DNA Research, 2016, 23(4), 403–414
doi: 10.1093/dnares/dsw017
Advance Access Publication Date: 21 April 2016
Full Paper

OXFORD

Full Paper

Identifying the genome-wide genetic variation between precocious trifoliate orange and its wild type and developing new markers for genetics research

Jin-Zhi Zhang, Sheng-Rui Liu, and Chun-Gen Hu*

Key Laboratory of Horticultural Plant Biology (Ministry of Education), College of Horticulture and Forestry Science, Huazhong Agricultural University, Wuhan 430070, China

*To whom correspondence should be addressed. Tel. +86 27-87281826. Fax. +86 27-87281826. E-mail: chungen@mail.hzau.edu.cn

Edited by Dr Satoshi Tabata

Received 13 December 2015; Accepted 21 March 2016

Abstract

To increase our understanding of the genes involved in flowering in citrus, we performed genome resequencing of an early flowering trifoliate orange mutant (*Poncirus trifoliata* L. Raf.) and its wild type. At the genome level, 3,932,628 single nucleotide polymorphisms (SNPs), 1,293,383 insertion/deletion polymorphisms (InDels), and 52,135 structural variations were identified between the mutant and its wild type based on the citrus reference genome. Based on integrative analysis of resequencing and transcriptome analysis, 233,998 SNPs and 75,836 InDels were also identified between the mutant and its wild type at the transcriptional level. Also, 272 citrus homologous flowering-time transcripts containing genetic variation were also identified. Gene Ontology and Kyoto Encyclopaedia of Genes and Genomes annotation revealed that the transcripts containing the mutant- and the wild-type-specific InDel were involved in diverse biological processes and molecular function. Among these transcripts, there were 131 transcripts that were expressed differently in the two genotypes. When 268 selected InDels were tested on 32 genotypes of the three genera of Rutaceae for the genetic diversity assessment, these InDel-based markers showed high transferability. This work provides important information that will allow a better understanding of the citrus genome and that will be helpful for dissecting the genetic basis of important traits in citrus.

Key words: citrus, flowering, genome resequencing, genetic variation, SNP

1. Introduction

Flower induction and initiation are the key developmental stages for flowering plants. Recently, extensive analyses performed in *Arabidopsis* have provided a complex picture of how these plants integrate environmental and endogenous signals to regulate the flowering transition.[1] Several flowering regulatory pathways function to promote or repress flowering depending on the environmental and endogenous conditions identified, such as photoperiod, vernalization, gibberellins, autonomous, and ageing pathways.[1,2] The balance of signals from these flowering pathways is integrated by a common set of genes, such as *FLOWERING LOCUS T* (*FT*), *FLOWERING LOCUS C* (*FLC*), *LEAFY* (*LFY*), and *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*), and could regulate the transition from the juvenile to the adult phase.[3,4] This knowledge has greatly accelerated research related to flowering in perennial plants.[5] However, when homologous genes from perennial plants were

analysed, differences regarding expression patterns and phenotypes were detected when compared with *Arabidopsis*, and these differences may reflect roles distinct from those described in model plants.[2,6] These studies indicated that the underlying molecular mechanism of flowering time may differ between perennial and model plants because of different flowering characteristics, such as long juvenile phase and seasonal flowering; some novel genes may also play a key role in these processes. Therefore, an understanding of these different characteristics requires identification and characterization of novel genes related to these characteristics in woody plants.

Citrus is one of the most economically important evergreen crops for the production of fresh fruit and juice; its economic and social impact on our society is tremendous.[7–9] However, the flowering transition of most citrus plants is 6–8 yr, and for some species such as seedling oranges and grapefruit it is 8–10 yr.[8] However, commercially important traits of citrus are expressed primarily in fruit tissue. This requires the trees to be capable of flowering and producing fruit to evaluate functions of these genes. Precocious trifoliate orange, a spontaneous mutant with a short juvenile phase from WT trifoliate orange (*Poncirus trifoliata* L. Raf.) was found in Yichang, Hubei Province, China.[10] Approximately 20% of the seedlings from the mutant seeds first flowered during the year after the spring planting; the juvenile period of the mutant has been greatly reduced to 1– 2 yr when compared with the WT plants, which have a juvenile period of 6– 8 yr.[10,11] Precocious trifoliate orange and its wild type have nearly the same morphology, except for the flowering habit, and no DNA polymorphism has been detected between them based on some molecular markers.[10–12] Therefore, precocious trifoliate orange is speculated to be a direct variant of the wild type. Previous transcriptional studies including suppression subtraction hybridization combined with macroarray and massively parallel signature sequencing were used to investigate transcriptome changes between the mutant (MT) and its wild type (WT), and several differentially expressed genes were detected.[11,13] However, it remains largely unknown what kind of mechanism results in early flowering.

The recent release of the citrus genome[9,14] and next-generation DNA sequencing technology[15] will dramatically enhance the efficiency of functional and comparative genomics research in citrus. The alignment of the short reads obtained from the MT and the WT to the citrus reference genome has provided the perfect opportunity to identify a large number of genetic variations between the MT and the WT, including single-nucleotide polymorphisms (SNPs), insertion and deletion polymorphisms (InDels), and structural variations (SVs). Here, we sequenced the genome of the MT and the WT; several putative genetic variations including SNPs, InDels, and SVs were identified. The frequency and distribution of these genetic variations in different regions of gene were also identified. In addition, we developed a great deal of effective DNA markers for exploring genetic diversity in the citrus genus.

2. Methods

2.1. Plant materials

All samples from 32 citrus species (fortunella; kumquat; trifoliate orange; precocious trifoliate orange; flying dragon trifoliate orange; citranges; Fenghuang pummelo; Kaopan pummelo; grapefruit; Hirado Buntan pummelo; guoqing no. 1 satsuma; Qingjiang Ponkan; Bendiguangju mandarin; Miyagawa wase satsuma; red tangerine; calamondin; Clementine mandarin; Newhall navel orange; Valencia orange; Jingcheng sweet orange; Honganliu sweet orange; sour orange; Cara Cara navel orange; Washington navel orange; blood orange; Ichang papeda; papeda; citron; bergamot; Mexican lime; lemon; and rough lemon) were collected in the experiment fields of the National Citrus Breeding Center at Huazhong Agricultural University. To identify flowering-related genetic variation at the transcriptional level, the terminal bud and the five following buds (the major node position for flower formation) from spring flushes of the MT and the WT 2-yr-old trees were also collected at flower initiation (after self-pruning). Previous cytological studies revealed that the floral buds of spring shoots in 2-yr-old precocious trifoliate orange initiated differentiation immediately after self-pruning, but its WT did not.[16] To eliminate the influence of the genetic background, the materials used for genome re-sequencing and RNA-seq from the same 2-yr-old trees in this study. All materials for RNA and DNA extraction were collected from three individual plants. All plant tissues were sampled according to the demands of each experiment, immediately frozen in liquid nitrogen, and stored at −80°C until used.

2.2. DNA isolation and genome sequencing

Plant DNA of all material was isolated from the leaf according to the cetyltrimethyl ammonium bromide method.[17] For genome resequencing, the DNA from the MT and the WT were randomly sheared. After electrophoresis, DNA fragments of the desired length were gel-purified. Adaptor ligation and DNA cluster preparation were performed and subjected to Solexa sequencing using Illumina Genome Analyzer II.[18] Low-quality reads (<20), reads with an adaptor sequence, and duplicated reads were filtered, and the remaining high-quality data were used for mapping. Raw sequence data obtained in our study have been deposited in the NCBI Short Read Archive with accession number SRP070975.

2.3. Detection of SNPs, InDels, and SVs

The sequencing reads from the MT and the WT were aligned to the Clementine genome (http://www.phytozome.net/clementine.php) separately using Burrows–Wheeler Aligner (BWA) software.[19] Reads that aligned to more than one position of the reference genome were filtered and used for determining reads mapping to multiple positions in the reference and unmapped reads. Average sequencing depth and coverage were calculated using the alignment results.[20] The mapped reads were then used to detect SNPs, InDels, and SVs using SOAPsnp and SOAPsv software with default parameter settings.[20–22] The SNPs identified were filtered based on the following stringent criteria: no less than two times for coverage depth (no less than three times in the heterozygous locus), no more than three times of average depth, distance of adjacent variation had to be >5 bp, and target mapping quality had to be >20. For the InDels, gaps supported by at least three pair-end (PE) reads were retained. For obtaining reliable SVs, the detected SVs must be returned to the PE alignments between resequencing genome and the reference genome and be validated under the following criteria: 2× to 100× for coverage depth and >20 for SVs quality.

2.4. Annotation of SNPs, InDels, and SVs

Localization of the SNPs, InDels, and SVs was based on the annotation of gene models of the Clementine genome (http://www.phytozome.net/clementine.php). The three types of polymorphisms in the gene region and other genome regions were annotated as genic and intergenic, respectively. The genic SNPs, InDels, and SVs were classified as CDS, UTR, and intron according to their localization. The SNPs in the CDS region were further separated into synonymous and non-synonymous amino substitutions using

Genewise.[23] The Gene Ontology (GO)/PFAM annotation data were further used to functionally annotate each gene including non-synonymous SNPs with 1- to 10-bp lengths.

