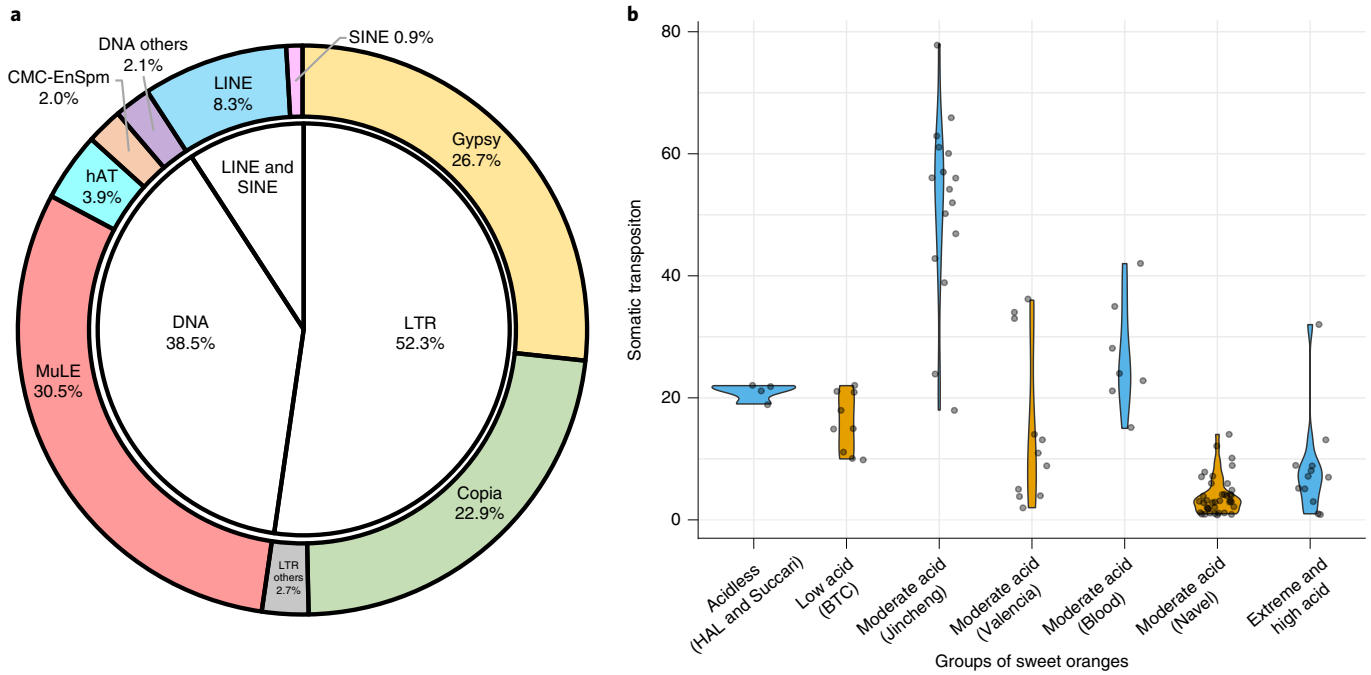


Fig. 4 | Categorization of somatic TE transposition events in sweet oranges. a, Types and percentages of different transpositions detected in the 114 sweet orange somatic mutants. LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements. **b**, Frequencies of somatic transposition in seven sweet orange groups (4 individuals in the acidless group, 7 in the low-acid group, 16 in the Jincheng group, 11 in the Valencia group, 7 in the blood orange group, 47 in the navel orange group and 12 in the extremely high-acid and high-acid group). Ten other accessions without signals of TE insertions are not shown.

transposon sequence and predicted that it is a Pack-MuLE transposon (Extended Data Fig. 5a,b and Supplementary Table 9). The sequence of the transposon is similar to that of a transposon found in the genome of a clementine mandarin mutant[37], although we found a few SNPs and deletions. The expression of the *CsNHX* gene was upregulated in the low-acid mutants (Extended Data Fig. 5c) and upregulated during the later stages of fruit development, with a peak of expression at 210 DAF and 240 DAF, when fruit acidity is decreasing (Extended Data Fig. 5d,e). Furthermore, data from the transient gene expression assays in kumquat (*F. margarita*) indicated that the *CsNHX* gene showed an average of 15.4-fold increase in expression (Extended Data Fig. 5f), which led to a significant increase of pH values and significant decrease in citric acid levels in overexpressed fruits relative to the control (Extended Data Fig. 5g,h).

A gene associated with acidlessness was also revealed by retrotransposon-mediated insertions in the *AN1* gene, which encodes a bHLH transcription factor (Fig. 5 and Supplementary Fig. 29), and in a gene encoding a BTB-domain-containing protein (Supplementary Fig. 30). In addition, we found a large allelic deletion specific to the acidless oranges (Supplementary Fig. 31). We sequenced the two allele-specific retrotransposons, with insertion sites spaced 108-bp apart, in the bHLH gene of the acidless cultivars (HAL and Succari) (Supplementary Table 9). One retrotransposon was highly similar to the *Tcs4* sequence[38]. Indeed, we found only six SNP differences between the two sequences. Another retrotransposon in the other allele was similar to *Tcs6x* in 'Vaniglia'[38,39]. We found only four SNPs and four deletions between the two sequences. Based on these data, we suggest that the acidless oranges (HAL collected in South China), Succari (cultivated in the Middle East) and Vaniglia (an Italian variety)[38,39] may share a common origin. In addition, we found a large deletion specific to Succari, which indicates that this variety arose after a mutant responsible for acidless appeared (Supplementary Fig. 32). The two retrotransposons clustered with the *Bianca* Copia superfamily of retrotransposons from sweet orange.

Based on the variation map, the genetic data also support the dispersal history of sweet orange (Fig. 5c). The prototype of sweet orange was developed in South China. The high acidity of the fruits was a pivotal advantage in long-distance dispersal, including to the Mediterranean area through the sea routes of India and Persia and to the Americas on Columbus's voyage. Many rounds of introduction may have occurred during the dispersal history of sweet orange. The data also suggest that blood oranges from Europe and China resulted from independent somatic variations, which was further supported by a validated deletion on chromosome 1 (Supplementary Fig. 10) and different retrotransposons in the promoter of the *Ruby1* gene on chromosome 5 (ref.[10]).

Discussion

We sequenced 114 sweet orange varieties that are cultivated worldwide, including a set of proto-sweet oranges with extremely high acidity from South China. We de novo assembled a high-quality *Citrus* reference genome with an average of three gaps per chromosome and a N50 contig of 24.2 Mb. By using both the double-haploid and heterozygous diploid, we constructed both haplotype genomes of sweet orange.

Using a population containing somatic mutants, we generated a map showing the global distribution of somatic variation. Single-nucleotide mutations, SVs and TEs were investigated in 114 sweet orange accessions. The results have broad implications for perennials because somatic mutants are frequently observed in such plants and used for breeding purposes. Approximately 80% of sweet orange and 60% of citrus varieties worldwide are derived from somatic mutants. In the past 40 years of breeding in China, 78% of citrus varieties, 80% of banana varieties, 32% of apple varieties, 22% of pear varieties and 18% of grape varieties were obtained from somatic mutants[40]. These data are important in that they add to our existing knowledge of somatic mutations associated with particular loci or phenotypes/cases, such as those in grape[11,41,42], peach[43], oak