2.5. RNA-Seq data analyses

To prepare a representative sample of total RNA, we pooled three individual plants from approximately equal numbers of the MT and the WT materials, respectively. Total RNA was extracted according to a previous protocol.[24] Library construction and sequencing were performed as described by Zenoni et al.[25] Low-quality reads (<20), reads with adaptor sequence, and duplicated reads were filtered, and the remaining high-quality data were used in the mapping. The PE sequencing reads were aligned to the citrus reference genome sequence separately using BWA software algorithm under the default parameters as described for genome sequencing. To identify flowering-related genetic variation in the two genotypes, we mapped the reads against the Clementine genome[9,14] using the SOAP2 software with default parameter settings as described for genome sequencing.[22] These identified SNPs and InDels from RNA-Seq were annotated the genic genetic variation by identified of genome sequencing. A candidate genetic variation was identified by both RNA-Seq and genomic DNA reads, for which the available flanking sequence matched 100% over the entire length at a single location. It will be considered a real variation at the transcriptional level. The RNA-Seq data from this study have been submitted to Gene Expression Omnibus under accession number GSE78810.

2.6. Functional assignments of the transcripts containing specific InDels

To assign putative functions to the MT and the WT containing specific InDel transcripts, Blast2Go program was run to BLAST against a reference database that stores UniProt entries, GO, Enzyme Commission (EC), and Kyoto Encyclopaedia of Genes and Genomes (KEGG) annotation.[26] The GO categorization results were expressed as three independent hierarchies for biological process, cellular component, and molecular function.[26]

2.7. New marker screening and polymorphism survey

For Sanger sequencing, primers spanning genomic regions predicted to contain InDels were used to amplify genomic DNA templates from the Clementine, the MT, and the WT. PCR amplification was conducted in 25-μl reactions containing 50 ng template DNA, 2.5 μM $MgCl_2$, 2.5 μl 10× PCR buffer, 0.5 mM of each primer, 0.5 U Taq DNA polymerase, and 2.5 mM dNTPs. The PCR cycling profile was 94°C for 5 min, 35 cycles at 94°C for 45 s, 60°C for 45 s, 72°C for 45 s, and a final extension at 72°C for 10 min. Target sequences were recovered and sequenced. For new marker screening, the quality of the PCR product from 32 citrus species was checked by a 6% polyacrylamide sequencing gel containing 7 M urea in 0.5× TBE buffer. Three microliters of PCR product was mixed to an equal volume of loading buffer containing 95% formamide, 0.25% bromophenol blue, 0.25% xylen cyanol, and 10 mM EDTA. This mixture was heated for 5 min at 94°C to denature the DNA before loading. Gels were stained with silver nitrate following the protocol described[27] for gel electrophoresis analysis and for comparison with the 10-bp DNA standard ladders (Invitrogen).

2.8. Genetic diversity and data analysis

POPgene (v1.32) was used to calculate the different statistical and genetic parameters,[28] such as the effective allele number,[29] Shannon's information index,[30] and expected heterozygosity.[31] The genetic distance of the InDel genotype was calculated based on Nei's genetic distance measure, with pairwise distance calculated by MEGA 4.[32] The polymorphic information content (PIC) value for the InDel marker was calculated thereafter.[33] A dendrogram was constructed based on the unweighted pair-group method, with arithmetic mean determined by MEGA4.

3. Results

3.1. Genome sequencing and mapping onto the Clementine reference genome

To identify the genome-wide genetic variation between precocious trifoliate orange and its WT, two resequencing libraries from the two genotypes of trifoliate orange were constructed. The genome size of the Clementine reference genome is 301.4 Mb.[9] We estimated that approximately 40-fold genome coverage should be sufficient for aligning most of the sequences. Therefore, 12.22 (143.78 million reads) and 13.16 Gb (159.56 million reads) of sequencing data were generated from the MT and the WT, respectively. Approximately 11.79 Gb (138.68 million clean reads) of filtered sequencing data for the MT and 12.64 Gb (153.23 million clean reads) of data for the WT were aligned to the Clementine genome using BWA software (Supplementary Table S1). The results showed that the genomic GC contents of the MT and the WT were 36.46 and 36.04%, respectively. A summary of the resequencing data is presented in Supplementary Table S1. For the MT and the WT, 61.73 and 63.13% of clean reads were uniquely mapped to the reference genome and translated into 24- and 26-fold effective coverage of the reference sequences, respectively (Supplementary Table S1). These clean reads of the MT and the WT covered approximately 81.16 and 82.30% of the reference sequence, respectively, indicating that the generated data set was highly relevant to the reference genome (Supplementary Table S1). Among citrus 33,929 transcripts, 6.95% and 7.01% reads of sequencing data were mapped 23,543 (69.38%) and 23,821 (70.21%) transcripts from the MT and the WT, respectively.

3.2. Detection and characteristics of genetic variations of the MT and the WT

Genome-wide SNPs and InDels between the MT and the WT were identified by using SOAP on the basis of comparisons with the Clementine genome. In the MT and the WT, 3,237,862 and 3,422,013 SNPs were detected, respectively. Obviously, the number of detected SNPs in the latter comparison was more than that in the former because of the high fold coverage for the reference sequence. The SNP detected between two genomes and the reference genome was classified as transition (G/A and C/T) or transversion (A/C, C/G, G/T, and T/A) based on nucleotide substitution (Fig. 1A). Both of the transition proportions were significantly higher than the transversion proportions in two genotypes (Fig. 1A). Among these transitions, the G/A proportion was slightly more than that of C/T; among the transversions, the T/A proportion was the most and the G/C proportion was the least (Fig. 1A). The transition-to-transversion ratios were 1.468 and 1.475 in the MT and the WT, respectively.

The MT and the WT yielded 867,854 and 895,643 InDels (including insertion/deletion), respectively. Among these InDels, 412,922 (47.58%) and 426,266 (47.59%) were insertions in the MT and the WT in comparison with the reference sequence, respectively (Fig. 1B). The length of insertion ranged from 1 to 32 bp, whereas that of deletion was up to 46 bp between the MT and the reference

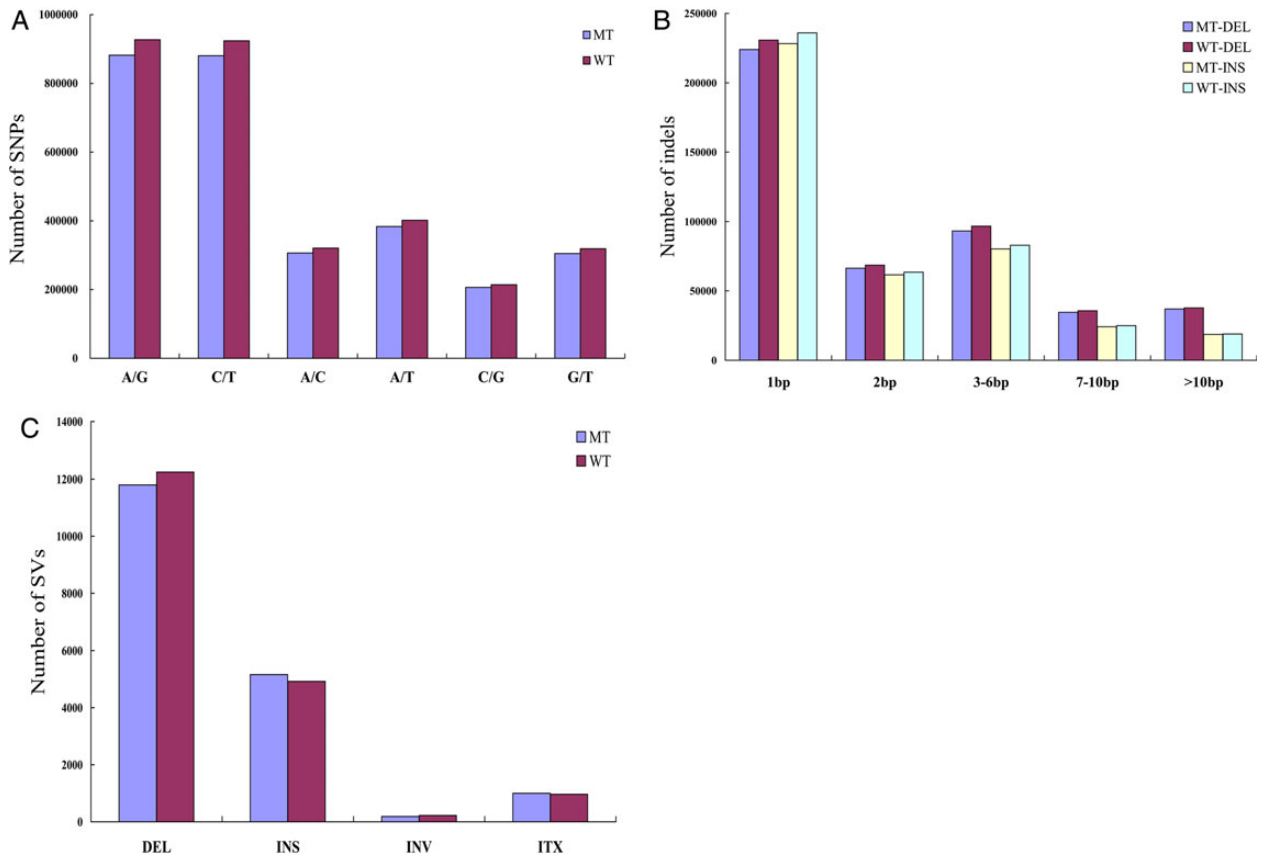

Figure 1. Annotation analysis of SNPs, InDels, and SVs. (A) Frequency of different substitution types in the identified SNPs in MT/Clementine (precocious trifoliate orange compared with the *C. clementina* reference genome) and WT/Clementine (wild-type trifoliate orange compared with the *C. clementina* reference genome). (B) Length distribution of InDels in MT/Clementine and WT/Clementine. (C) The number of different types of SVs. DEL, deletion; INS, insertion; INV, inversion; ITX, translocation. This figure is available in black and white in print and in colour at *DNA Research* online.