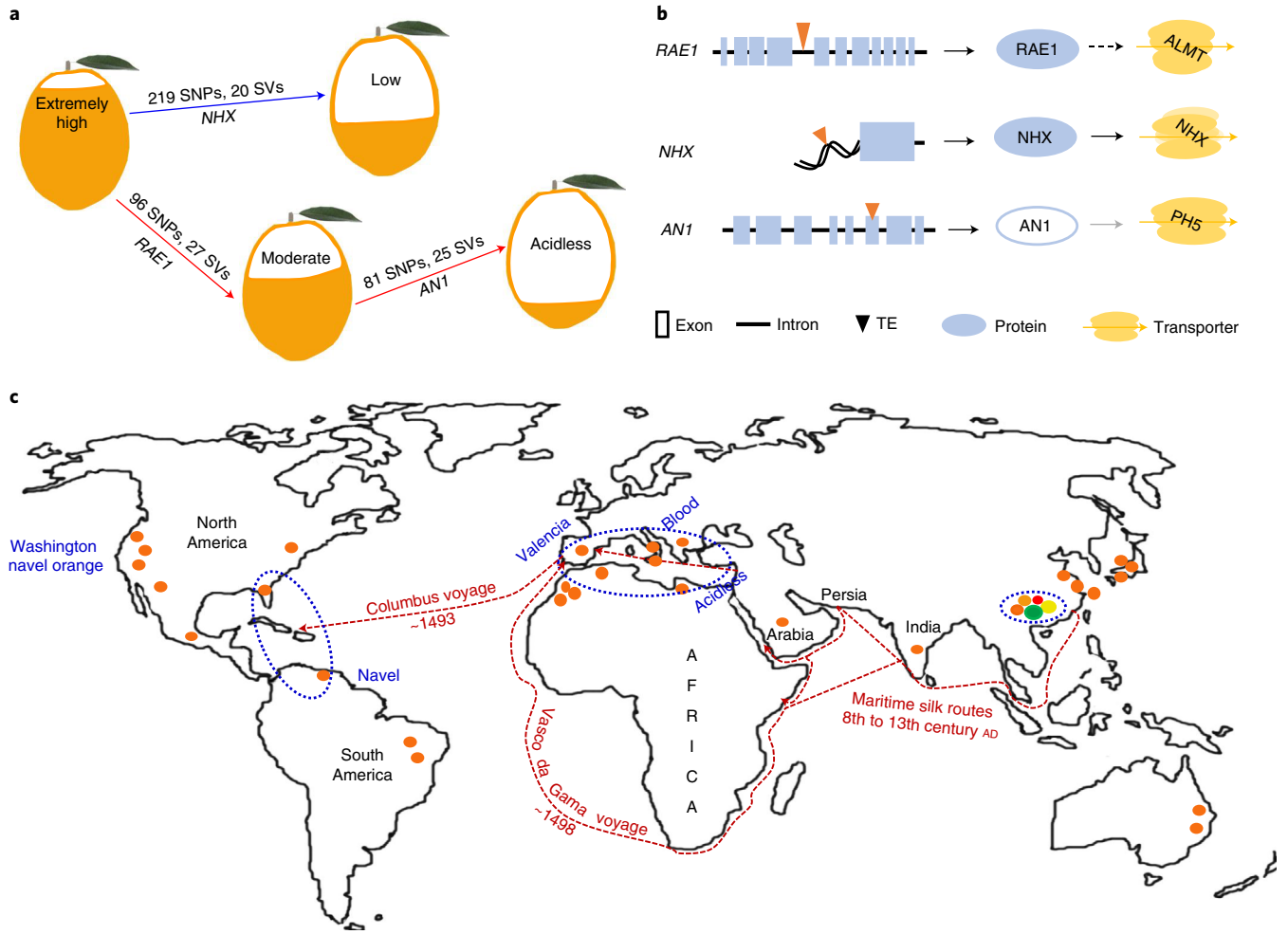


Fig. 5 | The diversification and dispersal history of sweet orange. a, Somatic variations and genes associated with the diversification of acidity during the dispersal history of sweet orange. The somatic variations indicated above the arrows are the group-specific variations indicated by the arrow. The genes indicated below the arrows are candidate genes harbouring the mutations discussed in the main text. BTC is used as the representative of the low-acid type, and blood orange is used as the representative of the moderate-acid type. Succari and HAL are used as representatives of the acidless type. **b**, Model for the relationship between the mutated genes and the transporter genes. The TE insertions are indicated with inverted triangles. *RAE1* encodes a F-box protein that modulates the expression of genes that encode ALMT transporters[31,32]. *NHX* encodes a Na^+/H^+ transporter that is reported to affect pH in *Arabidopsis*[35]. *AN1* encodes a bHLH transcription factor that regulates the expression of the *PH5* gene, which encodes a vacuolar proton pump that affects fruit acidity[37,38]. **c**, A possible dispersal history of sweet orange. The three major centres of diversity for sweet orange were South China, the Mediterranean region and the Americas (indicated by the dashed blue ovals). The possible routes used to spread sweet orange—data from ref. [24]—are indicated with red arrows. The red circles indicate the origins of the sweet orange mutants sequenced in this study.

trees[15,16] and annual plants[44]. By comparison, the somatic mutations discovered in long-lived trees are a type of neutral situation under natural conditions that is distinct from the somatic mutations in sweet orange because domesticated cultivars were used for mutation identification. The dataset from somatic mutants of sweet orange is a manifestation of both somatic mutation and human selection. Large variations, such as TE transposition, or large SVs, including deletion, duplication or even mitotic recombination, were observed in these mutants.

Our study revealed abundant transpositions in the 114 somatic mutants of sweet orange. The activity of TEs can induce somatic mutations[10,45]. In our study, we found 877 TE insertions, 52.3% of which were LTR retrotransposon insertions. Unexpectedly, 30.5% were MuLE-type DNA transposons. Although several cases of retrotransposons generating somatic mutations have been reported, the role of MuLE transposons in generating somatic mutations is less well known. For instance, a somatic mutation that changes the colour of grape skin from white to red is caused by the movement of a retrotransposon named Gret1 (refs. [11,41]). Additionally, the Copia-like retrotransposon inserted in the promoter of the *Ruby* gene, which encodes a MYB transcription factor that induces the accumulation of anthocyanins, is responsible for the red flesh colour of blood orange[10]. The Copia-like retrotransposons Tcs4 and Tcs6x inserted into the *AN1* gene[38] in Vaniglia sweet orange were recently reported to be highly associated with fruit acidity[38,39]. In an interesting case, a MuLE transposon was shown to be associated with a large deletion and somatic mutation in clementine mandarin[37].

Higher frequencies of somatic mutants were observed in species with high heterozygosity relative to species with low heterozygosity. Sweet orange—a frequency of 80% somatic mutants in the cultivars—is an inter-species hybrid between pummelo and mandarin[6]. Moreover, the frequencies are 72% in grapefruit varieties (a hybrid between pummelo and sweet orange) and 97% in satsuma mandarin varieties (a hybrid between two mandarins). These are markedly higher frequencies than we observed in basal citrus species, which

have a more homozygous background (for example, 16% in mandarin and 6% in pummelo). Indeed, an interesting study of spontaneous mutations during meiosis revealed a 3.5-fold higher mutation rate in heterozygotes than in homozygotes[46]. Our data also provide hints of large somatic variations (for example, >500 kb) in sweet oranges, and this may be due to the advantage of heterozygosity and the large variations that occurred in one diploid genome. Similar observations were reported in other hybrid citrus. For instance, a 2-Mb deletion was observed in clementine mandarin, which is a hybrid between mandarin and sweet orange[37]. Chromosome replacement and deletion in the anthocyanin gene cluster led to partial homozygosity and clonal polymorphism in the berry colour of grape[47]. SNP analysis also indicated that 23.34% of somatic mutations are homozygous mutations (that is, conversion from the heterozygous to the homozygous state). It was proposed that the loss of heterozygosity drove the clonal diversity of the microbe *Phytophthora capsici*[48]. A heterozygous genetic background may provide an evolutionary advantage, as it could buffer the effects of deleterious genetic mutations and increase the survival rate under stress[29].

The domestication and dispersal histories of sweet orange are probably associated with the selection and diversification of fruit acidity. We found a prototype of sweet orange with extremely high levels of fruit acidity and that citric acid was dramatically reduced in the widely cultivated sweet orange compared with its progenitor—the high-acid sweet orange from South China. Different levels of fruit acidity, including moderate, low and no acid, were detected in varieties domesticated in different locations during the dispersal history of sweet orange. Historically, there were many independent introductions of sweet orange from East Asia to Persia, East Africa and Europe via ancient maritime silk routes[24]. The first variety that was distributed outside China accumulated only low levels of acid and sugar in its fruits and was possibly distributed through the commercial route established by the Genoese[23]. Sweet orange was also mentioned in a medical book, entitled "Charaka sambita", in 1100 AD[49]. A variety introduced later to Europe, called "China orange", was a superior cultivar that accumulated high levels of both sugar and acid. The Portuguese brought this variety from China circa 1500 AD. Columbus subsequently introduced it to the Americas[23,24]. Generally, mutants that originated in the Mediterranean area and the Americas showed larger and more frequent variations than the mutants that originated in China. Large SVs frequently occurred in navel orange, which originated in the Americas. These results indicate that the different habitats and environments encountered during the spread of sweet orange may have had different influences on its somatic variation.

Citric acid has a major impact on fruit flavour and the juicing quality of sweet orange. It is also of industrial value and is widely used as an acidifier and as a flavouring and chelating agent. Recent studies[38,39] of two acidless mutants revealed that a vacuolar proton pump affects fruit acidity and that a bHLH transcription factor regulates the expression of the gene that encodes this proton pump. We detected a variety of mutations by collecting and comparing an extremely high-acid prototype of sweet orange to moderate-acid, low-acid and acidless sweet oranges. We found mutations in genes that encode a variety of transporters and that the types of transporters encoded by these genes were related to the level of acidity. The retrotransposon insertion in *RAE1* in the moderate-acid type probably modulates acidity by interacting with a zinc-finger transcription factor that regulates the expression of genes encoding the transporter ALMT[31]. The mutation leading to low acidity was associated with a DNA transposon inserted upstream of the Na^+/H^+ transporter gene, which was reported to affect vacuolar pH[35]. Gene expression data also supported our conclusion that the *NHX* and *RAE1* genes are upregulated during the later stages of fruit development when fruit acidity is decreasing. Retrotranspositions in acidless sweet oranges affected a gene encoding a bHLH transcription factor that is known to affect fruit acidity by regulating the transcription of a gene that encodes a vacuolar proton pump[38,39]. This series of transporter genes and regulatory genes provide candidate genes for future mechanistic studies of the regulation of fruit acidity and will ultimately help us fine-tune fruit taste and flavour.