(Fig. 1B). More than half of the InDels (52.10%) were mononucleotides, 14.74% were dinucleotides, 19.98% were 3- to 6-bp nucleotides, 6.78% were 7- to 10-bp nucleotides, and 6.40% were nucleotides >10 bp. Interestingly, the WT and the reference genome (Fig. 1B) showed the same trend. Meanwhile, 27,257 and 29,066 SVs were also found in the MT and the WT compared with the citrus reference genome, respectively. A comparative analysis showed that a total of 52,135 were identified between the MT and the WT; 4,188 were common and 47,947 were specific to the two genotypes. The types of SVs were deletion (75.86%), insertion (17.24%), translocation (5.38%), and inversion (1.52%) in the MT (Fig. 1C). In WT, the distribution trends were similar to those of MT. We also detected 16,487 SVs in two genotypes; 1,168 were present in both the MT and the WT, 7,600 were specific to the MT, and the other 7,719 were specific to the WT. The homozygous and heterozygous ratios were 4.12:1 for InDels and 2.22:1 for SVs in the MT, respectively. For the WT, the homozygous and heterozygous ratios were 4.00:1 and 2.12:1, respectively. The average densities of detected genetic variations between the two genotype genomes and the Clementine reference genome were 10,038.5 and 10,521.5 SNPs/Mb, 2,940.2 and 3,024.3 InDels/Mb, and 92.3 and 98.5 SVs/Mb in the MT and the WT, respectively.

3.3. Analysis of genetic variation differences between the MT and the WT

We compared the SNPs between the MT and the WT based on Clementine reference genome, and the result showed that a total of 3,932,628 were identified between the MT and the WT, 2,727,248 were common and 1,117,089 were specific to the two genotypes at the genome level. To further evaluate the differential SNPs between the MT and the WT in the genic region, the data from the two genotypes were combined, and we obtained the following: a total of 1,897,842 SNPs located in the genic region; 1,589,456 were found in both the MT and the WT, 141,195 were specific to the MT distributed in 20,023 transcripts, and 167,182 were specific to the WT distributed in 21,684 transcripts. Among these genic SNPs, there were 1,422,343 homozygous and 475,499 hemizygous SNPs. The positions of specific SNP were identified in CDS, intron, 5′ UTR, and 3′ UTR regions according to the reference genome. Among 308,377 specific SNPs, 51,589 (MT: 23,169 and WT: 28,419) were located in UTR regions, 169,373 (MT: 76,885 and WT: 92,488) in intron regions, and 87,416 (MT: 41,141 and WT: 46,275) in CDS regions. Non-synonymous coding SNPs are believed to have the higher impact on phenotype.[34] We therefore paid attention to specific non-synonymous SNPs in CDS regions in two genotypes. The number of non-synonymous SNPs was 25,769 (29.14%) and 29,018 (27.68%) in CDS regions of the MT and the WT, respectively. Such information will greatly enhance our understanding of the flowering process in the MT, as these are more likely to have phenotypic effects.

A comparative analysis showed that a total of 1,293,383 InDels were identified between the MT and the WT; 649,514 were common and 397,741 were specific to the two genotypes. When increasing or decreasing nucleotides in a coding sequence, it's possible to produce a frameshift mutation followed by character variation.[35] Therefore, the

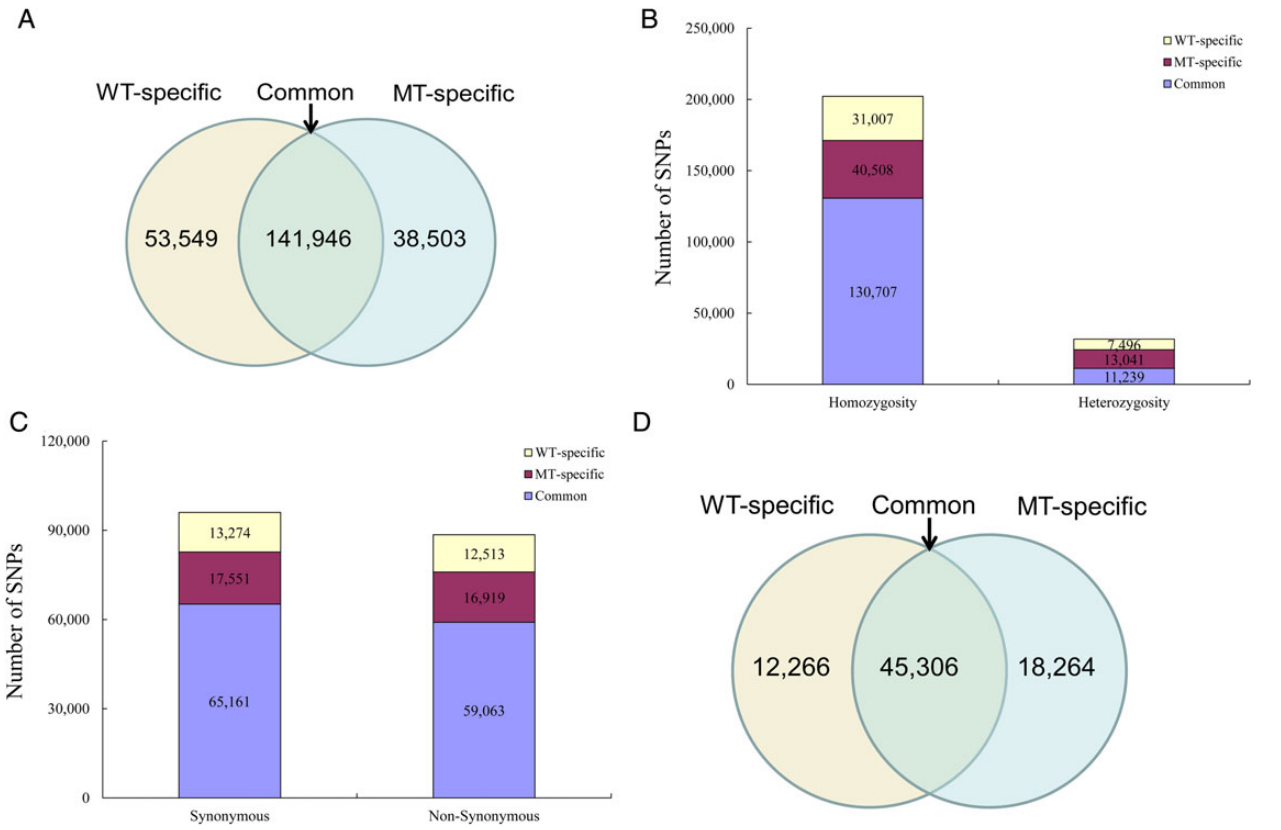

Figure 2. SNPs and InDels from transcriptional level distribution. (A) WT-specific SNPs (these SNPs are present only in the wild-type trifoliate orange genome) and MT-specific SNPs (these SNPs are present only in precocious trifoliate orange genome). The remaining SNPs are present in both genotypes. (B) Numbers of homozygous and heterozygous SNPs. (C) Numbers of synonymous and non-synonymous SNPs. (D) WT-specific InDels (these InDels are present only in the wild-type trifoliate orange genome) and MT-specific InDels (these InDels are present only in precocious trifoliate orange genome). The remaining InDels are present in both genotypes. This figure is available in black and white in print and in colour at *DNA Research* online.

InDels of genic regions were also identified in this study. After being carefully filtered, 429,301 InDels were obtained in genic regions in the two genotypes; 267,756 were found in both the MT and the WT distributed in 23,488 transcripts, 63,119 were specific to the MT distributed in 10,092 transcripts, and 98,426 were specific to the WT distributed in 10,518 transcripts. Among 161,545 specific InDels, 37,720 (MT: 14,945 and WT: 22,775) were located in UTR regions, 114,118 (MT: 44,578 and WT: 69,540) in intron regions, and 9,707 (MT: 3,596 and WT: 6,111) in CDS regions. Interestingly, the number of specific genetic variations (including SNPs and InDels) from the MT is less than the MT in this study. Within the gene body, the maximum number of specific SNPs (MT: 54.45% and WT: 55.30%) and InDels (MT: 70.62% and WT: 70.65%) was present in the introns of two genotypes. Only a small part of specific SNPs (MT: 29.14% and WT: 27.68%) and InDels (MT: 5.69% and WT: 6.21%) occurred in the CDS region.

3.4. Transcriptome analysis of flowering-related genetic variation in two genotypes

To reduce the identification of false-positive genetic variations at the transcriptional level, an integrative analysis of RNA-Seq and genome resequencing was performed for the two genotypes at the flowering transition stage (see Section 2.1). For the MT and the WT, 3.94 and 3.68 Gb of clean data were obtained, respectively. A total of 80.88% and 80.36% successful reads were perfectly matched to the Clementine genome for the MT and the WT, respectively. A summary of the resequencing data is presented in Supplementary Table S2. A total of 195,495 and 180,449 SNPs were identified in the MT and the WT distributed in 24,849 and 24,328 transcripts, respectively. Finally, a total of 233,998 SNPs were identified between the MT and the WT based on integrative analysis of RNA-Seq and genome resequencing. Of these, 53,549 were only present in the MT distributed in 20,366 transcripts, and 38,503 were only observed in the WT distributed in 18,388 transcripts (Fig. 2A). There were 202,222 homozygous SNPs and 95,986 synonymous SNPs among the total SNPs (Fig. 2B and C). We also found 333 SNPs that might lead to premature starting or stopping of transcription (220 starts gained and 113 stops gained), 7 SNPs related to intron splicing, and 22 mutations in start/stop codons (12 in start codons and 10 in stop codons) between the MT and the WT. Compared with the Clementine genome, 75,836 InDels were uncovered in the two genotypes; 18,264 were only present in the MT distributed in 10,996 transcripts, and 12,266 were only observed in the WT distributed in 8,164 transcripts (Fig. 2D).

For the MT, annotation analysis showed that 47,130 SNPs and 20,791 InDels were located in the intron regions based on the annotation of the reference gene models during the phase change stage. Meanwhile, 40,416 SNPs and 19,011 InDels were also located in the intron regions in the WT. These results indicated that alternative splicing and novel transcriptional events may have occurred during the gene expression process of the two genotypes. In the MT and the WT, InDels with a length of 1 bp accounted for >50% of the

whole genes. Most of the InDels in the coding sequence (CDS) regions were trinucleotides or hexanucleotides, which could not have been caused by frameshifts. However, mononucleotides were always the most common nucleotides in intergenic regions. Despite the minimally abundant distribution within critical sites, such as the CDS region and untranslated region (UTR) (12.1% of total InDels), these InDels can alter phenotypes through a variety of mechanisms.

3.5. Identification of genetic variation in flowering-related genes

A total of 244 transcripts representing putative homologs to flowering-related genes were identified in Clementine genome (http://www.phytozome.net/clementine.php) by BLAST searches against the *Arabidopsis* Information Resource (TAIR) and NCBI data sets, and a total of 178 read-mapped flowering-related genes were observed in two genotypes in this study (Supplementary Table S2). Annotation analysis showed that 1,073 SNPs, 382 InDels, and 50 SVs were located in 147, 97, and 29 transcripts (including UTR, CDS region, and intron region) related to flower induction, flower development, and flowering time, respectively (Supplementary Table S3). These genetic variations may directly represent functional changes in the gene products. Some of these genes were required for the photoperiod pathway and some encode regulatory proteins specifically involved in the control of flowering by vernalization pathway, whereas others encoded components of ageing pathways or were involved in the circadian clock function (Supplementary Table S2). These included *AGAMOUS*, *APETALA*, and *SEPALLATA* genes for flower development; *FRIGIDA*, *FRIGIDA INTERACTING PROTEIN1* gene, and *VERNALIZATION3/5-like* for the vernalization pathway; and *SQUAMOSA promoter binding protein–like* gene, GIGANTEA protein, and flowering-time control protein-related/FCA gamma-related for age and circadian clock developmental processes (Supplementary Table S2). Moreover, some additional flowering-time regulators containing genetic variations that have not been placed in any specific pathway were also identified, such as *EMBRYONIC FLOWER1* (*EMF1*), *RELATIVE OF EARLY FLOWERING 6*, *FLOWERING PROMOTING FACTOR 1*, and *RECEPTOR-LIKE KINASE IN FLOWERS 1* (Supplementary Table S2). We also found several genes encoding putative photoreceptor proteins, including phytochrome, phytochrome interacting factor, phytochrome and flowering-time regulatory protein, phytochrome-associated protein, and cryptochrome. These variations may provide important genetic variations to explain the phenotypic differences between the MT and the WT (Supplementary Table S2).

3.6. Functional clustering and expression analysis of the transcripts containing specific InDels from the flowering phase change stage

InDels that occur in functionally important regions of genes (typically CDS region) could be seen affecting gene function by frameshifts and structural changes of protein. Annotation analysis showed that the MT- and WT-specific InDels from the phase change stage were located in the CDS region of 1,137 and 721 transcripts, respectively. To explore and summarize the functional categories of these transcripts, we used Blast2GO to obtain the GO terms for the representation of molecular function, cellular component, and biological process. Approximately 902 (79.4%) and 588 (80.5%) of the MT and the WT transcripts had BLAST hits, respectively. Fig. 3 shows the transcript distribution for three main categories. Based on molecular function, these genes were finally classified into 12 categories; the three most over-represented GO terms were binding, hydrolase activity, and transporter activity (Fig. 3A). Categories based on biological processes revealed that the mutant genes were related to 14 biological processes (Fig. 3B); the three most frequent terms were metabolic process, cellular process, and single-organism process, suggesting that these mutative genes were involved in a broad range of physiological functions (Fig. 3B). As shown in Fig. 3C in cellular component category, the analysis revealed a high percentage of cells and organelles in cellular components. The biological interpretation of these transcripts was further examined using KEGG pathway analysis. In the MT and the WT, 84 and 71 different pathways were found, respectively. Those pathways mainly correlated with development involved with metabolism, hormone signal transduction, and transcriptional regulation, suggesting that some mutative genes may affect the distinguishing traits of the MT and the WT.

Using all of our sequence reads, the expression levels of 1,858 transcripts containing specific InDels were estimated during the phase change stage. There were 131 transcripts that were differently expressed between the MT and the WT, with $P \leq 0.005$ and the absolute value of $|\log_2 \text{ratio}| \geq 0.5$ used as the threshold. Of these, 97 were more abundant and 34 were less abundant in the MT compared with the WT, suggesting that many genes were enriched during the flowering transition (Supplementary Table S4). BLAST searches of the 131 transcripts showed that 102 transcripts (77.9%) were homologous to known genes, 9 transcripts (6.9%) were homologous to genes of unknown function, and 20 clones (15.3%) had no matches in the database. Among 102 homologous to known genes, many had high identity with known transcription and post-transcriptional regulatory genes, indicating that these genes may be key regulators controlling flower development by activating or repressing numerous genes. Moreover, some transcription factors, including MYB, MADS, and Zinc finger, were observed (Supplementary Table S3). In addition to transcription factors, differently expressed genes involving chromatin remodeling, hormone regulation, and other metabolic pathways were also observed.

3.7. Development of new DNA markers in citrus

To test cross-species/genera transferability, 268 newly identified InDel markers were tested on a panel of 32 diverse citrus accessions (Fig. 4). Of these primer pairs, 35 primer pairs cannot amplify any fragment, suggesting that these primers were not well designed; 72 pairs of primers amplify numerous non-target bands, suggesting that these primers were also problematic; and 87 pairs of primers did not exhibit any difference. Only 74 InDels were confirmed. The results indicated that a high level of genetic diversity in the vicinity of these InDels was discovered in 32 genotypes. The number of alleles per locus ranges from 2 to 10, with an average of 3.92 alleles. Ne ranges from 1.06 to 7.04, with an average of 2.34. Ho ranges from 0.0 to 0.97, with an average of 0.34. He ranges from 0.06 to 0.87, with an average of 0.51. PIC ranged from 0.06 for the Ciclev10029752m locus to 0.48 for the Ciclev10010584m locus (Supplementary Table S5). The mean PIC value for citrus genera was 0.45 (range, 0.40–0.47), which was higher than 0.43 and 0.44 observed for trifoliate orange genera (range, 0.43–0.48) and fortunella genera, respectively (Table 1). The aforementioned InDel allelic data, when used to calculate the heterozygosity estimates, revealed highly significant differences between the observed and expected heterozygosity for citrus (mean Ho: 0.35; mean He: 0.43) and fortunella genera (mean Ho: 0.22; mean He: 0.23). Thus, the results suggest significant heterozygote deficiency among the three genera of Rutaceae (Supplementary Table S4). In addition,