Methods

Plant materials and population sequencing data. A total of 20 sweet oranges were collected in the regions containing the NYW region. In addition to the previous sequencing data of nine sweet oranges[50], we newly sequenced 105 types of sweet orange including landraces, well-known cultivars in Asia, the Americas and the Mediterranean (Supplementary Table 4). At least 10 μg of genomic DNA from each accession was used to construct a sequencing library. Paired-end sequencing libraries with an insert size of approximately 200–500 bp were constructed and sequenced on the Illumina platform.

De novo assembly and annotation of di-haploid sweet orange genomes. Callus of the double haploid sweet orange was used for the extraction of DNA[20] and genome assembly. A total of 27.0 Gb of data (approximately 73.5× genome coverage) for the di-haploid sweet orange were generated using the PacBio RSII platform. First, the raw reads were corrected and assembled using the hierarchical genome assembly process (HGAP)[51] in SMRT Analysis (v.2.3.0). Another round of polishing was performed using quiver and Pilon[52] to further improve the quality of the assembly. Next, mate-pair reads of 2 kb, 10 kb and 20 kb insert size libraries (110.9× genome coverage) were used to construct scaffolds with the SSPACE-STANDARD package[53]. Then, the GapCloser package[54] was utilized to close gaps in the original scaffolds and the SSPACE-STANDARD package was used to further extend gap-filled scaffolds.

The Illumina reads were mapped to the assembled genomes using BWA[55], with a mapping rate of over 90%, followed by the application of SAMtools[56] to call variants. Homozygous mismatches were regarded as assembly errors. The accuracy rates were higher than 99.99% for the genome. The completeness of the genomes was checked by mapping 1,440 orthologous genes from plants to the genomes using the BUSCO[57] software. A total of 1,371 orthologous genes that occupy 95% of the plant gene set were mapped to the genomes of *C. sinensis*. Finally, the assembled RNA sequences were mapped to the genome to check the transcript coverage. More than 90% of the assembled RNA sequences were mapped to the corresponding genomes.

To further improve the genome of sweet orange, a total of 34.6 Gb data from the nanopore ultra-long platform were generated. Necat[58] was used to assemble the genome, followed by three rounds of polishing using nanopore reads and Illumina reads using Recon[59] and Nextpolish[60], respectively. A Hi-C library was constructed for chromosome-level scaffolding, and 37.6 Gb data were generated. With the Hi-C library, the polished contigs were anchored to nine super-scaffolds using the 3d-dna pipeline[61] and juicer[62].

The TE libraries of *C. sinensis* were first constructed using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/), which is a de novo repeat families identification and modelling package. Subsequently, the library was combined with the Repbase plant repeat database. The TE libraries were finally used to mask the genome using the RepeatMasker software (http://www.RepeatMasker.org).

Gene models were annotated based on ab initio gene predictions, homology support and RNA sequencing evidence. For ab initio gene predictions, AUGUSTUS[63] and GlimmerHMM[64] were employed using the default parameters for *Arabidopsis thaliana* and *Oryza sativa*. The homology EST and protein databases were constructed by integrating the citrus EST and protein sequences from the NCBI databases and SwissProt databases. Then, the homology searches were done using exonerate[65] and AAT[66] software. In addition, RNA sequencing reads from mixtures of tissues were generated. Trinity software[67] was utilized to do genome-guided and de novo transcript assembly. All of the gene structures predicted using the aforementioned methods were combined using EVM software[68].

Genome assembly of six representative diploid sweet orange and phasing two haplotypes of Valencia sweet orange. PacBio reads, 10X Genomics linked reads and Illumina sequencing reads of Valencia sweet orange were mapped to our assembled genome using NGMLR[69], Long Ranger (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger) and BWA, respectively. SVs were identified using Sniffles[69] with the PacBio mapping result. Each SV was supported by at least ten reads. SNPs were also called using the GATK package with Illumina mapping results. Then, extractHairs tools from HapCUT2 software[70] were utilized on both PacBio reads and 10X Genomics linked reads. The output fragments were combined and used as the input for HapCUT2. The results were converted to a VCF file of phased SNPs. Finally, the phased SNPs, coupled with SVs and PacBio mapping results, were used to perform diploid genome assembly using Crossstitch software (https://github.com/schatzlab/crossstitch).

PacBio long reads were generated for TCPS1 and NHE. Nanopore long reads were generated for NW, BT2 and UKXC. A total of 12.7 Gb and 9.4 Gb of PacBio long reads were generated for TCPS1 and NHE. A total of 14.7 Gb, 27.1 Gb

and 37.4 Gb of nanopore long reads were generated for NW, BT2 and UKXC, respectively. The long reads were assembled using Nextdenovo (https://github.com/Nextomics/NextDenovo) and polished using Racon[59] and Nextpolish[60]. Then, the redundant sequences were removed using Pruge_haplotigs[71]. RaGOO[72] was finally used to anchor the contigs to nine super-scaffolds using the di-haploid sweet orange genome as reference.

Mapping and variant calling. Paired-end reads of all accessions were mapped to the sweet orange reference genome using BWA (v.0.7.5a-r405)[73] using the following parameters: 'aln -o 1 -e 10 -t 12 -l 32 -i 15 -q 15'. Duplicated mapping reads were removed using Samtools[56]. Then, 'AddOrReplaceReadGroups.jar' in the picard-tools package (v.1.105) (https://github.com/broadinstitute/picard) was used to add the read groups to each library. IndelRealigner in the GATK package[74] was used to perform local realignments around the InDels.

All of the genotype information from the polymorphic sites was retrieved using the GATK population method. This procedure yielded high-quality variations for each of the 70 individuals. These sets of SNPs were filtered based on sequence depth. The genotype of each particular individual was retained if the depth was between 4 and 150. We further filtered SNPs by retaining only non-singleton, biallelic SNPs with non-missing data across all sweet oranges.

The SNPs that we identified were further categorized based on their locations within the genome (for example, intergenic regions, untranslated regions, coding sequences and introns) using the gene model annotations for sweet orange. The SNPs were grouped into synonymous and nonsynonymous types.

Phylogeny and PCA. To analyse the phylogenetic relationships among all 114 sweet oranges, we used the somatic SNPs from whole genomes that we identified using the method described above to construct the phylogenetic tree using RaxmlHPC[75] (v.8.0.0). The PCA was performed using the population variant data from EIGENSTRAT[76] software.

Somatic SNPs and InDels. A population-level comparison of the genotypes was employed to predict the somatic variants. Due to the wide geographical distribution of the mutants, different mutations will occur and, therefore, different lineages have different variations. We suggest that the somatic variations in the sweet orange population are not specific to particular lines. Somatic mutations will be present at different frequencies in the mutant population.

Variant calling (SNP and InDel) was applied to the 114 sweet oranges using GATK[74] (-T UnifiedGenotyper(UG)). To increase the screen efficiency, 114 sweet oranges were randomly divided into four groups that contained approximately 30 individuals per group. For each group, read depth, mapping quality and genotype quality were used for quality control. The requirements were as follows: the lowest read depth (DP) in each group ≥10; mapping quality (MP) of ≥30; average genotype quality of ≥30; and neighbouring SNPs and InDels are separated by 150 bp. SNP and InDel sites in the regions of SV were excluded. Only biallelic SNP sites were used for downstream analysis.

Then, allele metrics and entire candidate SNP and InDel sites from the last step were generated using 'bam-readcount' (https://github.com/genome/bam-readcount) with the parameter '–min-mapping-quality 20–min-base-quality 15' for each genomic mapping file (bam format). The obtained metrics were used for further filtering for homozygous mutations and heterozygous mutations.

A heterozygous mutation was defined as a new allele that occurred in the homozygous background in the population and was considered a candidate somatic SNP. Homozygous mutations were also considered when a homozygous genotype occurred in a predominantly heterozygous background. Both candidate heterozygous and homozygous mutations were used for subsequent allele ratio tests. Reference allele ratios (reference allele depth/total allele depth) were used to evaluate the mutations.

High-quality candidate somatic SNP and InDel mutations met the following conditions: the mean base quality for the reads containing the allele was ≥40; at the mutation site, the reference allele ratio equals 0 or 100% for the homozygous genotype; and the reference allele ratio is ≥35% and ≤75% for the heterozygous genotype. Among the somatic population of 114 sweet orange varieties, both the mutant genotypes and background genotypes should fit the two above-mentioned requirements.

We also called InDels using the HaplotypeCaller (HC) algorithm implemented in GATK. The consensus sites based on both the final UG and HC results were defined as somatic InDels. Finally, the genotypes of the somatic SNPs and InDels were extracted using VCFtools[77].

Identification of candidate deleterious mutations. Amino acid substitutions based on the UniProt database and their effects on protein function were predicted with the SIFT algorithm. Particular amino acid substitutions were predicted to be deleterious if the score was ≤0.05 and tolerated if the score was ≥0.05.

We used a comparative genomics approach to predict the genomic constraints using the sweet orange reference with eight other species, including *Litchi chinensis*, *Carica papaya*, *A. thaliana*, *Linum usitatissimum, Manihot esculenta*, *Populus trichocarpa*, *Eucalyptus grandis* and *Gossypium hirsutum*. All genome sequences were downloaded from the Phytozome website (https://phytozome.jgi.doe.gov/pz/portal.html) with the exception of *L. chinensis*. Phylogenetic models were estimated using the results from the maximum likelihood (ML) phylogenetic tree made with 1,000 single-copy genes shared by nine species constructed using RaxmlHPC (v.8.0.0).

We used scripts in the CNSpipline[78] (https://github.com/liangpingping/CNSpipeline.git) to prepare the reference genome for subsequent alignments. LASTZ was used to align the assembled genome sequences to the sweet orange genome. An eight-way multiple alignment was processed following the method described by Hubisz[79]. Based on multiple alignments and estimated phylogenetic models, conservation scores were calculated for 100-kb non-overlapping windows using phastCon with the following parameters: –target-coverage 0.25–expected-length 12–rho 0.4 (ref. [79]).

Somatic SV identification. A large genomic region with a heterozygous SNP density continually decreasing relative to other sweet oranges was used as a signature of SV. We calculated the heterozygous SNP density in the sweet orange genome in 2-kb windows. Then, we compared the heterozygous SNP distribution among 114 sweet oranges.

Deletion and duplication were predicted by CNV-seq[80] and FREEC[81] and confirmed by the two allele frequencies. FREEC was utilized with a sliding window of 5 kb using default parameters. CNV-seq parameters were "–log2 0.6 –p 0.001 –bigger-window 1.5 –annotate–minimum-windows 4". If results from both software showed significant copy number variation and ratio of two allele frequencies significantly deviated from 1:1, we defined the copy number variation as deletion or duplication.

Somatic TE identification. Generally, the detection of TE insertions was based on clusters of paired reads whereby one read mapped onto a unique location of the genome and the other to a genomic repetitive region, which was used to determine the type of the large insertion. In addition, clipped reads have been utilized to call SVs at the single-nucleotide resolution. These combined features will predict the exact position of the insertion points and enable the identification of large SV signatures. Based on this method, we developed a population comparison scheme to detect LFI and somatic transposition (Supplementary Fig. 27). Five individuals were excluded: NHE, NW, SO1, SO2 and SO3. Read mapping quality over 30 was retained for downstream analysis.

We first extracted discordant read (DRs) and identified the candidate windows with LFIs. The DRs were extracted with the command "samtools view sample.mem.bam |awk '($7!='='){print}' > sample.unmapped.mem.txt". DRs in 1-kb windows were calculated for each individual. We merged these results in a table and sorted by the window position along the genome within the group. The following filters were used to detect the candidate window with LFIs. We determined the DR enrichment difference index (D). We limited the window in which the maximum DRs of an individual in the group were more than 10 and less than 150. The number of random DRs at each window of a normal individual was estimated to be 2. The total number of DRs in a window of the group is N_0. All DRs more than ten of individuals was T. The number of group individuals was G. The DR enrichment difference between the mutant and wild type was defined as D:

$$D = (G - \mathrm{Int}(T/10)) \times 2 - (N_0 - T)$$

Fisher's exact test was used to test the significance of the D value in the group. We used the ML method to estimate N_0, T and D of the window with DR enrichments in the equation. When D is between the $[-10, 50]$ and $P \leq 0.05$, we kept the window as the candidate with a LFI.

We then designed the following steps to detect breakpoints. The extractSplitReads script in Lumpy software was used to report the clipping reads. The command was "samtools view -h sample.mem.bam | extractSplitReads_BwaMem -i stdin | samtools view -Sb - > sample.splitters.unsorted.bam". The site in the candidate window with the LFI supported by most split reads was kept as the breakpoint of a large insertion. We performed BLAST searches of the repeats database of the reference genome using the reads as queries. Transposition type was predicted using the mate-pair location and the results of discordant reads in the repeats database.

Genome walking was employed to clone the full-length transposon sequence that was inserted in the candidate gene. We annotated the full transposon sequence using the LTR-finder (http://tlife.fudan.edu.cn/ltr_finder/) and Interproscan (online version http://www.ebi.ac.uk/interpro/search/sequence-search).

Quantitative PCR with reverse transcription analysis. Total RNA from all of the frozen tissues was extracted using RNAiso Plus (Total RNA extraction reagent; TaKaRa). Complementary DNA was synthesized using 1 μg of total RNA and a HiScriptII Q RT SuperMix for qPCR (with gDNA wiper) kit (Vazyme). Quantitative PCR with reverse transcription was performed with a Roche LightCycler 480 II instrument (Roche Applied Science) using Hieff qPCR SYBR Green Master Mix according to the manufacturer's instructions (YEASEN). RNA extraction and cDNA synthesis were performed with three biological replicates for each sampling point. The citrus β-actin gene was used as the internal control. The primers used are listed in Supplementary Table 16.

Transient overexpression in kumquat fruits. The transient overexpression assay using *F. margarita* (kumquat) fruit at the green mature stage was performed using the procedure developed for citrus fruits[82] with slight modifications. Full-length coding sequences of target genes (*CsNHX* and *CsRAE1*) were amplified with specific primers (Supplementary Table 16) and cloned into the pK7WG2D vector with the GFP marker using Gateway Technology. *Agrobacterium tumefaciens* strain EHA105 was separately transformed with these constructs and the empty vector. These *A. tumefaciens* strains containing the empty vector (pK7WG2D) or the vector harbouring the target genes were infiltrated into two uniform sections of kumquat fruit. Three days after infiltration, samples were collected from the infiltrated regions with fluorescence. A total of 20 fruits were used for this assay, which were separated into four pools with each containing five independent infiltrated fruit pieces as a biological replicate for the quantification of pH value and gene expression levels.

Sugar and organic acid determinations. Citrus fruits were squeezed to produce juice that was filtered before use. Soluble solid content was determined using a saccharimeter (ATAGO). Acidity was measured by titration with 0.1 M NaOH using phenolphthalein as an indicator. Three biological replicates and two technical replicates were used to analyse each sample. The pH value was estimated with a STARTER 3100 pH metre (OHAUS).

Compositions and concentrations of both soluble sugars and organic acids were determined using gas chromatography (5%-phenyl-methyl polysiloxane; 30 m × 320 μm i.d. of 0.25 μm) as previously described[83].

Heterozygosity estimation. The raw sequencing data were filtered and the high-quality reads were used to estimate heterozygosity based on k-mer analysis. The heterozygosity was estimated using GCE software[84] with 17-bp k-mer lengths. The heterozygosity was defined as the proportion of heterozygous base pairs in the genome.

Statistics and reproducibility. Validation of SV was carried out by PCR experiment and gel scans. The primers of representative experiments and the times of the independent experiments with similar results obtained are provided in Supplementary Tables 9 and 10.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Genome data for di-haploid *C. sinensis* v.3.0 and v.4.0 have been deposited at DDBJ/ENA/GenBank under accession numbers MORK00000000 and JAFBAU000000000, respectively. The genome data for six diploid sweet oranges have been deposited at NCBI under accession PRJNA321100. All of the genome sequencing data and transcriptome sequencing data have been deposited at the Sequence Read Archive (SRA) database at NCBI. The PacBio and nanopore sequencing data for *C. sinensis* were deposited under the SRR accession number SRR5838837. The sequencing data that support the findings of this study have been deposited in the SRA database under accession PRJNA321100. The SRR accessions for whole-genome sequencing data and six diploid sweet oranges can be found in Supplementary Table 4. Sweet orange genome sequences are also available from our website at http://citrus.hzau.edu.cn/orange. All supporting data are included in the Supplementary Information. Source data are provided with this paper.