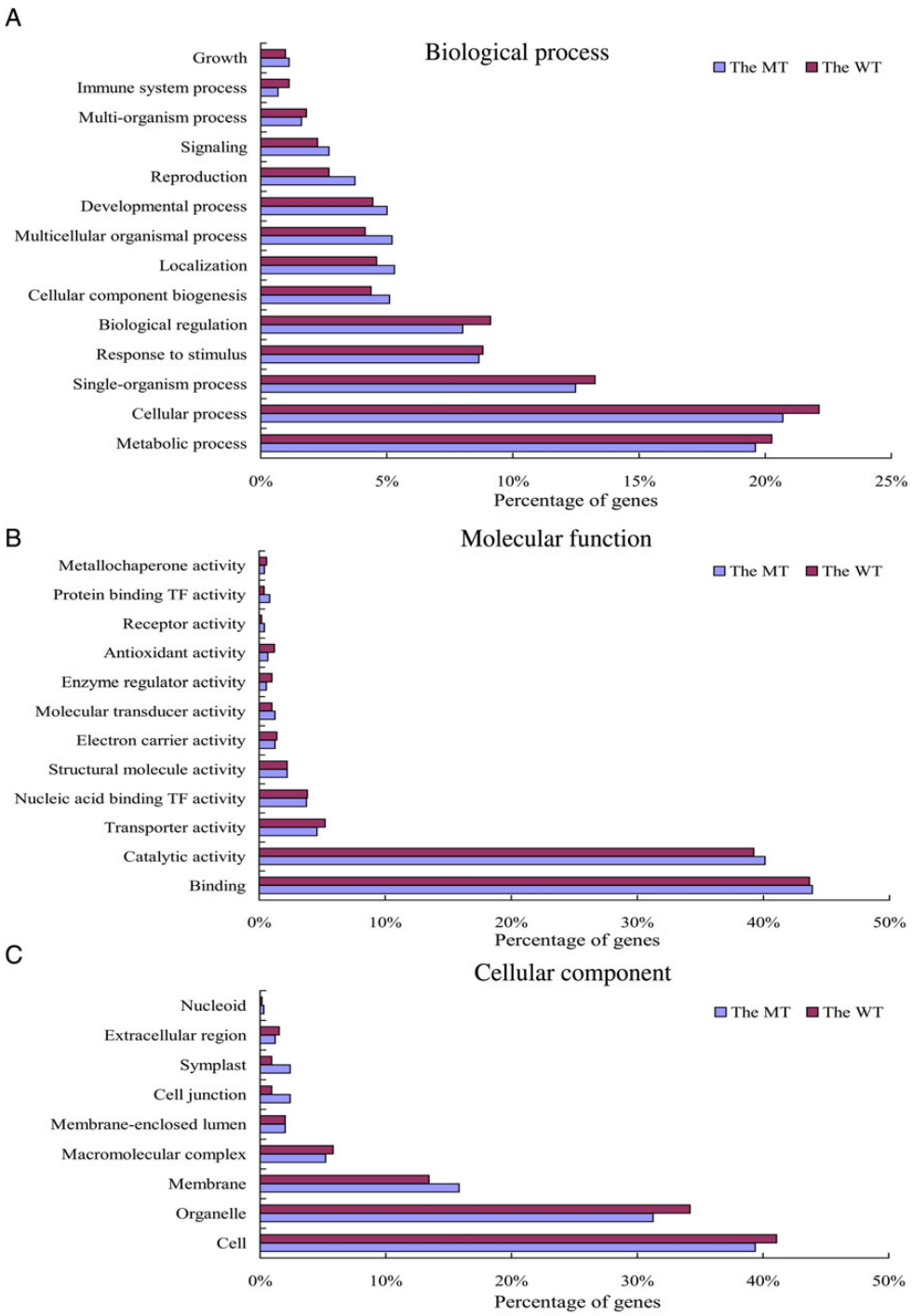

Figure 3. Functional categorization of the transcripts containing InDels. These genes were categorized based on GO annotation, and the proportion of each category is displayed based on (A) cellular component, (B) molecular function, or (C) biological process. This figure is available in black and white in print and in colour at *DNA Research* online.

it is noteworthy that some band patterns from the above markers were almost the same between the MT and flying dragon trifoliate orange (Fig. 4). However, precocious trifoliate orange cannot be derived from flying dragon trifoliate orange based on their huge morphological differences and previous studies.[10,12,13] There might be one possible explanation. Because precocious trifoliate orange and its WT have nearly the same morphology other than flowering regulation, these InDel markers from the MT and the WT may be related to flowering. Furthermore, flying dragon trifoliate orange shows a certain degree of early flowering compared with the mutant's WT.

3.8. Genetic diversity and dendrogram in citrus genus by new DNA markers

Despite a relatively low level of polymorphism, the 74 new markers from InDels were also examined for their potential use in genetic

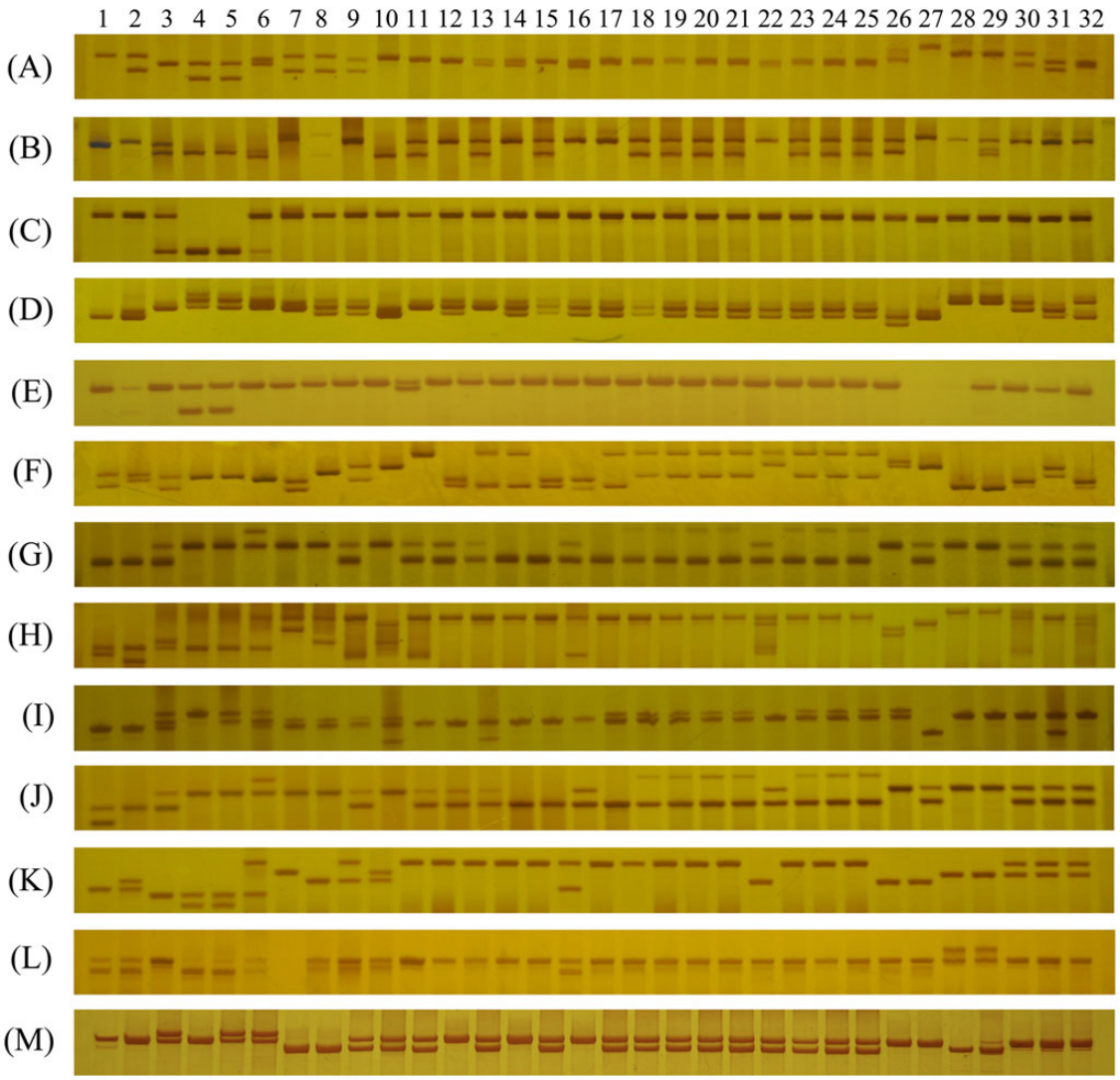

Figure 4. Polymorphism detected by 13 InDel markers among 32 genotypes of three genera of Rutaceae. 1: fortunella (*F. japonica*); 2: kumquat (*F. hindsii* var. *chintou* Swing); 3: trifoliate orange (*P. trifoliata* [L.] Raf.); 4: precocious trifoliate orange (*P. trifoliata* [L.] Raf.); 5: flying dragon trifoliate orange (*P. trifoliata Raf.* var. *monstrosa*); 6: cotranges (*P. trifoliata Raf* ´ *C. sinensis*); 7: Fenghuang pummelo (*C. grandis*); 8: Kaopan pummelo (*C. grandis*); 9: grapefruit (*C. grandis*); 10: Hirado Buntan pummelo (*C. grandis*); 11: Guoqing no. 1 satsuma (*C. unshiu*); 12: Qingjiang ponkan (*C. reticulata*); 13: Bendiguangju mandarin (*C. reticulata* cv. *Succosa*); 14: Miyagawa wase Satsuma (*C. unshiu*); 15: red tangerine (*C. reticulata*); 16: calamondin (*C. mitis*); 17: Clementine mandarin (*C. clementina*); 18: Newhall navel Orange (*C. sinensis*); 19: Valencia orange (*C. sinensis*); 20: Jingcheng sweet orange (*C. sinensis*); 21: Honganliu sweet orange (*C. sinensis*); 22: sour orange (*C. aurantium*); 23: Cara Cara navel orange (*C. sinensis*); 24: Washington navel orange (*C. sinensis*); 25: blood orange (*C. sinensis*); 26: Ichang papeda (*C. ichangensis*); 27: papeda (*C. honghensis*); 28: citron (*C. medica*); 29: bergamot (*C. bergamia* Risso); 30: lemon (*C. limon*); 31: Mexican lime (*C. aurantifolia*); 32: rough lemon (*C. jambhiri*). This figure is available in black and white in print and in colour at *DNA Research* online.

diversity analysis among 32 citrus species (Fig. 5). The UPGMA phylogenetic tree revealed three major clusters including 4, 3, and 25 genotypes in each cluster (trifoliate orange, fortunella, and citrus, respectively). The UPGMA-based clustering showed the grouping of all 4 species (including the MT and the WT) of trifoliate orange in one group (trifoliate orange cluster), 3 species of fortunella, kumquat, and calamondin in one group (fortunella cluster), and 25 species in one group (citrus cluster). Traditionally, calamondin should be clustered citrus cluster. There might be two possible explanations: at first, calamondin with a short juvenile of 1–2 yr and flowered twice or three times per year similar to precocious trifoliate orange. Second, these new InDel markers may be related to flowering. In the citrus cluster, citrus limon, citrus medica, citrus grandis, and sweet orange formed one subcluster and were more closely related than the other subclusters of citrus. Nei's genetic distance values ranged from 0 (*C. sinesis* Newhall, Jingcheng, Cara Cara, Washington, and blood orange) to 0.41 (*P. trifoliata* Raf. × *C. sinensis* vs. *C. grandis* Kaopan), with an average value of 0.25 ± 0.07 (Fig. 5). The out-group trifoliate orange and fortunella showed relatively larger amounts of average genetic distance (0.39 and 0.31) from all the genotypes of citrus.