Received: 25 January 2020; Accepted: 11 May 2021;
Published online: 17 June 2021

References

1. Miller, A. J. & Gross, B. L. From forest to field: perennial fruit crop domestication. *Am. J. Bot.* **98**, 1389–1414 (2011).
2. Mckey, D., Elias, M., Pujol, B. & Duputié, A. The evolutionary ecology of clonally propagated domesticated plants. *New Phytol.* **186**, 318 (2010).
3. Gaut, B. S., Diez, C. M. & Morrell, P. L. Genomics and the contrasting dynamics of annual and perennial domestication. *Trends Genet.* **31**, 709–719 (2015).
4. Shamel, A. D. & Pomeroy, C. S. Bud mutations in horticultural crops. *J. Hered.* **27**, 487–494 (1936).
5. Mendel, K. Bud mutations in *Citrus* and their potential commercial value. *Int. Soc. Citriculture* **1**, 86–89 (1981).
6. Poduri, A., Evrony, G., Cai, X. & Walsh, C. A. Somatic mutation genomic variation and neurological disease. *Science* **341**, 1237758 (2013).
7. Li, M. et al. Characterization of salt-induced epigenetic segregation by genome-wide loss of heterozygosity and its association with salt tolerance in rice (*Oryza sativa* L.). *Front. Plant Sci.* **8**, 977 (2017).
8. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
9. Yao, J., Dong, Y. & Morris, B. A. Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. *Proc. Natl Acad. Sci. USA* **98**, 1306–1311 (2001).
10. Butelli, E. et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* **24**, 1242–1255 (2012).
11. Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. Retrotransposon-induced mutations in grape skin color. *Science* **304**, 982 (2004).
12. Fernandez, L., Torregrosa, L., Segura, V., Bouquet, A. & Martinez-Zapater, J. M. Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine. *Plant J.* **61**, 545–557 (2010).
13. Carbonell-Bejerano, P. et al. Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiol.* **175**, 786–801 (2017).
14. Hiltunen, M., Grudzinska-Sterno, M., Wallerman, O., Ryberg, M. & Johannesson, H. Maintenance of high genome integrity over vegetative growth in the fairy-ring mushroom *Marasmius oreades*. *Curr. Biol.* **29**, 2758–2765 (2019).
15. Schmid-Siegert, E. et al. Low number of fixed somatic mutations in a long-lived oak tree. *Nat. Plants* **3**, 926–929 (2017).
16. Plomion, C. et al. Oak genome reveals facets of long lifespan. *Nat. Plants* **4**, 440–452 (2018).
17. Yu, L. et al. Somatic genetic drift and multilevel selection in a clonal seagrass. *Nat. Ecol. Evol.* **4**, 952–962 (2020).
18. 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).
19. Wu, G. A. et al. Genomics of the origin and evolution of *Citrus*. *Nature* **554**, 311–316 (2018).
20. Xu, Q. et al. The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **45**, 59–66 (2013).
21. Talon, M. & Gmitter, F. G.Jr. Citrus genomics. *Int. J. Plant Genomics* **2008**, 528361 (2008).
22. Zhou, K. L. & Ye, M. M. *Chinese Fruit Tree: Citrus* (China Forestry Publishing House, 2010).
23. Spiegel-Roy, P. & Goldschmidt, E. E. in *The Biology of Citrus* (eds Spiegel-Roy, P. & Goldschmidt, E. E.) 4–18 (Cambridge University Press, 1996).
24. Webber, H. J., Batchelor, L. D. & Reuther, W. in *The Citrus Industry* (eds Reuther, W. et al.) 1–39 (Univ. California Press, 1967).
25. Etienne, A., Genard, M., Lobit, P., Mbeguie, A. M. D. & Bugaud, C. What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells. *J. Exp. Bot.* **64**, 1451–1469 (2013).
26. Jiang, T. M. Preliminary study on selection of sweet orange buds in Qianyang region. *South China Fruits* **2**, 9–12 (1980).
27. Wang, L. et al. Genome of wild mandarin and domestication history of mandarin. *Mol. Plant* **11**, 1024–1037 (2018).
28. Moore, G. A. Oranges and lemons: clues to the taxonomy of *Citrus* from molecular markers. *Trends Genet.* **17**, 536–540 (2001).
29. Ramu, P. et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
30. Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D. & Gaut, B. S. Evolutionary genomics of grape (*Vitis vinifera* ssp. vinifera) domestication. *Proc. Natl Acad. Sci. USA* **114**, 11715–11720 (2017).
31. Zhang, Y. et al. F-box protein RAE1 regulates the stability of the aluminum-resistance transcription factor STOP1 in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **116**, 319–327 (2019).
32. Liu, M. Y. et al. Two citrate transporters coordinately regulate citrate secretion from rice bean root tip under aluminum stress. *Plant Cell Environ.* **41**, 809–822 (2018).
33. Fan, L. G. et al. Na^+, K^+/H^+ antiporters regulate the pH of endoplasmic reticulum and auxin-mediated development. *Plant Cell Environ.* **41**, 850–864 (2018).
34. Bassil, E. et al. Cellular ion homeostasis: emerging roles of intracellular NHX Na^+/H^+ antiporters in plant growth and development. *J. Exp. Bot.* **63**, 5727–5740 (2012).
35. Bassil, E., Zhang, S., Gong, H., Tajima, H. & Blumwald, E. Cation specificity of vacuolar NHX-type cation/H^+ antiporters. *Plant Physiol.* **179**, 616–629 (2019).
36. Zhang, M. et al. A HAK family Na^+ transporter confers natural variation of salt tolerance in maize. *Nat. Plants* **5**, 1297–1308 (2019).
37. Terol, J. et al. Involvement of a citrus meiotic recombination TTC-repeat motif in the formation of gross deletions generated by ionizing radiation and MULE activation. *BMC Genomics* **16**, 69 (2015).
38. Butelli, E. et al. *Noemi* controls production of flavonoid pigments and fruit acidity and illustrates the domestication routes of modern citrus varieties. *Curr. Biol.* **29**, 158–164 (2019).
39. Strazzer, P. et al. Hyperacidification of *Citrus* fruits by a vacuolar proton-pumping P-ATPase complex. *Nat. Commun.* **10**, 744 (2019).
40. Deng, X. et al. Retrospection and prospect of fruit breeding for last four decades in China (in Chinese). *J. Fruit Sci.* **36**, 514–520 (2019).
41. Lijavetzky, D. et al. Molecular genetics of berry colour variation in table grape. *Mol. Genet. Genomics* **276**, 427–435 (2006).

42. Vondras, A. M. et al. The genomic diversification of grapevine clones. *BMC Genomics* **20**, 972 (2019).
43. Wang, L. et al. The architecture of intra-organism mutation rate variation in plants. *PLoS Biol.* **17**, e3000191 (2019).
44. Lovell, J. T., Williamson, R. J., Wright, S. I., McKay, J. K. & Sharbel, T. F. Mutation accumulation in an asexual relative of *Arabidopsis*. *PLoS Genet.* **13**, e1006550 (2017).
45. Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **47**, 1435–1442 (2015).
46. Yang, S. et al. Parent–progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**, 463–467 (2015).
47. Pelsy, F., Dumas, V., Bevilacqua, L., Hocquigny, S. & Merdinoglu, D. Chromosome replacement and deletion lead to clonal polymorphism of berry color in grapevine. *PLoS Genet.* **11**, e1005081 (2015).
48. Hu, J. et al. Genetically diverse long-lived clonal lineages of *Phytophthora capsici* from pepper in Gansu, China. *Phytopathology* **103**, 920–926 (2013).
49. Calabrese, F. in *Citrus: The Genus Citrus (*eds Dugo, G. & Di Giacomo, A*)* 1–15 (Taylor & Francis, 2002).
50. Wang, X. et al. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nat. Genet.* **49**, 765–772 (2017).
51. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
52. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
53. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2010).
54. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
55. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
56. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
57. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
58. Chen, Y. et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun.* **12**, 60 (2021).
59. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 737–746 (2017).
60. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
61. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
62. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
63. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
64. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
65. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
66. Huang, X. Q., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37–45 (1997).
67. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
68. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
69. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
70. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2016).
71. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
72. Alonge, M. et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
73. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
74. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
75. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
76. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
77. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
78. Liang, P., Saqib, H. S. A., Zhang, X., Zhang, L. & Tang, H. Single-base resolution map of evolutionary constraints and annotation of conserved elements across major grass genomes. *Genome Biol. Evol.* **10**, 473–488 (2018).
79. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).
80. Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
81. Boeva, V. et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2011).
82. Li, S. J. et al. Citrus *CitNAC62* cooperates with CitWRKY1 to participate in citric acid degradation via up-regulation of *CitAco3*. *J. Exp. Bot.* **68**, 3419–3426 (2017).
83. Liu, Q. et al. A novel bud mutation that confers abnormal patterns of lycopene accumulation in sweet orange fruit (*Citrus sinensis* L. Osbeck). *J. Exp. Bot.* **58**, 4161–4171 (2007).
84. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Preprint at http://arxiv.org/abs/1308.2012 (2012).

Acknowledgements

We thank Y. Zhang from Chongqing Academy of Agricultural Sciences and W. Song from Zigui Agricultural Bureau, Yichang for sampling support. We also thank L. Chen for suggestions on the bioinformatics analysis. This project was financially supported by the National Key Research and Development Program of China granted to Q.X. (number 2018YFD1000101), the National Natural Science Foundation of China granted to Q.X. (numbers 31925034 and 31872052), the Fundamental Research Funds for the Central Universities granted to Q.X. (number 2662015PY109) and the support from Agricultural Research Service, US Department of Agriculture (number 8062-21000-043-02S to E.S.B.). L.W. was supported by the China Postdoctoral Science Foundation (number 2020M672375).

Author contributions

Q.X. conceived and designed the project. L.W. developed the method for the bioinformatics analyses of the somatic mutant, designed primers for experiments, prepared the figures and coordinated teamwork. Y.H. assembled the sweet orange genomes and performed gene annotation. Z. Liu carried out the somatic variant validation experiments (with contribution by J.H.). Z. Liu and J.H. performed gene expression. Z. Liu, Z. Lu and J.H. performed the transient overexpression experiments. F.H., X.J., S.Y., P.C., B.Z., L.K. and Z.X. collected and evaluated the samples. Z. Liu, F.H. and J.H. measured the fruit quality. Z. Liu, H.Y. and L.K. performed the DNA and RNA extraction experiments. D.J. provided partial sweet-orange samples. E.S.B. and R.M. supervised the bioinformatics analyses. Q.X., L.W., Y.H. and R.M.L. wrote the manuscript with contributions from X.D. and R.M.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41477-021-00941-x.

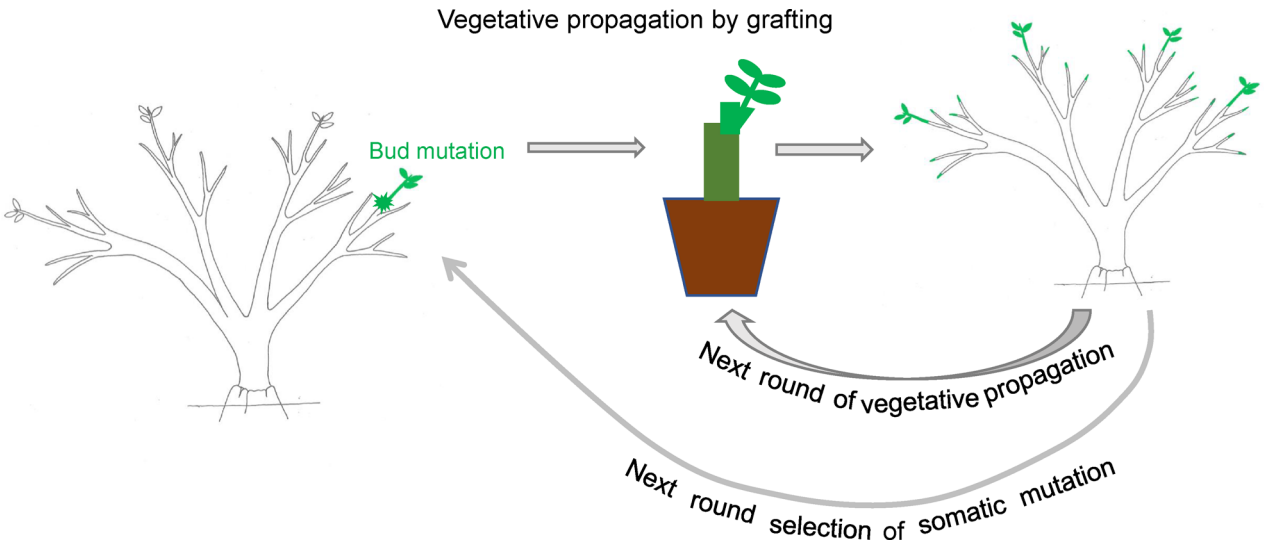
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-021-00941-x.

Correspondence and requests for materials should be addressed to Q.X.

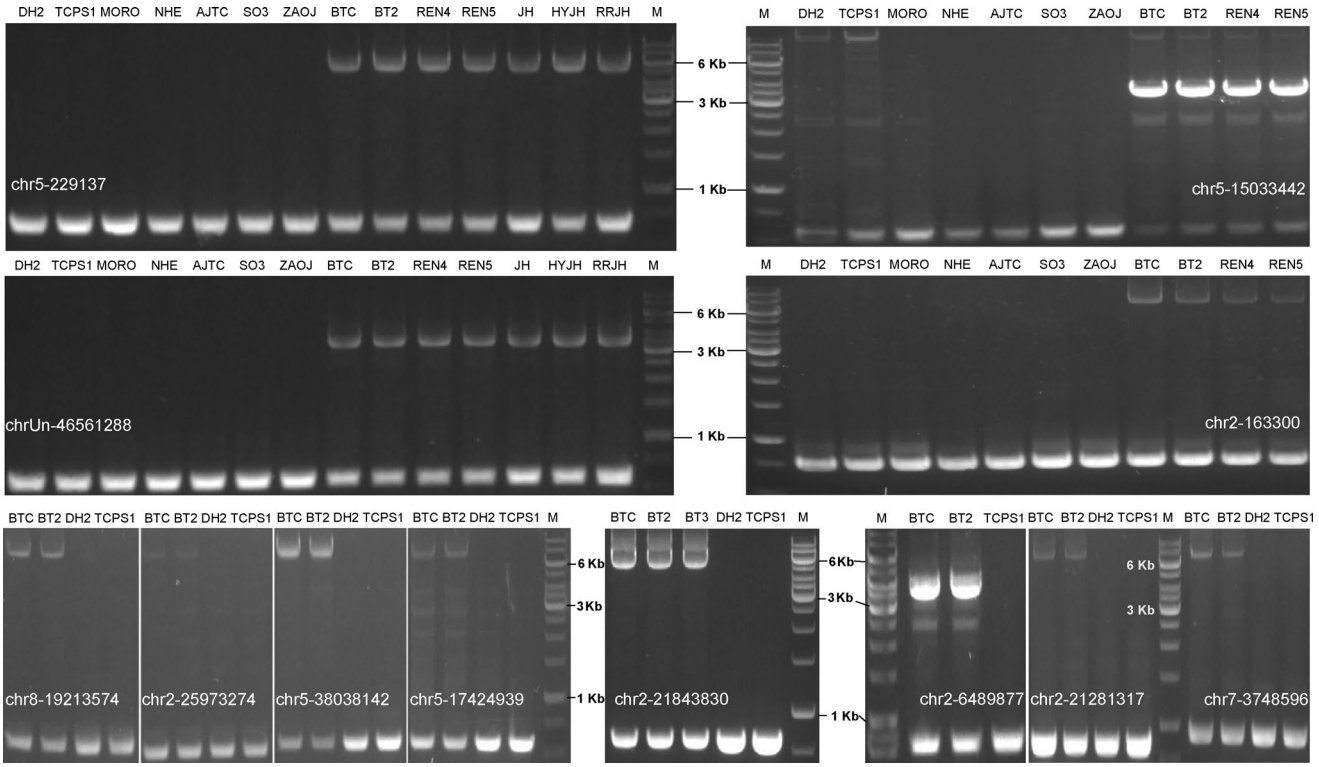
Peer review information *Nature Plants* thanks Olivier Panaud, Dacheng Tian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