4. Discussion

To understand genetic variation involved in flowering development in citrus, we used genome resequencing technology to analyse the genetic variations between precocious trifoliate orange and its WT. Between the MT and the WT, 3,932,628 SNPs, 1,293,383 InDels, and

Table 1. Cross-species transferability of new InDel marker among different genomes in three genera of Rutaceae

Genus	Common name	Latin name	Ploidy	Cultivated/ wild	No. of InDel	% of amplified	Absent of amplified	PIC
Fortunella Swingle	Fortunella	*F. japonica*	2×	Cultivated	73	98.7	1	0.43
Fortunella Swingle	Kumquat	*F. hindsii* var. *chintou* Swing	2×	Cultivated	73	98.7	1	0.43
Trifoliate orange	Trifoliate orange	*P. trifoliata* [L.] Raf.	2×	Cultivated	72	97.4	2	0.45
Trifoliate orange	Precocious trifoliate orange	*P. trifoliata* [L.] Raf	2×	Cultivated	74	100	0	0.43
Trifoliate orange	Flyingdragon trifoliate orange	*P. trifoliata Raf.* var. *monstrosa*	2×	Wild	74	100	0	0.44
Trifoliate orange	Citranges	*P. trifoliata Raf x C. sinensis*	2×	Cultivated	74	100	0	0.48
Citrus	'Fenghuang' pummelo	*C. grandis*	2×	Cultivated	74	100	0	0.43
Citrus	'Kaopan' pummelo	*C. grandis*	2×	Cultivated	71	96.1	3	0.43
Citrus	Grapefruit	*C. paradisi*	2×	Cultivated	73	98.7	1	0.47
Citrus	Hirado Buntan pummelo	*C. grandis*	2×	Cultivated	73	98.7	1	0.44
Citrus	'Guoqing no.1' Satsuma	*C. unshiu*	2×	Cultivated	73	98.7	1	0.45
Citrus	'Qingjiang' Ponkan	*C. reticulata*	2×	Cultivated	74	100	0	0.44
Citrus	Bendiguangju mandarin	*C. reticulata cv. Succosa*	2×	Cultivated	74	100	0	0.44
Citrus	'Miyagawa wase' Satsuma	*C. unshiu*	2×	Cultivated	74	100	0	0.44
Citrus	Red tangerine	*C. reticulata*	2×	Cultivated	74	100	0	0.43
Citrus	Clementine mandarin	*C. clementina*	2×	Cultivated	74	100	0	0.46
Citrus	Newhall Navel Orange	*C. sinensis*	2×	Cultivated	73	98.7	1	0.47
Citrus	Valencia orange	*C. sinensis*	2×	Cultivated	74	100	0	0.47
Citrus	'Jingcheng' sweet orange	*C. sinensis*	2×	Cultivated	74	100	0	0.47
Citrus	'Honganliu' sweet orange	*C. sinensis*	2×	Cultivated	74	100	0	0.47
Citrus	Sour orange	*C. aurantium*	2×	Wild	74	100	0	0.45
Citrus	'CaraCara' navel orange	*C. sinensis*	2×	Cultivated	74	100	0	0.47
Citrus	Washington navel orange	*C. sinensis*	2×	Cultivated	74	100	0	0.47
Citrus	Blood orange	*C. sinensis*	2×	Cultivated	74	100	0	0.47
Citrus	Ichang papeda	*C. ichangensis*	2×	Cultivated	74	100	0	0.44
Citrus	Papeda	*C. honghensis*	2×	Wild	68	92.1	6	0.42
Citrus	Calamondin	*C. mitis*	2×	Cultivated	74	100	0	0.43
Citrus	Citron	*C. medica*	2×	Wild	73	98.7	1	0.40
Citrus	Bergamot	*C. bergamia*	2×	Wild	73	98.7	1	0.42
Citrus	Mexican lime	*C. aurantifolia*	2×	Cultivated	74	100	0	0.46
Citrus	lemon	*C. limon*	2×	Cultivated	73	98.7	1	0.47
Citrus	Rough lemon	*C. jambhiri*	2×	Cultivated	74	100	0	0.47

52,135 SVs were identified. In previous studies, precocious trifoliate orange is speculated to be a direct variant of the WT.[10,12,13] Such a large amount of genetic variation is found owing to the following reasons. For example, naturally occurring mutants are extensively found under the influence of environmental factors in woody perennials and humanity[36]; it was reported that there was many genetic variation in different tissues of the human body such as liver and lung.[37,38] On the other hand, the trifoliate orange produces polyembryonic seeds containing both sexual and apomictic embryos. About 10–20% of seedlings from open-pollinated trifoliate orange seeds develop from zygotic embryos; some genetic variations may be generated. Therefore, the genetic variations gradually increase from generation to generation.[39] In this study, >20% of the SNPs were heterozygous in two trifoliate oranges. Further supporting the observation that the InDels was also highly heterozygous, approximately 31.07 and 32.04% of the InDels were heterozygous in the MT and the WT, respectively. This observation suggested a notably high heterozygosity rate in the trifoliate orange genera. High levels of heterozygosity introduce additional challenges for the identification of various kinds of DNA polymorphisms, especially complex SVs.[37] Heterozygosity decreases the proportion of reads mapping to a unique genomic location in a manner dependent on the degree of heterozygosity, so only 61.73 and 63.13% of clean reads were uniquely mapped to the citrus reference genome. In previous studies, heterozygosity has been a common feature in most eukaryotic organisms and has shown important biological functions in woody plants.[40,41] Although citrus is generally perceived to have highly heterozygous traits,[14] the amount of heterozygosity in the whole genome is not clear. The results presented in this study further characterize the extent of genome heterozygosity and its functional effects at both the whole genome and the global transcriptome levels for two trifoliate oranges.

Knowing genomic positions of genetic variations in genetic markers is important.[42] Many population genetic and genetic mapping applications rely on unlinked markers. In this study, these genetic variations showed only minimal distribution in CDS regions, consistent with the results of SNP in tomato.[43] This might be related to the increased size of the intron in citrus. Interestingly, there were more SNPs than InDels and SVs in the CDS regions. This difference can be explained by the fact that InDels and SVs are more deleterious than SNPs in the CDS regions, as indicated by InDels and SVs that cause frameshift mutations and amino acid substitutions that cause major changes in gene function.[44,45] However, SNPs often produce synonymous mutations that have little or no impact on gene expression and function.[46] In this study, 16,919 MT-specific and 12,513 WT-specific SNPs predicted to cause non-synonymous amino acid substitutions were identified at the transcriptional level based on integrative analysis. The ratio of non-synonymous to synonymous substitutions was 2.22 and 2.19 in the WT and the MT, respectively. These SNPs may represent causal genetic variation contributing to phenotype variation in the MT and the WT. Without additional experimental evidence, we cannot yet say whether these SNPs affect the regulation of flowering-related genes or have another direct effect on

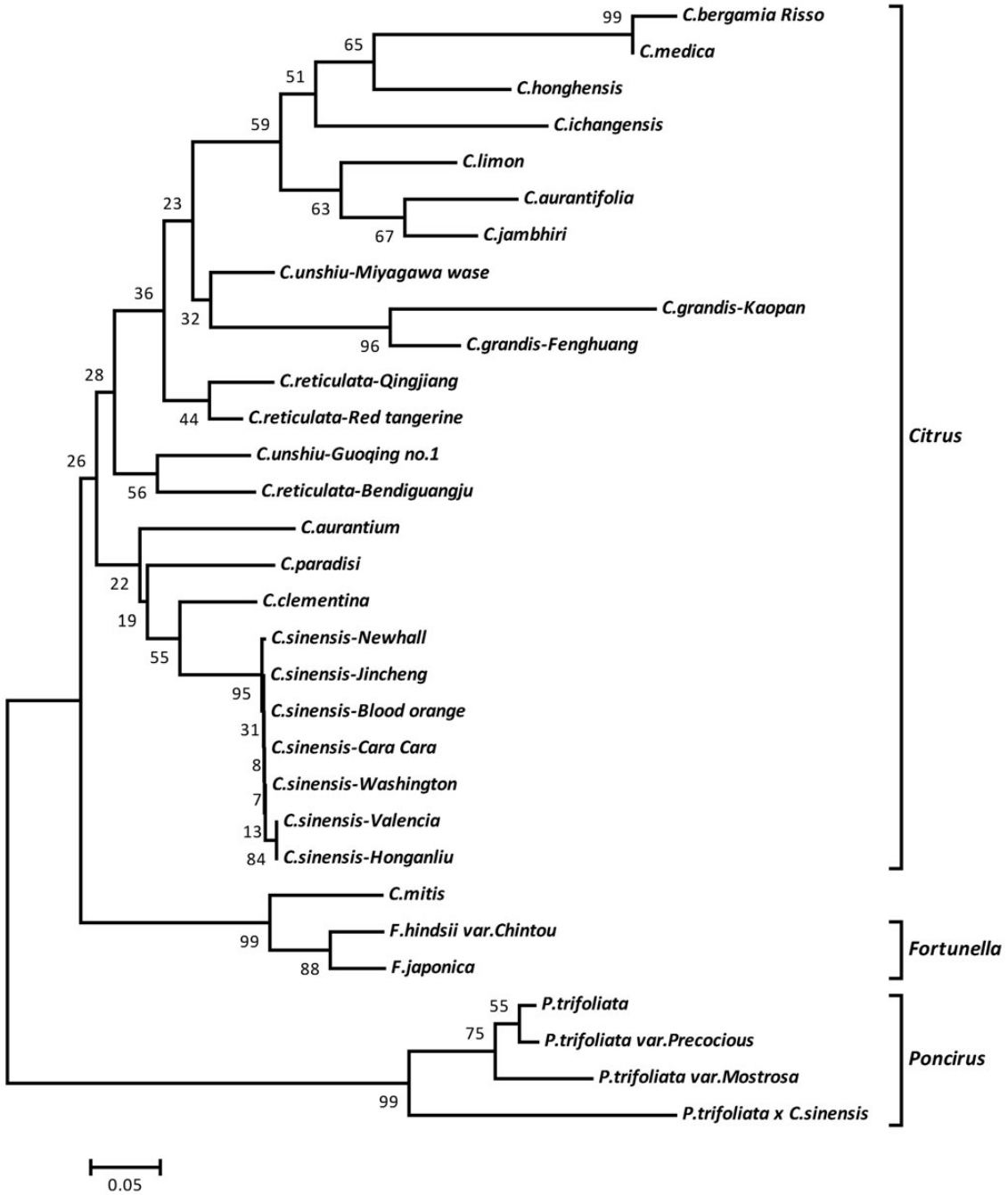

Figure 5. UPGMA tree of 32 citrus genotypes of three genera of Rutaceae based on Nei's genetic distance using 74 new InDel markers. *C. bergamia* Risso: bergamot; *C. medica*: citron; *C. honghensis*: papeda; *C. ichangensis*: Ichang papeda; *C. limon*: lemon; *C. aurantifolia*: Mexican lime; *C. jambhiri*: rough lemon; *C. unshiu*-Miyagawa wase: Miyagawa wase Satsuma; *C. grandis*-Kaopan: Kaopan pummelo; *C. grandis*-Fenghuang: Fenghuang pummelo; *C. grandis*-Hirado: Hirado Buntan pummelo; *C. reticulate*-Qingjiang: Qingjiang ponkan; *C. reticulate*-red tangerine: red tangerine; *C. unshiu*-Guoqing no.1: Guoqing no. 1 satsuma fortunella; *C. reticulate*-Bendiguangju: Bendiguangju mandarin; *C. aurantium*: sour orange; *C. grandis*: grapefruit; *C. clementina*: Clementine mandarin; *C. sinensis*-Newhall*:* Newhall navel Orange; *C. sinensis*-Jingcheng: Jingcheng sweet orange; *C. sinensis*-blood orange: blood orange; *C. sinensis*-Cara Cara: Cara Cara navel orange; *C. sinensis*-Washington: Washington navel orange; *C. sinensis*-Valencia: Valencia orange; *C. sinensis*-Honganliu: Honganliu sweet orange; *C. mitis*: calamondin; *F. hindsii* var. *chintou*: kumquat; Fortunella: *F. japonica*; *P. trifoliate*: trifoliate orange; *P. trifoliate* var. precocious: precocious trifoliate orange; *P. trifoliata* var. *monstrosa*: flying dragon trifoliate orange; *P. trifoliata* Raf × *C. sinensis*: cotranges.

mutant phenotypes. Nevertheless, using this SNP set to perform genome-wide association in two genotype genomes would be more efficient than using a general SNP set to identify causal gene mutations.

Combining genome and transcriptome profiling data has been used as a powerful approach for the identification of functional genetic variations and candidate genes for traits of interest.[47] To investigate the candidate genetic variation responsible for early flowering, we integrated the RNA-Seq and genome resequencing of genetic variation data. A total of 362 large-effect SNPs were found between the MT and the WT and were found to affect the integrity of encoded proteins.

These large-effect SNPs included disruption of splice sites, loss of translation initiation codon, introduction of premature stop codon, and loss of stop codon. Likewise, we identified 1,137 and 721 specific InDels in the CDS region of the MT and the WT, respectively, during the phase stage, which may cause frameshift, disruption of splice sites, or introduction of premature stop codon in the two genotypes of trifoliate orange. The significant proportion of genes containing large-effect SNPs and InDels suggested that these genes may represent differences in gene sequences between the MT and the WT. GO analysis indicated that these genes predicted to contain specific InDels were more commonly associated with binding than with other functionality in this study. This may suggest that proteins with the function of binding may play a significant role in the early flowering process of the MT. In addition, the significantly differentially expressed genes contain specific InDels (Supplementary Table S3) that can be directly linked to trifoliate orange flowering transition. Many known transcription factors related to flowering regulation were also differentially expressed, including MADS-box, ring zinc finger, MYB, and GRAS (Supplementary Table S3). The transcript levels of these transcription factors were higher during the phase stage when floral buds began flowering compared with the WT, suggesting that its main role was associated with flowering time in the MT. These differentially expressed genes covered a broad range of genes related to flowering regulation, providing helpful information for understanding the genetic mechanisms underlying the signalling and regulation of the transition from the vegetative to the reproductive phase.

However, structural alterations of genes are generally believed to modify their function and expression. We found that at least 2,288 and 2,260 specific SVs showed changes in the CDS region of the MT and the WT, respectively. Although further studies are required to identify the specific gene(s) responsible for the early flowering trait, the genes could be interesting candidates for investigation in further studies. In addition, we determined whether the genes related to the flowering pathway controlling flowering time were significantly altered in the MT genome. Annotation analysis showed that 1,073 SNPs, 382 InDels, and 50 SVs were located in 272 flowering-related transcripts, respectively. This finding might suggest that the early flowering of the MT was correlated to genetic variations of these flowering pathway genes. An alternative hypothesis is that the phenotype was caused by the changes in the function of the genes through protein structure alterations. Differences in the expression patterns of flowering pathway genes were observed in the MT compared with the WT, and this might support this hypothesis. Although the exact function of these mutated genes, especially with respect to flowering, was not determined, these genes might be interesting candidates for further studies.

Recently, a large number of InDels, SNPs, and SVs have been generated using the genome resequencing platform in citrus.[9,40,41] These genetic variation markers may be assayed using the same separation and detection technologies as simple sequence repeat (SSR) markers. In fact, some InDels may be caused by SSRs. *Citrus trifoliata* (representing *P. trifoliata*) has consistently been one of the most important rootstock species used in the citrus industry, and it has even been used as a model species for citrus molecular biology and genomic studies. These new InDel markers are specific to trifoliate orange. Therefore, they will play important roles in marker-assisted selection and citrus breeding. In addition, these markers may be related to the flowering they ascribed to their high distribution, and whole-genome polymorphisms have been applied to high-resolution genetic mapping and map-based cloning in flowering genes.[48,49] However, the usefulness of genetic variation has not been explored in genetic and genomic research for citrus. To verify that these genetic variations were suitable for use as new DNA markers, they were used to successfully design PCR primers. In this study, we selected 268 InDels to develop PCR-based markers, of which 74 (26.7%) were polymorphic either between the two resequenced trifoliate orange or among 30 accessions from different subspecies of citrus, indicating that the bioinformatics tools were inadequate for detecting InDels in the trifoliate orange genome. However, the genetic variation will be useful to select candidate genetic variation that could be associated with unique phenotypes or agronomical traits in citrus. Furthermore, our collection of genetic variations that differentiate the MT from the WT can be used to guide the search for pure trifoliate orange types (or to recognize other cryptic species) among the hundreds of known cultivars and other germplasm accessions. More importantly, because the coordinates of these loci are known in relation to a reference genome, it is possible to develop genetic markers within specific genome regions to assist in efficient construction of genetic maps in the future.

Acknowledgements

We are grateful to Yong-Ping Li for his helpful discussion and help in bioinformatics analysis during revising this manuscript.

Supplementary Data

Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This research was supported financially by the National Natural Science Foundation of China (31130046, 31471863, 31372046, 31221062, and 31101528), the Fundamental Research Funds for the Central Universities (2013PY083), and the International Foundation for Science (C/5148-2). Funding to pay the Open Access publication charges for this article was provided by the National Science Foundation of China (grant no. 31130046).

References

1. Liu, C., Xi, W., Shen, L., Tan, C. and Yu, H. 2009, Regulation of floral patterning by flowering time genes, *Dev. Cell*, **16**, 711–22.
2. Khan, M.R.G., Ai, X.-Y. and Zhang, J.-Z. 2014, Genetic regulation of flowering time in annual and perennial plants, *Wiley Interdiscip. Rev. RNA*, **5**, 347–59.
3. Mouradov, A., Cremer, F. and Coupland, G. 2002, Control of flowering time interacting pathways as a basis for diversity, *Plant Cell*, **14**(Suppl. 1), S111–30.
4. Boss, P.K., Bastow, R.M., Mylne, J.S. and Dean, C. 2004, Multiple pathways in the decision to flower: enabling, promoting, and resetting, *Plant Cell*, **16**(Suppl. 1), S18–31.
5. Ferrario, S., Immink, R.G. and Angenent, G.C. 2004, Conservation and diversity in flower land, *Curr. Opin. Plant Biol.*, **7**, 84–91.
6. Tan, F.C. and Swain, S.M. 2006, Genetics of flower initiation and development in annual and perennial plants, *Physiol. Plant.*, **128**, 8–17.
7. Tan, F.C. and Swain, S.M. 2007, Functional characterization of AP3, SOC1 and WUS homologues from citrus (Citrus sinensis), *Physiol. Plant.*, **131**, 481–95.
8. Gmitter, F.G. Jr, Chen, C., Rao, M.N. and Soneji, J.R. 2007, Citrus fruits. In: *Fruits and Nuts, Genome Mapping & Molecular Breeding in Plants*. Springer: Heidelberg, pp. 265–79.
9. Wu, G.A., Prochnik, S., Jenkins, J., et al. 2014, Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication, *Nat. Biotechnol.*, **32**, 656–62.
10. Liang, S., Zhu, W. and Xiang, W. 1999, Precocious trifoliate orange (Poncirus trifoliata L. Raf.) biology characteristic and its stock experiment, *ZheJiang Citrus*, **16**, 2–4.
11. Zhang, J.-Z., Ai, X.-Y., Sun, L.-M., et al. 2011, Transcriptome profile analysis of flowering molecular processes of early flowering trifoliate orange

mutant and the wild-type [Poncirus trifoliata (L.) Raf.] by massively parallel signature sequencing, *BMC Genomics*, **12**, 63.
12. Pang, X., Deng, X. and Hu, C. 2003, Construction of AFLP fingerprint of 36 Poncirus accessions, *Acta Horticulturae Sinica*, **30**, 394–8.
13. Zhang, J.-Z., Li, Z.-M., Yao, J.-L. and Hu, C.-G. 2009, Identification of flowering-related genes between early flowering trifoliate orange mutant and wild-type trifoliate orange (Poncirus trifoliata L. Raf.) by suppression subtraction hybridization (SSH) and macroarray, *Gene*, **430**, 95–104.
14. Gmitter, F.G. Jr, Chen, C., Machado, M.A., et al. 2012, Citrus genomics, *Tree Genet. Genomes*, **8**, 611–26.
15. Metzker, M.L. 2010, Sequencing technologies-the next generation, *Nat. Rev. Genet.*, **11**, 31–46.
16. Li, Z.-M., Zhang, J.-Z., Mei, L., Deng, X.-X., Hu, C.-G. and Yao, J.-L. 2010, PtSVP, an SVP homolog from trifoliate orange (Poncirus trifoliata L. Raf.), shows seasonal periodicity of meristem determination and affects flower development in transgenic Arabidopsis and tobacco plants, *Plant Mol. Biol.*, **74**, 129–42.
17. Cheng, Y.-J., Guo, W.-W., Yi, H.-L., Pang, X.-M. and Deng, X. 2003, An efficient protocol for genomic DNA extraction from Citrus species, *Plant Mol. Biol. Rep.*, **21**, 177–8.
18. Datta, S., Datta, S., Kim, S., Chakraborty, S. and Gill, R.S. 2010, Statistical analyses of next generation sequence data: a partial overview, *J. Proteomics Bioinform.*, **3**, 183–90.
19. Li, H. and Durbin, R. 2010, Fast and accurate long-read alignment with Burrows–Wheeler transform, *Bioinformatics*, **26**, 589–95.
20. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The sequence alignment/map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
21. Li, R., Li, Y., Fang, X., et al. 2009, SNP detection for massively parallel whole-genome resequencing, *Genome Res.*, **19**, 1124–32.
22. Li, R., Yu, C., Li, Y., et al. 2009, SOAP2: an improved ultrafast tool for short read alignment, *Bioinformatics*, **25**, 1966–7.
23. Birney, E., Clamp, M. and Durbin, R. 2004, GeneWise and genomewise, *Genome Res.*, **14**, 988–95.
24. Zhang, J.-Z., Li, Z.-M., Liu, L., Mei, L., Yao, J.-L. and Hu, C.-G. 2008, Identification of early-flower-related ESTs in an early-flowering mutant of trifoliate orange (Poncirus trifoliata) by suppression subtractive hybridization and macroarray analysis, *Tree Physiol.*, **28**, 1449–57.
25. Zenoni, S., Ferrarini, A., Giacomelli, E., et al. 2010, Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq, *Plant Physiol.*, **152**, 1787–95.
26. Schmid, R. and Blaxter, M.L. 2008, annot8r: GO, EC and KEGG annotation of EST datasets, *BMC Bioinform.*, **9**, 180.
27. Liu, S.-R., Li, W.-Y., Long, D., Hu, C.-G. and Zhang, J.-Z. 2013, Development and characterization of genomic and expressed SSRs in citrus by genome-wide analysis, *PLoS One*, **8**, e75149.
28. Yeh, F.C., Yang, R., Boyle, T., Ye, Z. and Mao, J.X. 1999, POPGENE, version 1.32: the user friendly software for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Edmonton, AB, Canada.
29. Kimura, M. and Crow, J.F. 1964, The number of alleles that can be maintained in a finite population, *Genetics*, **49**, 725–38.
30. Lewontin, R.C. 1972, Testing the theory of natural selection, *Nature*, **236**, 181–2.
31. Nei, M. 1973, Analysis of gene diversity in subdivided populations, *Proc. Natl. Acad. Sci. USA*, **70**, 3321–3.
32. Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2007, MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0, *Mol. Biol. Evol.*, **24**, 1596–9.
33. Geuna, F., Toschi, M. and Bassi, D. 2003, The use of AFLP markers for cultivar identification in apricot, *Plant Breed.*, **122**, 526–31.
34. Kumar, P., Henikoff, S. and Ng, P.C. 2009, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, *Nat. Protoc.*, **4**, 1073–81.
35. Subbaiyan, G.K., Waters, D.L., Katiyar, S.K., Sadananda, A.R., Vaddadi, S. and Henry, R.J. 2012, Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing, *Plant Biotechnol. J.*, **10**, 623–34.
36. Zhang, M. and Deng, X. 2006, Advances in research of citrus cultivars selected by bud mutation and the mechanism of formation of mutated characteristics, *J. Fruit Sci.*, **23**, 871–6.
37. Alkan, C., Coe, B.P. and Eichler, E.E. 2011, Genome structural variation discovery and genotyping, *Nat. Rev. Genet.*, **12**, 363–76.
38. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. and Kinzler, K.W. 2013, Cancer genome landscapes, *Science*, **339**, 1546–58.
39. Khan, I. and Roose, M. 1988, Frequency and characteristics of nucellar and zygotic seedlings in three cultivars of trifoliate orange, *J. Am. Soc. Hortic. Sci.*, **113**, 105–10.
40. Jiao, W.-B., Huang, D., Xing, F., et al. 2013, Genome-wide characterization and expression analysis of genetic variants in sweet orange, *Plant J.*, **75**, 954–64.
41. Xu, Q., Chen, L.-L., Ruan, X., et al. 2013, The draft genome of sweet orange (Citrus sinensis), *Nat. Genet.*, **45**, 59–66.
42. Xu, X., Liu, X., Ge, S., et al. 2012, Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes, *Nat. Biotechnol.*, **30**, 105–11.
43. Kobayashi, M., Nagasaki, H., Garcia, V., et al. 2014, Genome-wide analysis of intraspecific DNA polymorphism in 'Micro-Tom', a model cultivar of tomato (Solanum lycopersicum), *Plant Cell Physiol.*, **55**, 445–54.
44. Clark, R.M., Schweikert, G., Toomajian, C., et al. 2007, Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana, *Science*, **317**, 338–42.
45. Lai, J., Li, R., Xu, X., et al. 2010, Genome-wide patterns of genetic variation among elite maize inbred lines, *Nat. Genet.*, **42**, 1027–30.
46. Mills, R.E., Pittard, W.S., Mullaney, J.M., et al. 2011, Natural genetic variation caused by small insertions and deletions in the human genome, *Genome Res.*, **21**, 830–9.
47. Jones, D.B., Jerry, D.R., Forêt, S., Konovalov, D.A. and Zenger, K.R. 2013, Genome-wide SNP validation and mantle tissue transcriptome analysis in the silver-lipped pearl oyster, Pinctada maxima, *Mar. Biotechnol.*, **15**, 647–58.
48. Arai-Kichise, Y., Shiwa, Y., Nagasaki, H., et al. 2011, Discovery of genome-wide DNA polymorphisms in a landrace cultivar of japonica rice by whole-genome sequencing, *Plant Cell Physiol.*, **52**, 274–82.
49. Păcurar, D.I., Păcurar, M.L., Street, N., et al. 2012, A collection of INDEL markers for map-based cloning in seven Arabidopsis accessions, *J. Exp. Bot.*, **63**, 2491–501.