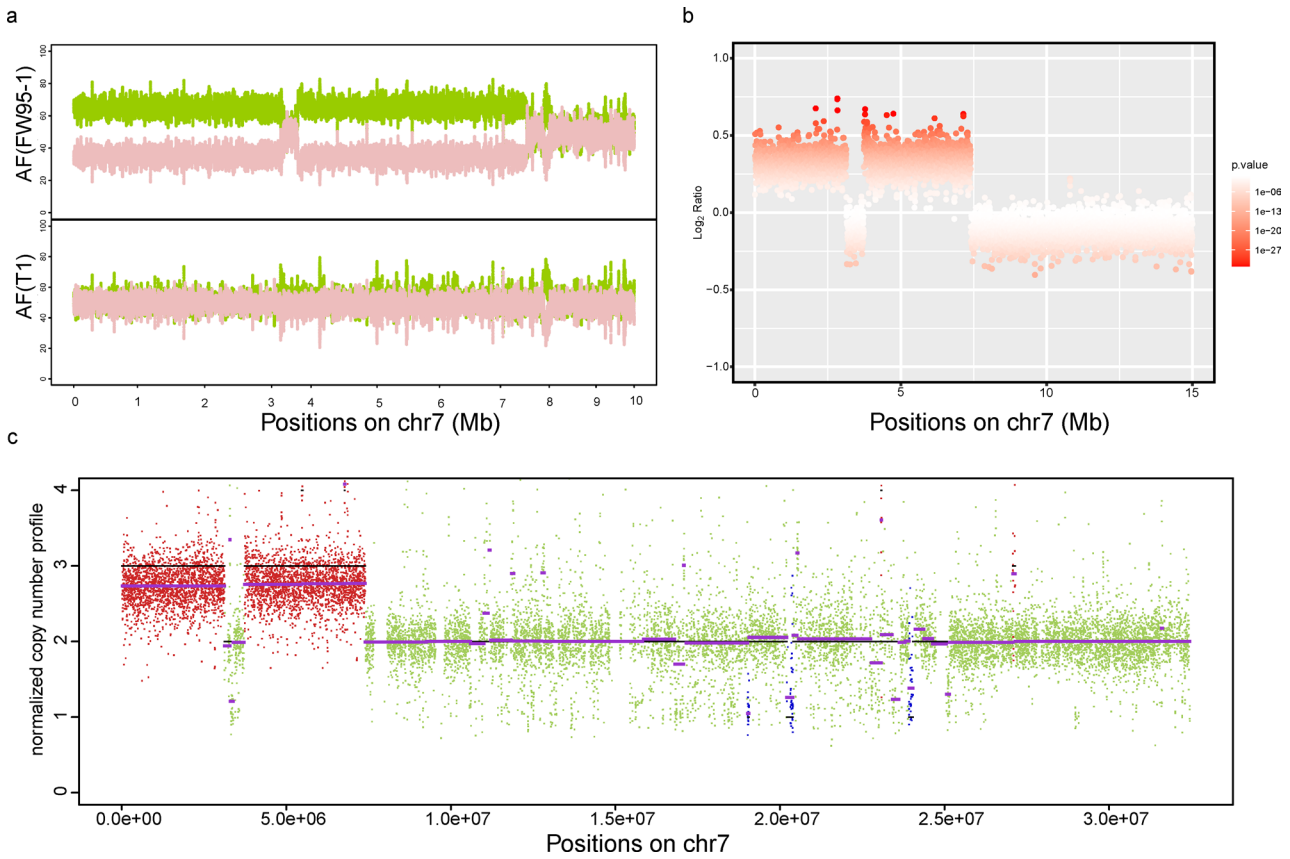
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



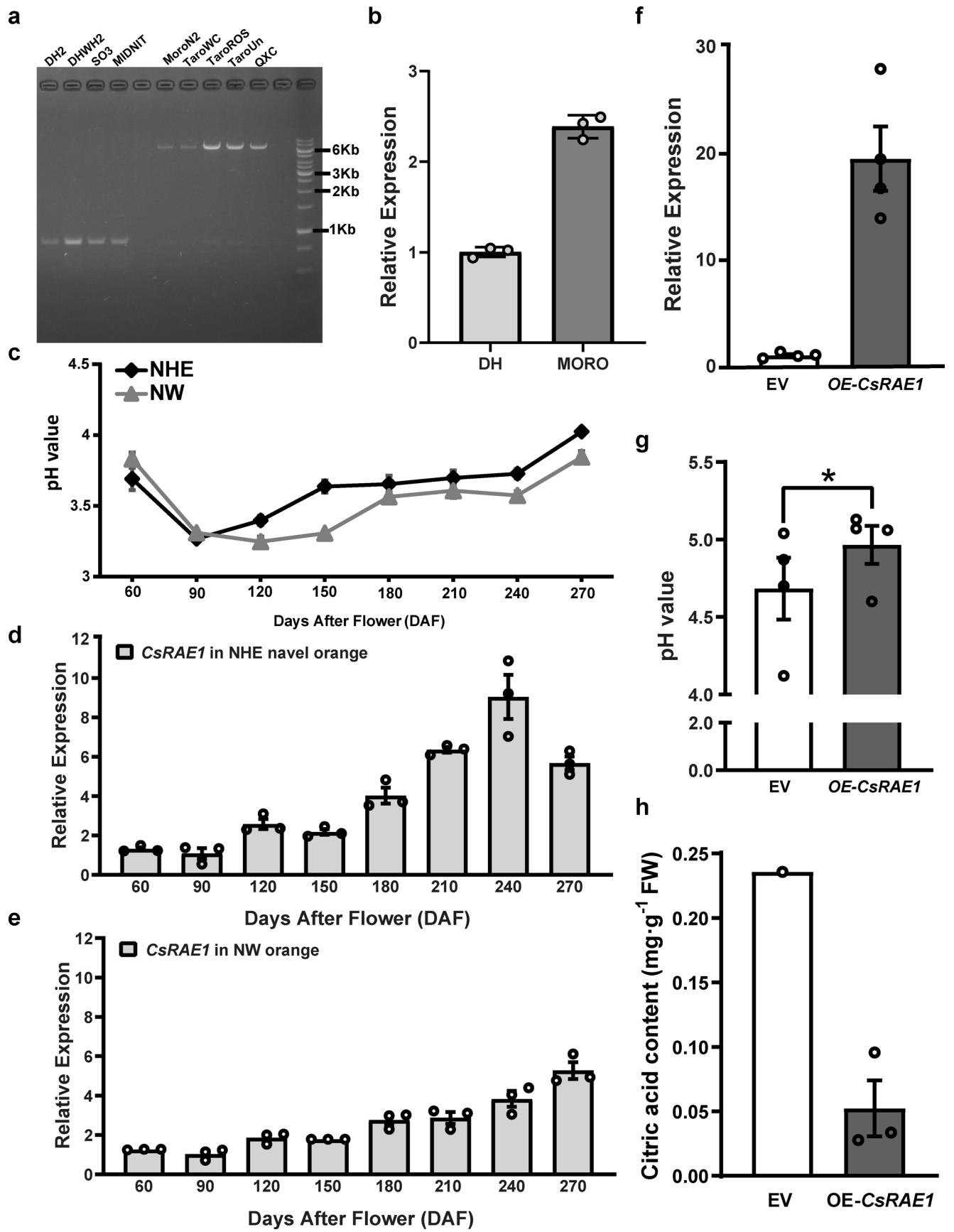
Extended Data Fig. 1 | Citrus bud mutation and asexual propagation. The mutation occurred somatically on a bud of one branch of the tree. If this mutation was observed by human, the mutated branch will be grafted on rootstock. Then this mutant was further propagated if developed as cultivars. The whole process is on somatic level.



Extended Data Fig. 2 | Validation of 12 TE insertions in low acid sweet orange (BTC) by PCR experiments. DH2, TCPS1, MORO, NHE, AJTC, SO3, ZAOJ are control sweet oranges; BTC, BT2, REN4, REN5, JH (accession name: JHBTC), HYJH (accession name: HYJHBTC), RRJH (accession name: RRJHBTC) are Bingtangcheng. The accession name was provided in the Supplementary Table 4 and the primers and reproducibility of gel validation experiments was provided in Supplementary Table 9.

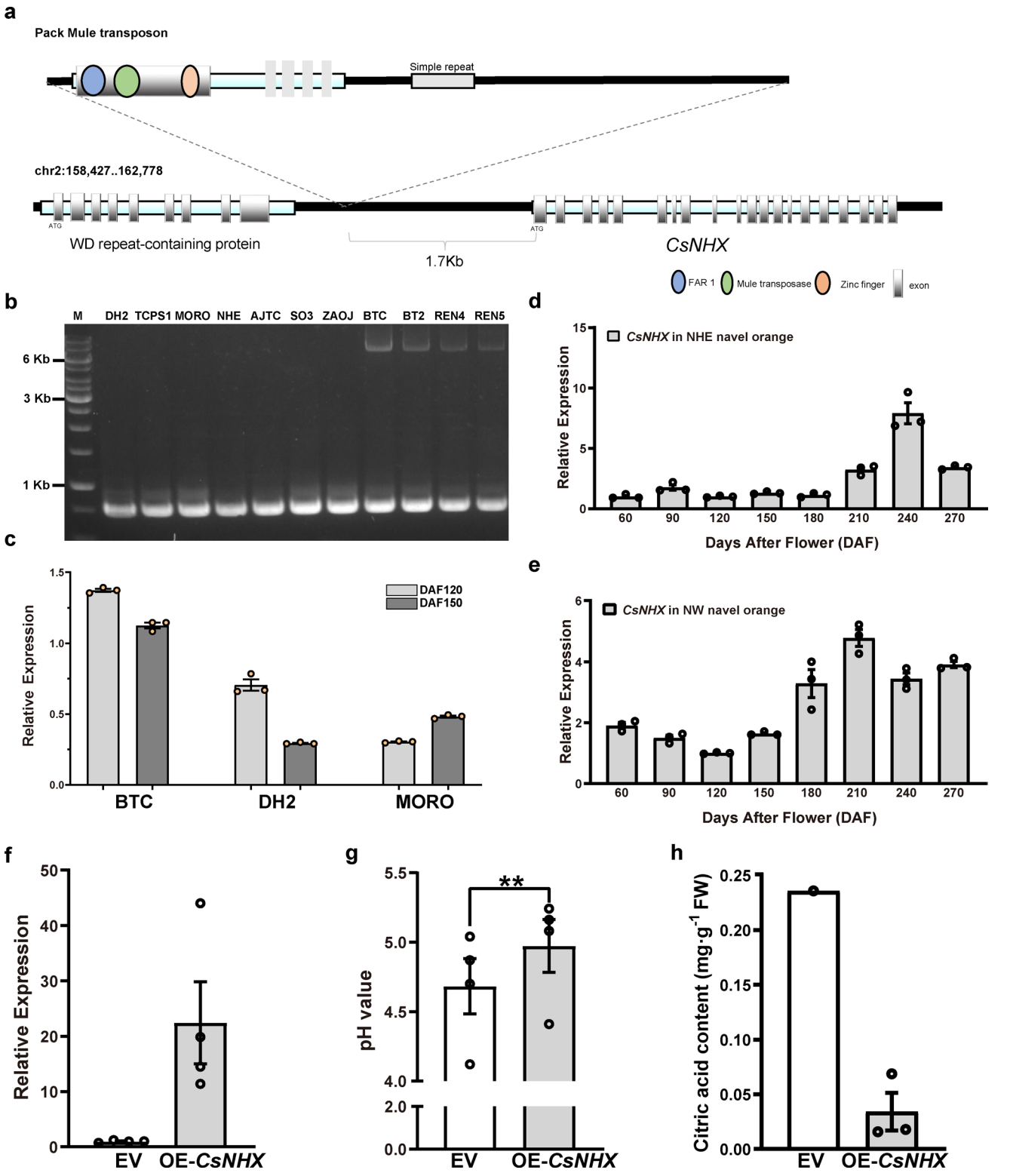


Extended Data Fig. 3 | Feature of the large duplication at 0..7.4M on chromosome 7. a. The allele frequency in the mutants FW95-1 and the control (T1) Statistical source data was provided. **b** Copy number ratios between FW95-1 and the control (T1). Windows in increasing red color tones with significance P values correspond to the signal of CNV. **c**. the copy number profile results of FREEC. Window with red represent the signal of copy number increase.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Validation the TE insertion in *CsRAE1* gene in the blood orange, transient gene transformation assay, and gene expression analysis of *CsRAE1* gene. a. DH2 and DHWH2 are high acid oranges; SO3 and MIDNIT are Valencia oranges; MoroN2, TaroWC, TaroROS, TaroUn and QXC are blood oranges. All the accession name was provided in the Supplementary Table 4. Nine independent experiments were repeated with similar results. Primer design information and experiments reproducibility was provided in Supplementary Table 9. **b**. Expression of *RAE1* in blood orange (XC), a moderate sweet orange and high acid sweet orange (DH, Dahong). Values are means ± S.E.M (n = 3 biological independent samples), **c**. the pH value in the fruit development of Newhall navel (NHE) and late maturing orange (NW). Values are means ± S.E.M (n = 3 biological independent samples), **d-e** Gene expression of the *RAE1* in the NHE (d) and NW (e). Values are means ± S.E.M (n = 3 biological independent samples), **f**. The expression of the *CsRAE1* gene in the overexpression (OE) lines and the control, **g**. the citric acid content in the OE lines and the control (EV), Values are means ± S.E.M (n = 4 biological independent samples), **h**. pH value in the OE lines of *CsRAE1* and EV, Values are means ± S.E.M (n = 4 biological independent samples). Asterisks indicate significant difference (*$p \leq 0.05$, P = 0.025, one-sided t-test,). All primer pairs were listed in Supplementary Tables 9 and 16.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Validation the TE insertion in promoter of *NHX* gene in the low acid orange (Bingtangcheng), transient gene transformation assay, and gene expression analysis of *CsNHX* gene. a. The structure of Mule transposon sequence and the *CsNHX* (Na^+/H^+ transporter) gene. **b**. PCR confirmation of the TE insertion. BTC, BT2, REN4, REN5 are low acid mutants (Bingtangcheng). DH2, TCPS1 are high acid oranges; Valencia (SO3) and blood orange (MORO) are moderate acid; AJTC is the acidless mutant. All the accession name was provided in the Supplementary Table 4. Seven independent experiments were repeated with similar results. Primer design information and experiments reproducibility was provided in Supplementary Table 9. **c**. Expression of *CsNHX* in different citrus varieties. Values are means ± S.E.M (n = 3 biological independent samples). XC means blood orange, a moderate sweet orange; DH (Dahong) is a high acid sweet orange. DAF, days after flowering. **d-e** Gene expression of the *NHX* in the Newhall navel orange (d) and Lanlate late-maturing orange (e). Values are means ± S.E.M (n = 3 biological independent samples). **f**. The expression of the *CsNHX* gene in the overexpression (OE) lines and the control (EV), Values are means ± S.E.M (n = 4 biological independent samples). **g**. the citric acid content in the OE lines and the EV, Values are means ± S.E.M (n = 3 biological independent samples). **h**. the pH value in the overexpression line of *CsNHX* and the EV, Values are means ± S.E.M (n = 4 biological independent samples). Asterisks indicate significant difference (**$p \leq 0.01$, P = 0.0093, one-sided t-test). All primer pairs were listed in Supplementary Tables 9 and 16.

nature research

Corresponding author(s): Qiang Xu

Last updated by author(s): Mar 16, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Raw sequencing data were generated from PacBio RSII, Oxford Nanopore, Illumina platform.
Data analysis	SMRT-analysis (version 2.1), Pilon (version 1.18), SSPACE-STANDARD (version 3.0), GapCloser (version 1.12), Necat, Nextpolish (version 1.0.4), NextDenovo (version 2.0-beta.1), 3D-DNA (version 180922), Juicebox (version 1.11.08), Recon (version 1.3.1), HapCUT2 (version 1.1), long ranger (version 2.2.22), extractHairs (version 1.1), BUSCO (version 3.0.2), RepeatModeler (version 1.0.11), Repeatmasker (version 4.0.7), LTR_FINDER (version 1.05), Augustus (version 2.4), GlimmerHMM (version 3.0.4), EVM (version 1.1.1), Exonerate (version 2.2.0), AAT (version 03052011), Trinity (version 2.9.1), GATK (version 3.8.1), BEDTools (version 2.13.3), snpEff (v4_3k), SAMTools (version 0.1.19), BWA (version 0.7.8-r455), R (version 3.5.3), Picard tools, GEC (version 0.2), vcftools (version 0.1.11), MEGA (version 7.0), ClustalW (version 2.1), NGMLR(version 0.2.7), Minimap2 (version 2.10), MUMmer (version 3.23), CNV-seq(version 0.2-8), FREEC(version 9.5), Lumpy(version 0.2.13), Pindel(version 0.2.4t), Perl(v5.16.3), bam-readcount(v.0.8), SIFT (version 2.0.0), LASTZ (1.03.54), CNS pipeline (https://github.com/liangpingping/CNSpipeline.git), phast (version 1.5), TreeBeST (version 1.9.2), Evolview (version 2), RaxmlHPC (version 8.0.0), EIGENSTRAT (version 4.2), iTOL (version 3), PLINK (version 1.90), FastQC (version 0.11.7), Trimmomatic (version 0.36), Circos (version 0.69-6), Sniffles, crossstitch (default), Ragoo (default), purge_haplotigs (default).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Genome data for di-haploid Citrus sinensis version 3.0 and version 4.0 have been deposited at DDBJ/ENA/GenBank under accession numbers MORK00000000 and JAFBAU000000000, respectively. The sequencing data and the Genome data for six diploid sweet oranges has been deposited at NCBI under accession PRJNA321100. The PacBio and Nanopore sequencing data for Citrus sinensis was deposited under the SRR accession number SRR5838837. Sweet orange genome sequences are also available from our website at http://citrus.hzau.edu.cn/orange.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	114 sweet orange somatic mutants was collected and used in our study, which can cover common varieties of sweet orange and represent the genetic polymorphism of the somatic mutants in sweet orange.
Data exclusions	No data was excluded in the analysis of somatic SNP, InDel, large deletion, Copy number variation. 10 samples was excluded in the large insertion analysis due to the low confidence of the results in our analysis.
Replication	In the fruit quality assay of the sweet orange landraces, two to six biologically independent replicates were used. In the gene function assay by transient overexpression, three or four biologically independent replicates were used for gene expression, pH measurement and citric acid measurement. In the gene expression experiments, three or four biological replications were used.
Randomization	In the identification of somatic variation, we divided the 114 sweet oranges into four random groups.
Blinding	The landrace samples were harvested blindly without access to the genotype identities.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging