



ARTICLE

<https://doi.org/10.1038/s41467-019-09518-x>

OPEN

# A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour

Liyi Zhang<sup>1</sup>, Jiang Hu<sup>2</sup>, Xiaolei Han<sup>1</sup>, Jingjing Li <sup>2</sup>, Yuan Gao<sup>1</sup>, Christopher M. Richards <sup>3</sup>, Caixia Zhang<sup>1</sup>, Yi Tian<sup>1</sup>, Guiming Liu<sup>4</sup>, Hera Gul<sup>1</sup>, Dajiang Wang<sup>1</sup>, Yu Tian<sup>2</sup>, Chuanxin Yang<sup>2</sup>, Minghui Meng<sup>2</sup>, Gaopeng Yuan<sup>1</sup>, Guodong Kang<sup>1</sup>, Yonglong Wu<sup>1</sup>, Kun Wang<sup>1</sup>, Hengtao Zhang<sup>5</sup>, Depeng Wang<sup>2</sup> & Peihua Cong<sup>1</sup>

A complete and accurate genome sequence provides a fundamental tool for functional genomics and DNA-informed breeding. Here, we assemble a high-quality genome (contig N50 of 6.99 Mb) of the apple anther-derived homozygous line HFTH1, including 22 telomere sequences, using a combination of PacBio single-molecule real-time (SMRT) sequencing, chromosome conformation capture (Hi-C) sequencing, and optical mapping. In comparison to the Golden Delicious reference genome, we identify 18,047 deletions, 12,101 insertions and 14 large inversions. We reveal that these extensive genomic variations are largely attributable to activity of transposable elements. Interestingly, we find that a long terminal repeat (LTR) retrotransposon insertion upstream of *MdMYB1*, a core transcriptional activator of anthocyanin biosynthesis, is associated with red-skinned phenotype. This finding provides insights into the molecular mechanisms underlying red fruit coloration, and highlights the utility of this high-quality genome assembly in deciphering agriculturally important trait in apple.

<sup>1</sup>Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Research Institute of Pomology, Chinese Academy of Agricultural Science, 125100 Xingcheng, Liaoning, China. <sup>2</sup>Nextomics Biosciences Institute, 430000 Wuhan, Hubei, China. <sup>3</sup>USDA-ARS National Center for Genetic Resources Preservation, Fort Collins, CO 80521, USA. <sup>4</sup>Beijing Agro-Biotechnology Research Center, Beijing Academy of Agriculture and Forestry Sciences, 100097 Beijing, China. <sup>5</sup>Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Science, 450009 Zhengzhou, Henan, China. These authors contributed equally: Liyi Zhang, Jiang Hu. Correspondence and requests for materials should be addressed to D.W. (email: [wangdp@grandomics.com](mailto:wangdp@grandomics.com)) or to P.C. (email: [congpeihua@caas.cn](mailto:congpeihua@caas.cn))

The apple genome is a foundation of genetic research and DNA-informed breeding that drives innovations for sustainable apple production<sup>1,2</sup>. Although the availability of the current high-quality genome of Golden Delicious and the resequencing of major genotypes enable rapid progress in apple genomics and breeding studies<sup>3–6</sup>, only a single reference genome together with short-read resequencing data presents some limitations in the discovery of new genes and characterisation of genomic variations, which may substantially contribute to genome evolution and the genetics of complex traits<sup>3,7,8</sup>. A large-scale survey of genomic variations will provide insights into the potential biological mechanisms of key traits, which in turn will aid the development of genetic markers for marker-assisted selection (MAS) breeding in apple. Recent studies have demonstrated that an extensive range of functional genomic variation can be readily uncovered through direct comparative analyses of the several high-quality genomes<sup>9,10</sup>. In addition, apple (*Malus domestica* Borkh.) is among the most diverse and economically important fruit species in the Rosaceae family<sup>3,6</sup>, but we still lack an in-depth understanding of genetic basis of its many economically important traits. In the case of red skin colouration, although the developmental and environmental regulatory mechanisms of the anthocyanin biosynthetic pathway have been well characterised and the corresponding genes have been identified<sup>11–14</sup>, the genetic basis for the regulation of fruits colouration is not yet fully understood. Dissection of the genetic determinants of this crucial trait appears to be difficult without the availability of high-quality genome sequences.

All cultivated apple genotypes are a highly heterozygous and ancient autotetraploid of 17 chromosomes, presenting enormous challenges for genome analyses and breeding<sup>3</sup>. Thus, the anther-derived homozygous genotype HFTH1 was developed for sequencing (Fig. 1a). This homozygous line has the advantage of simplifying genome assembly<sup>3</sup>, and its parentage Hanfu (Dongguang x Fuji) is a very influential cultivar in China, that has several desirable traits, including bright red skin, cold-resistance and long shelf-life<sup>15</sup>. Here, we present a high-quality reference genome of HFTH1 using a combination of SMRT sequencing, Hi-C sequencing and optical mapping. We then use this genome to perform a comparative genomic analysis with the existing apple genomes. We track the highly dynamic evolution of transposons, and discover an LTR retrotransposon that is associated with the red-skinned phenotype and thus serves as a valuable tool for MAS breeding. This additional reference genome provides a foundation for functional genomics and transposon biology, and enhances our understanding of the genome variation that shapes phenotypic diversity in apple.

## Results

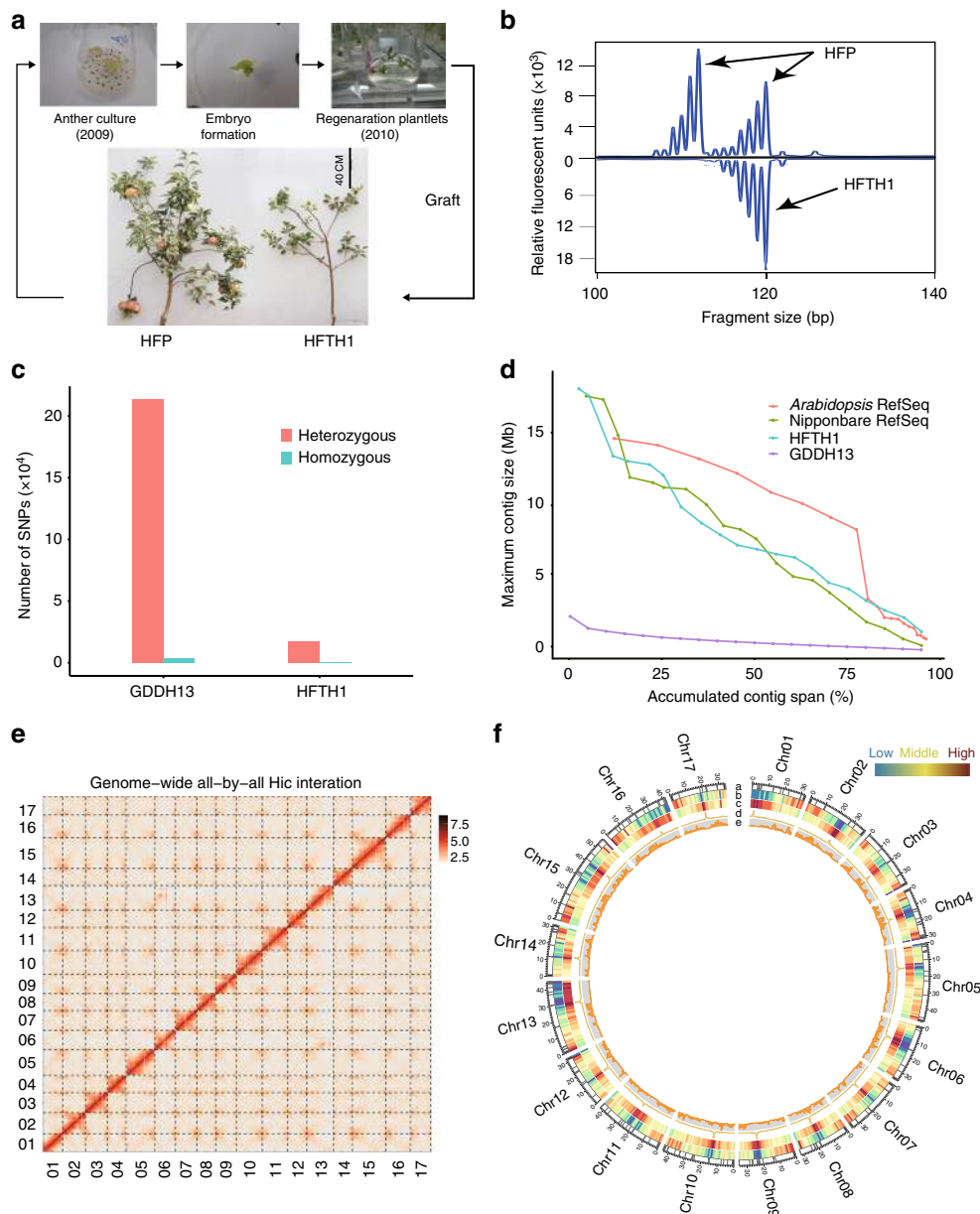
**Genome sequencing and assembly.** To minimise the complexity of assembly, an anther-derived trihaploid Hanfu line (HFTH1) was used for genome sequencing, while its donor Hanfu was a highly heterozygous diploid cultivar (HFP, Fig. 1a, b and Supplementary Fig. 1). The homozygosity of the HFTH1 genome was confirmed by microsatellite markers and k-mer spectrum analysis (Fig. 1b and Supplementary Figs. 1, 2). The HFTH1 genome was determined as a triploidy by flow cytometry analysis, suggesting that HFTH1 had undergone spontaneous chromosome duplication during *in vitro* culture (Supplementary Fig. 3). Compared with the recently published genome of double haploid Golden Delicious line<sup>3</sup> (GDDH13), HFTH1 was more homozygous, as demonstrated by calling heterozygous SNPs from Illumina reads obtained from the corresponding individual (Fig. 1c). This higher level of homozygosity is more favourable for improving the quality of the genome assembly.

Our sequencing of HFTH1 resulted in coverage of ~117-fold PacBio single-molecule long reads (77 Gb with an average length of 13.1 kb), 66-fold Illumina paired-end short reads (43.3 Gb), 224-fold optical map data (147.8 Gb with an average length of 178.9 kb) and 145-fold Hi-C data (95.5 Gb, Supplementary Table 1). The assembly was performed in a stepwise fashion<sup>16</sup>, and the initial assembly of the PacBio-only data generated a 656.52 Mb genome size with a contig N50 of 4.63 Mb (Supplementary Table 2). The initial contigs were polished with PacBio long reads and Illumina short reads. Subsequently, the polished contigs were scaffolded using optical map data, and during this step four contigs containing conflicting connections were identified and split to resolve conflicts, and 58.5% gaps that were introduced in this step were closed by subsequent gap filling procedure. Finally, scaffolding with Hi-C data allowed the accurate clustering and ordering of 17 pseudo-chromosomes covering the 658.90 Mb assembly, with a contig N50 of 6.99 Mb and a maximum contig length of 18.01 Mb (Supplementary Table 3 and Fig. 1d, e). The assembly size was close to the estimated genome size of GDDH13<sup>3</sup>, but represented 92.99% of our estimated genome size (708.54 Mb) for HFTH1 by k-mer analysis, and ~97.89% of the Illumina reads of HFTH1 could be mapped to our assembly (Supplementary Table 4). In addition, the 160,068 bp chloroplast genome and 396,939 bp mitochondria genome were assembled into two complete contigs (Supplementary Fig. 4).

**Assessment of genome quality.** The quality and completeness of the assembly were evaluated using several different strategies. For the base accuracy of the sequencing, the quality value (QV) of the assembly was estimated to be at least 41, which compared very favourably with those of two published mammalian genomes<sup>16,17</sup> (QV35 for gorilla and QV34.5 for goat) that were also assembled with PacBio data. For the structural accuracy of assembly, ~98.71% of the mapped Illumina reads of HFTH1 could be mapped with the correct orientation and estimated insert size, versus 94.36% of the mapped reads of GDDH13 (Supplementary Table 4). Furthermore, the whole-genome alignment of HFTH1 and GDDH13 showed strong collinearity and consistency (Supplementary Fig. 5).

Our assembly captured 22 long stretches of telomeric sequences (5'-TTTAGGG-3') at both ends of seven chromosomes and at a single end of eight chromosomes, with repeat numbers ranging from 294 to 1073 (Table 1 and Fig. 1f). Moreover, based on a benchmark of 1440 conserved plant genes<sup>18</sup>, ~97.0% complete BUSCO genes and 98.02% of the expressed sequence tags (ESTs) of *Malus domestica* from GenBank could be detected in the assembly (Supplementary Tables 3 and 5), confirming the high completeness of the assembly. In addition, the contig N50 value of our assembly was comparable to that of the rice genome<sup>19</sup> (RGAP7), and the HFTH1 genome covered approximately the same BUSCOs as the *Arabidopsis* genome<sup>20</sup> (TAIR10) and RGAP7 genome (Supplementary Table 3 and Fig. 1d), which demonstrated that the level of completeness of our genome was at a similar to that of the model plant genomes, even though HFTH1 had the largest genome size and the highest proportion of repetitive sequences compared with the genomes of *Arabidopsis* and rice (Supplementary Table 3). The completeness of our genome provided an opportunity to comprehensively assess genome variations between HFTH1 and GDDH13.

**Genome annotation.** We identified 44,677 high-confidence protein-coding genes, and ~4.29% of annotated genes were located in the gap regions of the GDDH13 genome. The annotated genes covered 95.9% of the complete BUSCO genes



**Fig. 1** Overview of the assembly quality and characteristics of the HFTH1 genome. **a** Regenerated plantlets derived from Hanfu anther culture and the dramatic changes between the HFTH1 phenotype and the heterozygous donor (HFP) genotype. **b** Homozygosity analysis of chromosome 01 of HFTH1 and HFP using simple sequence repeats (SSRs, the results of the analysis of all chromosomes are shown in Supplementary Fig. 1). **c** Counts of SNPs detected in the GDDH13 and HFTH1 genomes. SNPs were detected using Illumina reads from the same individual that was used for the assembly. The heterozygous SNP (red) represents the heterozygosity of a genome and the homozygous SNP (blue) represents the potential error for an assembly. **d** Distribution of the contig length in the assemblies of *Arabidopsis* (TAIR10), Nipponbare (RGAP7), HFTH1 and GDDH13. Each contig size (y-axis) represents the minimum contig size that covered the cumulative percentage of the assembly (x-axis) after the assembled contigs were sorted from the largest to the smallest. **e** Hi-C interactions among 17 chromosomes with a 100-kb resolution. Strong interactions are indicated in dark red and weak interactions are indicated in yellow. **f** Circular diagram depicting the characteristics of the HFTH1 genome. The tracks from outer to inner circles indicate the following: **a** chromosomes (Chr.), gaps and telomeres, the black regions at the end of each chromosome represent assembled the telomere sequences and the grey regions represent gap regions; **b** gene density (window size of 1 Mb); **c** repeat density (window size of 1 Mb); **d** Copia-7; **e** Golden Delicious shared SNPs (window size of 500 kb)

(Supplementary Table 3), and ~92.28% of the annotated genes were expressed in at least one tissue or homologous to known proteins, which suggested that our genes annotation was very complete.

We employed a combination of de novo and homology-based approaches to annotate repetitive sequences. Approximately 393.88 Mb and 362.18 Mb transposable elements (TEs) were identified in the HFTH1 and GDDH13 genomes, respectively

(Supplementary Tables 3 and 6). The difference in the repeat sequence content (~32 Mb) can account for 93.09% of the additional non-N bases between the HFTH1 and GDDH13 genomes. Although we were unable to identify any significantly enriched tandem centromeric repeat elements in the HFTH1 genome using existing approaches<sup>21,22</sup>, 17 blocks (one block per chromosome) with particularly high proportions of repeat elements (>90%) in the HFTH1 genome, were located near the

**Table 1 Statistics of HFTH1 and GDDH13 genome assemblies**

Chromosome	HFTH1			GDDH13		
	Length (bp)	# Gaps	Telomere	Length (bp)	# Gaps	Telomere
Chr01	32,944,118	12	Single	32,625,452	85	Single
Chr02	38,449,405	8	Both	37,577,729	87	-
Chr03	37,138,690	4	Single	37,524,076	80	-
Chr04	31,012,745	7	Both	32,301,874	64	Single
Chr05	47,891,858	13	Single	47,952,461	107	-
Chr06	35,567,198	5	Both	37,137,259	88	Single
Chr07	35,934,761	5	Both	36,691,129	75	-
Chr08	31,511,015	7	Single	31,609,270	66	Single
Chr09	34,800,404	9	Single	37,604,908	79	Single
Chr10	43,815,736	13	Both	41,762,413	82	Single
Chr11	42,456,296	14	Single	43,059,885	90	-
Chr12	32,285,079	8	Single	33,050,054	74	-
Chr13	44,866,511	12	Single	44,339,518	118	Single
Chr14	31,515,206	5	Both	32,513,452	61	Single
Chr15	56,644,392	15	-	54,945,402	128	Single
Chr16	41,670,059	14	-	41,389,449	92	-
Chr17	33,998,825	7	Both	34,748,701	75	Single
mtDNA	396,939	0	NA	396,947	0	NA
cpDNA	160,068	0	NA	160,068	0	NA
Unanchored	7,992,922	326	-	52,728,359	839	-

NA not available

middle of chromosomes (Supplementary Table 7 and Fig. 1f). These regions can be assumed to represent part of putative heterochromatin regions on the HFTH1 chromosomes. The most abundant repeat element family in these regions was Copia-7 (ID from Repbase<sup>23</sup>), and most of the Copia-7 elements were concentrated in these regions (Fig. 1f). Although *HODOR* (a high-copy Golden Delicious repeat) was identified as the most repetitive consensus sequence in the apple genome<sup>3</sup>, the Copia-7 elements showed low similarity with *HODOR*.

**Gap filling for the reference Golden Delicious genome.** Owing to the genomic congruence between GDDH13 and HFTH1, we used the HFTH1 genome to fill gaps in the GDDH13 genome, as a similar approach has been applied to the genomes of human<sup>24</sup>, gorilla<sup>17</sup> and goat<sup>16</sup>. In total, 488 gaps (adjacent gaps were merged and counted as one gap) in the GDDH13 genome, with average and median lengths of 78,864 bp and 15,647 bp, were completely closed (Supplementary Fig. 6a). Among these, the longest gap closure was 2,859,572 bp and spanned 256 genes with 33.98% tandem duplicated genes. Approximately 97.75% of the closed gaps were located in repeat regions (Supplementary Fig. 6b), and most of these gaps had been assembled into multiple short fragments or filled with ambiguous (N) bases in previously published genomes<sup>3–5</sup>. In particular, ~39.34% sequences of the filled genomic gaps in GDDH13 could be found on its unanchored Chr00. For example, one 719,872 bp gap (the expected gap size was 728,347 bp) could be completely closed using the HFTH1 assembly, whereas this gap was assembled into 2 and 11 fragments in the assemblies reported by Velasco et al.<sup>4</sup> and Li et al.<sup>5</sup>, respectively (Supplementary Fig. 6c). In addition, we filled nine genomic gaps with an average length of 42,360 bp for the HFTH1 genome using the GDDH13 genome. These additional sequences will help to improve genomic annotation and discover thousands of functional genes and regulatory elements.

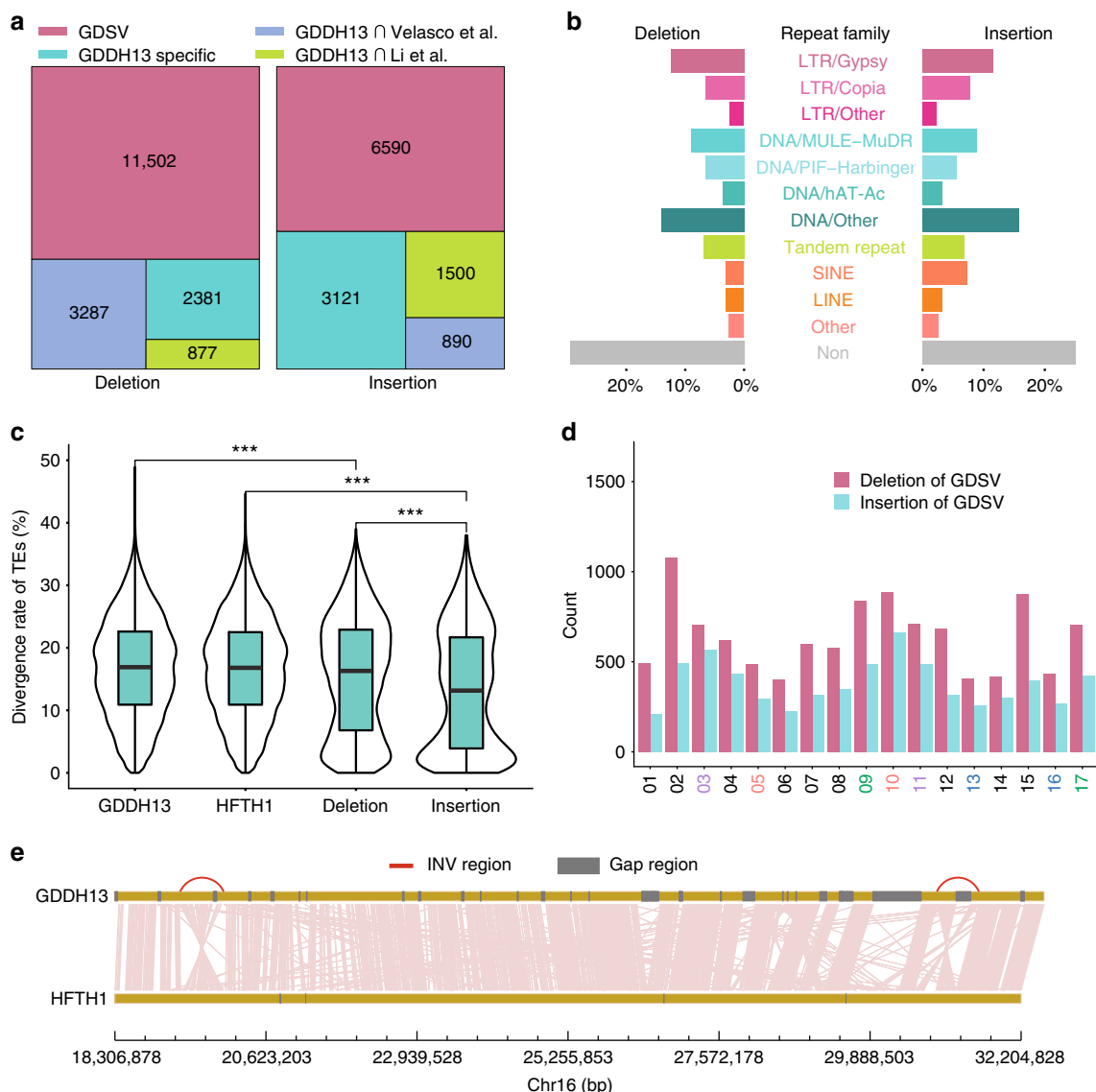
**Genome comparison between Golden Delicious and Hanfu.** Genome variations, such as insertions, deletions, inversions and duplications, are an important source of the genetic diversity that

shapes phenotypic variations. Hanfu exhibits an obviously different phenotype from that of Golden Delicious, including red skin, a strong cold-resistant habit, a high resistance to alternaria leaf spot and branch ring-rot, long fruit storability, and an obvious short branching character<sup>15</sup>. The comparison of the Golden Delicious genome with our HFTH1 reference genome, allowed us to directly catalogue the extent of the genomic variation between these two cultivars.

First, an average density of 2.15 Golden Delicious shared SNPs per kilobase was identified (Fig. 1f). Approximately 3.34% of the SNPs were located within 31.69% of protein-coding genes containing non-synonymous substitutions. An enrichment analysis of the InterPro domains of genes with non-synonymous SNPs, showed that these genes were significantly correlated with the disease-resistant domains (Supplementary Table 8), which demonstrated that these genes may evolve under different selection pressures in two cultivars, and provide resistance to various environmental stresses.

We identified 18,047 deletions and 12,101 insertions (including absence/presence variations) that were had a length greater than 100 bp in length. Of these, the presences of 63.73% of the deletions and 54.46% of the insertions in all published genomes of Golden Delicious (Fig. 2a), were defined as Golden Delicious shared structural variations (GDSVs). An enrichment analysis of the genes associated with GDSVs agreed with the results of the non-synonymous SNP analysis (Supplementary Table 9). For instance, one GDSV that includes a short interspersed nuclear element (SINE) was inserted in the 3'UTR of a disease-resistance gene (Supplementary Fig. 7a). Another GDSV was a deletion in the 5'UTR of a C-repeat/DRE binding factor (*MdCBF2*), which is a master regulator of cold acclimation<sup>25</sup> (Supplementary Fig. 7b).

In addition, among the GDSVs, the average lengths of the deletions (508 bp) and insertions (519 bp) were very similar (Supplementary Fig. 8), and >70% of GDSVs were associated with TEs. The TE distribution patterns of insertions and deletions were similar, with the exception that the percentage of SINEs associated with insertion was approximately twofold higher than that found for deletions, which accounted for only 3–8% of



**Fig. 2** Characterisation of structural variants. **a** The overlap of structural variations (> 100 bp) between GDDH13 and other published assemblies of Golden Delicious. **b** Classification of repeat elements associated with structural variations. A repeat element was defined as being associated with a structural variation if it overlaps with the structural variation. LTR, long terminal repeat; DNA, DNA transposon; SINE, short interspersed element; LINE, long interspersed element; Other, other types of repeat; Non, no repeats were detected. **c** Divergence rate of transposons of different sources. Deletion and insertion represent transposons associated with deletions and insertions, respectively. RepeatMasker was used to calculate the divergence rate from the consensus sequence of Repbase for each transposon. The middle hinge of all boxes is the median, the lower and upper hinges correspond to the 25th and 75th percentiles, and the whiskers represent the 1.5 inter-quartile range (IQR) extending from the hinges. Wilcoxon rank sum test, \*\*\* $p < 0.001$  and  $n = 435,202, 456,919, 6525, 4010$  for GDDH13, HFTH1, Deletion, Insertion, respectively. **d** Count of structural variations detected in each chromosome. Duplicated chromosomes are shown in the x-axis with the same colour. **e** Syntenic view of two large inversions with a length longer than 600 bp in chromosome 16 (19,782,372–20,450,041 and 31,388,523–32,032,862). Source data are provided in Source Data file 1

GDSVs (Fig. 2b). We found that the divergence rates of TEs displayed a significant difference ( $p$ -value  $< 2.2e-16$ ; Welch two-sample  $t$ -test), and that TEs associated with insertions were less divergent from the consensus TEs found in Repbase<sup>23</sup> and younger than the TEs associated with deletions (Fig. 2c), which suggests that most of the Golden Delicious shared insertions were induced by young TEs insertion in the HFTH1 genome, after Hanfu and Golden Delicious diverged from a common ancestor.

Taking a whole-genome view of the distribution of SNPs and GDSVs, we found chr.15 had the highest average SNP density (2.99 per kilobase, Fig. 1f), whereas chr.2 exhibited the most structural variations (Fig. 2d). Most of the duplicated

chromosomes from the same chromosome ancestor after a recent whole-genome duplication<sup>4</sup> (WGD, chrs.3 and 11, 9 and 17, 13 and 16) showed similar contents of SNPs and GDSVs, whereas chrs.10 and 5 showed significantly different contents of SNPs (2.52:1, Fig. 1f) and GDSVs (1.66:1, Fig. 2d), particularly in one end of chr.5, which showed a lower diversity than the average diversity of whole-genome. However, chrs.10 and 5 had a similar gene content (Fig. 1f), suggesting that chr.5 may have undergone introgression and fixation during domestication and breeding process after the recent WGD.

Additionally, we identified 14 large inversions (INVs) with length longer than 100 kb, two of which on chr.16 were longer

than 600 kb (Fig. 2e). We found that all the breakpoints of these two INVs were located within TE elements, which may have induced these inversion events.

**Dynamic evolution of LTR retrotransposons in Hanfu.** TEs play an important role in driving adaptive evolution<sup>7</sup>, and >75% of the TEs from the apple genome were LTR-RTs. To track the highly dynamic evolution of LTR-RTs, we identified 7313 intact LTR-RTs with an average length of 7868 bp (Fig. 3a) in the HFTH1 genome. Approximately 62% of these LTR-RTs were present in the Golden Delicious genome based on its different assemblies (Fig. 3b and Supplementary Fig. 9), suggesting that these LTR-RTs may have been inserted in the common ancestor of the Hanfu and Golden Delicious lines. In the Rosaceae family, pear and apple share a highly conserved genomes. However, because of the fragmented genome assembly of the pear genome (contig N50 of 33.76 kb), only 40 intact LTR-RTs were found in the pear genome<sup>26</sup> (Supplementary Fig. 9). Further analysis showed that most existing intact LTR-RTs in the apple genome may have been inserted after the divergence of apple and pear, which is in accordance with the findings that the average insert time (0.8 MYA) of these LTR-RTs was notably less than the estimated divergence time (8.1 MYA) of apple and pear (Fig. 3c and Supplementary Fig. 10). More than half of the shared LTR-RTs were highly similar between the HFTH1 and GDDH13 genomes (identity  $\geq 0.99$ , Fig. 3b), and the average cumulative nucleotide substitutions (CNS) of the shared LTR-RTs and their flanking regions was less than the average level across the whole-genome (Fig. 3d) and most of the reverse-transcriptase domains (*gag*, *pol*, and *env*) of the shared LTR-RTs were not expressed in the ten tested samples (Supplementary Fig. 11a).

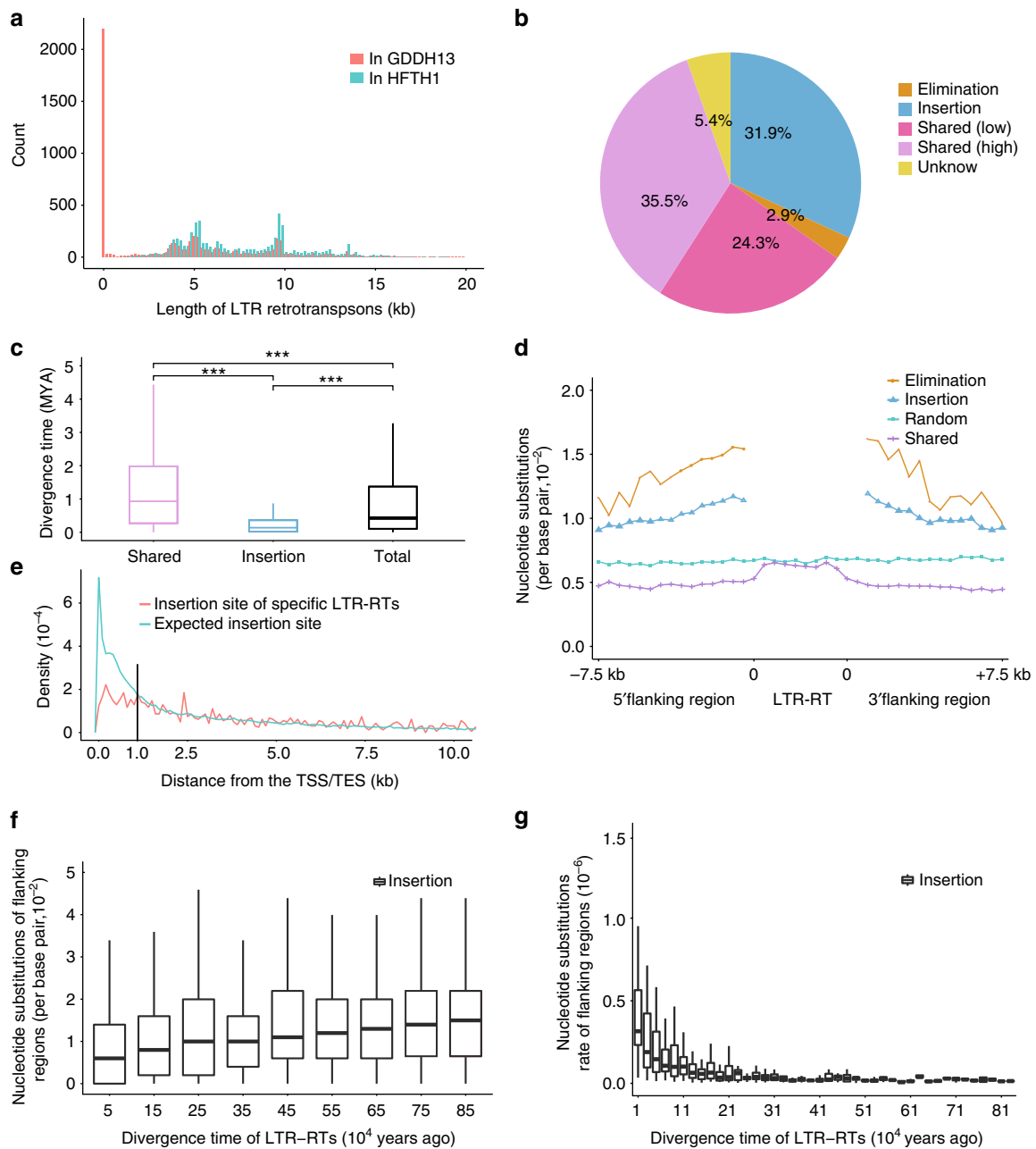
In addition, specific LTR-RTs (~31.9% of the total LTR-RTs) with two short target site duplications (TSDs) on their both sides, were only found on the HFTH1 genome (Fig. 3b), but only one of two TSDs was found in the corresponding position of the GDDH13 genome, suggesting that these specific LTR-RTs were inserted after the divergence of Hanfu and Golden Delicious. These specific insertion events, especially those near or inside genes, may be under strong selection for fixing, because protein-coding genes (coding plus intron regions) can account for 22.22% of whole-genome, but only 12.05% of the insertion events were located inside genes. The observed insertion events near genes were also lower than expected (Fig. 3e). Besides, the average expression level of genes near these insertion sites was less than that of the total gene set (Supplementary Fig. 11b). However, the selection pressure may be weakened or lost when the insertions located at least 1 kb from a gene (Fig. 3e). We found that 82.54% of specific LTR-RTs were expressed at least in one tested samples, compared with 64.39% of shared LTR-RTs ( $p$ -value <  $2.2e-16$ ; chi-squared test), which indicated that most specific LTR-RTs were young and active. The insertion events not only affect nearby gene expression, but also increase the mutation rate in the vicinity of the insertion site (~1.4- to 1.8-fold higher than the rate of the global nucleotide substitutions), which was in agreement with prior investigations<sup>27,28</sup>. Furthermore, the data also showed that the CNSs decreased as the distance from an LTR-RT increased; whereas the average CNSs in sequences surrounding shared inactive LTR-RTs was substantially lower (Fig. 3d). We found that the average CNSs increased gradually as the divergence time of LTR-RTs increased (Fig. 3f). Furthermore, the average CNSs rate appeared to slow and reach a bottleneck as the transposons were gradually inactivated and became other non-functional sequences (Fig. 3g). This result was consistent with the finding that most of the CNSs in the shared LTR-RTs were less than those of the specific LTR-RTs (Fig. 3d) even

though the shared LTR-RTs had a greater evolutionary age (Fig. 3c). In addition, we found that ~2.9% of the LTR-RTs in the HFTH1 genome may have been eliminated in the GDDH13 genome, because 66.27% of these LTR-RTs had a length shorter than 100 bp at the corresponding site in the GDDH13 genome. In addition, the others had no blast hits (BLASTN  $e$ -values  $\leq 10$ ) to the corresponding LTR-RT sequence in the HFTH1 genome, and most of these sequences appear to be the remains of the transposition occurrence of other TEs in the GDDH13 genome (Fig. 3b). The average CNSs of the sequences surrounding these eliminated LTR-RTs was the highest in the genome (Fig. 3d), suggesting elimination events might have made a greater impact than insertions did during the evolution of TEs. The dynamic evolution of TEs likely created the unique genetic and phenotypic characteristics of HFTH1.

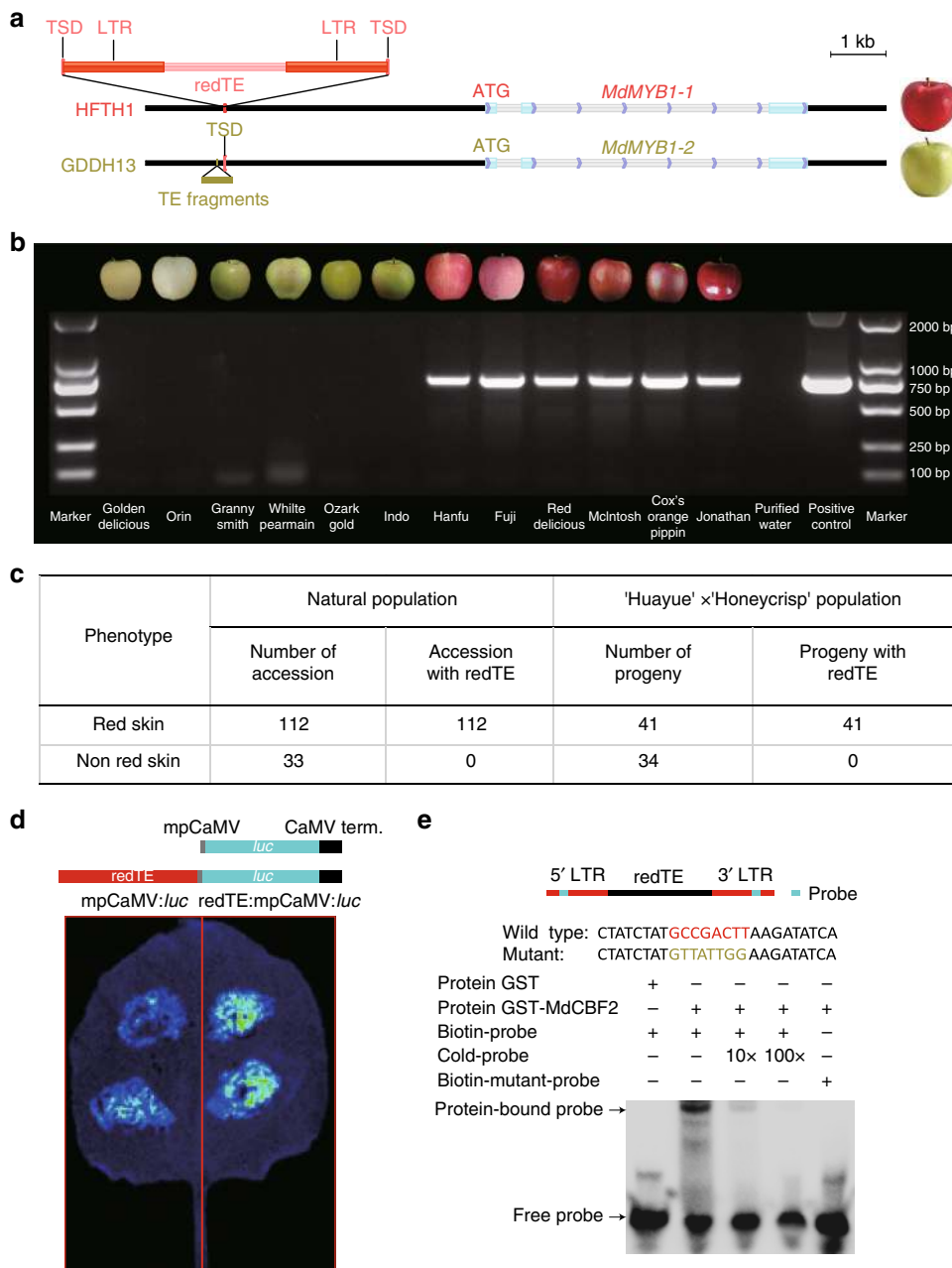
**Retrotransposon as an enhancer of *MdMYB1* expression.** To evaluate the potential of our genome assembly for the genetic dissection of agriculturally important traits, we focused on red phenotype of apple fruit, which is a key determinant of consumer preference<sup>12,13</sup>. Genetic evidence has confirmed that *MdMYB1*, a core transcriptional regulator of the anthocyanin biosynthesis pathway, is responsible for the red phenotype<sup>11,13,29</sup>. In this study, we detected a significant difference between the transcript levels of *MdMYB1* and anthocyanin-related structural genes in the skins of ripening fruits of Hanfu and Golden Delicious by quantitative reverse transcription PCR (qRT-PCR) (Supplementary Fig. 12). *MdMYB1* had at least three types of alleles, namely, *MdMYB1-1*, *MdMYB1-2* and *MdMYB1-3*, and *MdMYB1-1* was a single dominant allele controlling anthocyanin synthesis in apple skin. The *MdMYB1-2* and *MdMYB1-3* alleles in non-red-skinned cultivars show a limited expression under intense light and low-temperature<sup>12</sup>. Furthermore, it has been shown that the coding region differences of these alleles do not affect their functional activity<sup>13</sup>. The reason for the significant differences in the expression levels among *MdMYB1* alleles has not been fully elucidated.

The availability of two high-quality genomes allows a more precise comparative analysis. Specifically, we performed a sequence alignment of *MdMYB1* in the HFTH1 and GDDH13 genomes. The results showed that the coding sequences of *MdMYB1* were identical, but one SNP was detected in the intron regions. In addition, fifteen SNPs and five indels in the upstream region were found. Among of five indels, one 501 bp insertion occurs in the GDDH13 at a distance of -3394 bp upstream of the ATG initiation codon of *MdMYB1*, and is highly divergent from the neighbouring LTR-RTs (Fig. 4a); Another 4097 bp insertion is present in the HFTH1 at a distance of -3297 bp upstream of the ATG initiation codon of *MdMYB1*, and is a gypsy-like LTR retrotransposon (denoted redTE) with two TSDs (CATAT, Fig. 4a). Its two flanking LTR sequences (1274 bp) were completely identical, indicating that it was a recent insertion event. Although at least 3913 intact Gypsy-like retrotransposons were identified in the HFTH1 genome, only one of them had a 96.26% global identity with redTE (redTE-like, Supplementary Fig. 13), and the others showed <75.10% identity. Interestingly, redTE-like was only found in the HFTH1 genome, and not in GDDH13 through a genome-wide scan, and its two flanking LTR sequences (1262 bp and 1298 bp) harboured more mutations, suggesting that redTE-like was older than redTE.

To determine whether these differences were associated with fruits skin colour, we retrieved the sequences of these differential loci using data from public databases (Supplementary Table 10). The results showed that sixteen SNPs, two small indels in HFTH1 were also present in non-red-skinned accessions, and one 2-bp



**Fig. 3** Evolution of intact LTR retrotransposons in the HFTH1 genome. **a** Length distribution of intact LTR-RTs from the HFTH1 genome. A large number of LTR-RTs in the GDDH13 genome have a length <100 bp, which indicated that these LTR-RTs were newly inserted into the HFTH1 genome or eliminated from the GDDH13 genome. **b** Pie chart showing the classification of the LTR-RTs in the HFTH1 genome (see Methods for the definition of each type). **c** Estimated divergence time of each category of LTR-RTs in the HFTH1 genome. The divergence time was calculated using a substitution rate of  $1.3 \times 10^{-8}$  substitutions per site per year. Wilcoxon rank sum test,  $***p < 0.001$  and  $n = 3468, 1850, 5800$  for Shared, Insertion, Total, respectively. **d** Distribution of cumulative nucleotide substitutions of LTR-RTs and flanking sequences. The nucleotide substitutions of both flanking regions were calculated in a 400 bp sliding window, and the nucleotide substitutions of LTR-RT inside regions were calculated in ten non-overlapping equal bins. Three thousand blocks in the HFTH1 genome with a length 20 kb were randomly selected and used as a control. **e** Distribution of insert sites of specific LTR-RTs in the GDDH13 genome in 100-bp windows located at an increasing distance from genes. The distribution of intergenic distance was used as a control. **f** Distribution of cumulative nucleotide substitutions of flanking sequences over the divergence time of LTR-RTs. The flanking sequences were defined as 500 bp fragments located the upstream and downstream of LTR-RTs. The cumulative nucleotide substitutions were calculated in a sliding window of 100,000 years. **g** Rate of cumulative nucleotide substitutions of flanking sequences relative to the divergence time of LTR-RTs (see Methods for details). The middle hinge of all boxes is the median, the lower and upper hinges correspond to the 25th and 75th percentiles, and the whiskers represent the 1.5 inter-quartile range (IQR) extending from the hinges. Source data of Fig. 3a, b, c and e are provided in Source Data file 1. Source data of Fig. 3d, f and g are provided in Source Data file 2



**Fig. 4** Red phenotype of apple associated with an LTR retrotransposon. **a** Molecular structure of *MdMYB1-1* and *MdMYB1-2* alleles with flanking sequences. The insertion sites upstream of *MdMYB1-1* and *MdMYB1-2* are indicated by a red line (HFTH1) and golden yellow line (GDDH13), respectively. **b** Images of 12 well-known apple varieties with non-red or red skin colour (upper panel) and PCR-based screen showing the absence (right) or presence (left) of the LTR retrotransposon insertion in the upstream of *MdMYB1*. A 750 bp fragment corresponding to the partial of redTE that is absent in non-red-skinned varieties (lanes 1 to 6) and is present only in red-skinned varieties (lanes 7 to 12). Lane 13, control check (purified water was used as the template), Lane 14, positive control. **c** PCR-based analysis of the redTE insertion in 145 accessions and the F1 segregating population from the cross of Huayue x Honeycrisp. **d** Constructs and transient expression assays showing that redTE obviously enhanced the luciferase expression levels. Upper panel, the mpCaM:luc (up) and redTE:mpCaM:luc (down) construct backbone consists of the minimal promoter from the cauliflower mosaic virus (mpCaMV, grey box), luciferase ORF and cauliflower mosaic virus terminator (black box). Lower panel, Luciferase image of *Nicotiana benthamiana* leaves 72 h after infiltration with the Agrobacterial strains containing mpCaM:luc (left), and redTE:mpCaM:luc (right), respectively. **e** MdCBF2 binds directly to the cis-acting element GCCGACTT. Source data of Fig. 4b, d are provided in Source Data file 1

indel located in 17 repeat T bases in HFTH1 have a complex polymorphism in red apple accessions. In addition, the 501 bp insertion in the GDDH13 genome was also found in red-skinned accessions by PCR verification (Supplementary Fig. 14). Subsequently, we investigated the association of redTE with red skin colour, because it has been reported that retrotransposons often play crucial roles in tissue-specific expression patterns of pigment

genes in plants<sup>30</sup> and animals<sup>31</sup>. First, we retrieved resequencing data of the cultivars from GDR (<ftp.bioinfo.wsu.edu>, Supplementary Table 11) using the junction sequence GGATTTTATAT ATGTGTTGACCCTA of redTE. The results showed that this target junction sequence were able to be found from the data of only red-skinned cultivars. Next, we screened 112 red-skinned accessions and 33 non-red-skinned accessions with known



phenotypes using a PCR marker specific to redTE. The results showed that all the tested red-skinned accessions were completely associated with redTE, but redTE was not found in the non-red-skinned accessions (Fig. 4b, Supplementary Fig. 15 and Supplementary Data 1), suggesting that redTE insertion may be responsible for red phenotype in apple. Subsequently, we screened 75 progenies, including 41 with red skin and 34 without red skin, from the cross of Huayue (non-red skin) and Honeycrisp (red skin) (Fig. 4c, Supplementary Fig. 16 and Supplementary Data 2). The progenies analysis of this cross also confirmed the perfect co-segregation of redTE with red phenotype. These results suggested that redTE insertion was responsible for red phenotype in apple, which further supported the hypothesis of a single dominant mutation model underlying skin colour. Subsequently, because red-fleshed apple usually also have a red skin, we tested whether they also possess redTE insertion. To that end, three red-fleshed accessions with homozygous R6 genotypes were analysed by specific PCR marker, and the results showed that red-fleshed apples did not harbour this retrotransposon insertion (Supplementary Fig. 15 and Supplementary Data 1), suggesting that the red skin of red-fleshed apples is caused by the constitutive expression of *MdMYB10* binding to R6 motifs of its own promoter in an auto regulatory-loop manner<sup>8</sup>, independent of redTE.

Some studies have shown that TEs are an abundant source of enhancer activity in plants<sup>32</sup>. Based on its inserted position, redTE appears to provide local enhancer activities that modulate the light-responsive gene expression of the *MdMYB1*. To definitively assess whether redTE acts as an enhancer, we performed transient assays in *Nicotiana benthamiana* leaves to test the effects of redTE on the gene expression of firefly luciferase. The results showed that the construct with redTE led to a significant increase in reporter gene expression, relative to the construct with the minimal promoter alone (Fig. 4d). This observed enhancement of gene expression by redTE is consistent with the known higher level of *MdMYB1* expression in red apple, which is similar to the role of a functional transposon (hopscotch) as a long-distance enhancer of *tb1* gene expression in maize<sup>33</sup>.

In addition, many red cultivars originating from bud sports with different red patterns, such as Hanfu and HanM, Gala and GaleGala, etc., are pretty commonly used in apple production<sup>34</sup> (Supplementary Fig. 15 and Supplementary Data 1). These cultivars and their bud-sport mutations have redTE in our studies. It is well-known that transposon-induced epigenetic changes often affect the differential expression of neighbouring genes and create novel patterning<sup>35,36</sup>. Additionally, a recent research showed that striped pigmentation of Honeycrisp apple fruits is associated with hyper methylation in the promoter of *MdMYB1*<sup>14</sup>. Therefore, we selected the fruits of Hanfu (red stripe) and its sports HanM (fully red) at the same time (Supplementary Fig. 17a), and detected the DNA methylation status of redTE and the promoter region of *MdMYB1* using McrBC-PCR. The results showed that the Hanfu DNA was methylated in the MR3 and MR7 regions, in agreement with previous studies<sup>14,34</sup>. The redTE was heavily methylated in the MR8-MR11 regions, while the degree of methylation in Hanfu is much higher than that of HanM in the region MR12 (Supplementary Fig. 17b), which indicated that redTE-induced epigenetic changes may be associated with the variable colour patterns.

Given that redTE may control *MdMYB1* expression in red apple, we aimed to identify regulatory networks of redTE and other transcription factors under different environmental stresses. In blood orange, a retrotransposon controls the fruit-specific accumulation of anthocyanins in response to cold stress<sup>30</sup>. In apple, a relatively low ambient temperature can also promote

fruits coloration, which is presumably regulated via the recruitment of cold acclimation-related transcription factors. In particular, redTE contains the core cis-acting element (GCCGACTT) for cold acclimation transcription factor DREB/CBF binding<sup>37</sup>. We selected the low-temperature-inducible transcription factor MdCBF2 for electrophoretic mobility shift assay (EMSA)-based binding analysis. The result confirmed that MdCBF2 was capable of binding to the GCCGACTT element (Fig. 4e), which indicated it is potentially involved in fruit coloration through redTE regulatory networks under a relatively low ambient temperature.

## Discussion

A high-quality genome assembly is valuable for identifying structural variants, integrating phenotype-genotype associations, resulting in insights into the mode and tempo of genome evolution and elucidating the genetic architecture of important traits<sup>9,10</sup>. The generation of our reference genome provides a high-quality complement to the GDDH13 genome, demonstrating the utility of both whole-genome sequences for the accurate identification of the large and complex SVs, and providing a basis for the comparative genomic investigation of the unique biological characteristics and intraspecific genome diversity in apple. Our comparative genomic results indicate that dynamic changes in TEs can lead to large amounts of SVs, which might impact the genotypes. The discovery of redTE, a locus-specific LTR-RT insertion in the HFTH1 genome, will inspire researchers to further discuss the importance of TEs as a creator of major phenotypic variations. In fact, genome-scale analyses of TE-induced effects are only just beginning to be explored in humans and other crop species<sup>32</sup>. Therefore, the annotation of TEs in the HFTH1 and GDDH13 genomes will allow detailed studies of the functional effects and dynamic activity of more TEs polymorphisms in the long-term evolutionary background among different apple genotypes.

The evolution of fruits colour in apple is important and intriguing. In this study, our finding of a redTE insertion upstream of the *MdMYB1* promoter was associated with red colouration. Thus, redTE was regarded as an enhancer, controlling the development of red colouration by lowering the threshold value of the light response. Cultivars, without this enhancer, fail to effectively produce anthocyanin. This phenomenon could well explain why non-red-skinned cultivars still show a faint red colouration under intense light. In addition, red apple fruits display rich skin colours ranging from stripe to blush to dark red, and these diverse phenotypes are putatively the result of that redTE-mediated control the distribution patterns of anthocyanin through the creation genetic and epigenetic alleles to manipulate the function of *MdMYB1* under natural conditions. However, in apple, the mechanism through which redTE mediates how genetic and epigenetic variation contributes to red phenotypic diversity and adaptation to light changes remains to be investigated. Furthermore, site-specific TEs can also be used to trace the genetic relationships and origins<sup>38</sup>. A study of worldwide genetic diversity and pedigree records in apple estimated that one yellow-skinned cultivar Golden Delicious (1916) and four red-skinned cultivars [Cox's Orange Pippin (1850), Red Delicious (1880), Jonathan (1826) and McIntosh (1870)] were the core founders of modern apple breeding<sup>39,40</sup>. These four red-skinned cultivars contain redTE insertion (Fig. 4b), which suggests that they also had a common parental origin. Notably, the accession *M. sieversii* also contains redTE, suggesting the redTE likely originated from its supposed primary wild ancestor in Xinjiang (Supplementary Data 1). Although, a recent genome resequencing revealed *M. sieversii* in Xinjiang, China, is an ancient isolated ecotype not

directly contributing to apple domestication<sup>6</sup>, it's possible that early human activities spreaded *M. sieversii* with eye-catching red skin, from Xinjiang, China, to other geographical areas. The Chinese soft apple Huacaiping that was domesticated from Xinjiang wild apple with >2000-year cultivation history in China<sup>41</sup>, harbours redTE, which indicates that red apple once spreaded westward and eastward along the old Silk Road. This result appears to agree well with the results of breeding and geographical investigations of the origins and history of domesticated apple<sup>41,42</sup>. Subsequently, red apples from the redTE-induced mutation ancestor have become increasingly popular by artificial selection. Interestingly, a recent study of blood oranges<sup>30</sup> showed that two different TE insertions produced similar effects in controlling the cold-induced expression of Ruby, which modulates fruit colour. This finding raises the possibility that other TE insertions are present in apple whose colouring effects have not been characterised. Here, it remains to be seen whether an analogous redTE insertion enhances MYB transcription in pear and other fruits in the Rosaceae.

Additionally, owing to the long and laborious selection process of apple breeding, the ultimate goal of assembled genome is to serve as a guideline in developing tools for MAS breeding in apple. We identified a surprisingly abundant number of structural variations that were dispersed across the whole-genome, which will be highly useful in developing functional markers for apple breeding. For instance, the redTE-based specific marker, is a particularly valuable tool for pre-selection of hybrid seedlings with objective skin colour in apple breeding programmers, because it is more efficient and precise than the previously available markers<sup>11,13,43</sup>. The marker may greatly reduce the costs of apple breeding by eliminating a large number of non-target hybrid seedlings. In addition, this genome of HFTH1 from a resistant parent may provide an excellent basis for development of markers of cold-resistance and disease-resistance, such as branch ring-rot and alternaria leaf spot in apple breeding<sup>44</sup>. Overall, this near-complete genome and other genomic resources will aid the mining of genes and functional markers and support the translation of research findings into genetic improvements for sustainable apple production.

## Methods

**Plant materials and DNA sequencing.** The homozygous line HFTH1 was derived from an in vitro anther culture of a widely grown cultivar (HFP) in the cold region of northern China. HFTH1 and its donor parent, grafted on GM256 rootstock in 2010, were grown in the greenhouse and field at the Research Institute of Pomology, Chinese Academy of Agricultural Science. The DNA of HFTH1 was extracted from young leaves using the phenol-chloroform method. Two libraries with insert sizes of 300 bp and 20 kb were separately constructed using Illumina TruSeq Nano DNA Library Prep Kits and SMRTbell Template Prep Kits, and the 300 bp and 20 kb libraries were subsequently sequenced using an Illumina HiSeq X Ten instrument and a PacBio RS II instrument with the P6-C4 sequencing reagent, respectively. To obtain the BioNano optical mapping data, DNA was extracted from fresh young leaves of HFTH1, and embedded in a thin agarose layer for labelling at Nt.BspQI sites using the IrysPrep Reagent Kit protocol and subjected to optical scanning on the BioNano Irys platform. The Hi-C library, including cellular crosslinking, chromatin digestion, labelling of DNA ends, DNA ligation, purification and fragmentation, was constructed using the standard procedure as follows. Firstly, nuclear DNA from young leaves was cross-linked in situ, extracted, and digested with a restriction enzyme. The sticky ends of the digested fragments were biotinylated, diluted, and then ligated randomly. The biotinylated DNA fragments were enriched and sheared again to generate sequencing library, which was subsequently sequenced on Illumina HiSeq 4000 (<http://en.annoroad.com/>).

**Genome assembly.** Falcon<sup>45</sup> (v0.4) was used for constructing initial contigs using the following parameters: `length_cutoff = 13,000 length_cutoff_pr = 14,000 pa_DBSplit_option = -x1000 -s250 -a pa_HPCdaligner_option = -v -dal128 -t12 -e.75 -k20 -h320 -l1800 -s1000 falcon_sense_option = --output_multi --min_idt 0.75 --min_cov 2 --local_match_count_threshold 2 --max_n_read 400 --output_dformat ovlp_DBSplit_option = -x1000 -s200 ovlp_HPCdaligner_option = -v -dal100 -t12 -k18 -h280 -e.96 -l1800 -s1000 overlap_filtering_setting = --max_diff 50 --max_cov 80 --min_cov 2 --bestn 10`. The initial polishing was performed with

Quiver<sup>46</sup> using PacBio-only long reads, and then Pilon<sup>47</sup> (v1.20) was utilised to further correct the PacBio-corrected contigs with accurate Illumina short reads. The BioNano data was first assembled to a consensus map using the IrysView software with a molecular length threshold of 150 kb and a minimum labels per molecule of 8, and hybrid scaffolding of the PacBio-corrected contigs and BioNano-based consensus map was performed using the hybrid scaffolding module within IrysView software with manufacturer's suggested parameters. After scaffolding, PBjelly from PBSuite<sup>48</sup> (v14.9.9) was performed to close gaps in the hybrid assembly. We re-performed error correction procedures to polish the sequences in the gap regions. Subsequently, the mitochondrial and chloroplast scaffolds or contigs were removed through alignment to mitochondrial and chloroplast references of apple, and any scaffolds or contigs for which at least 80% of the total length was aligned and that showed an identity larger than 90% were discarded as mitochondrial or chloroplast sequences. The Hi-C sequencing data were first aligned to the assembled genome using the bowtie2 end-to-end algorithm<sup>49</sup>, then the assembled scaffolds are clustered, ordered and directed onto the pseudo-chromosomes using Lachesis<sup>50</sup>. Finally, the pseudo-chromosomes predicted by Lachesis were cut into bins with equal length of 100 kb and used to construct a heatmap based on the interaction signals that generated by valid mapped read pairs to do validate and correct manually.

To obtain the genome of two organelles, we first used BLASR to map all raw PacBio long reads to the organelle references of *Malus* (downloaded from GenBank, accession NC\_018554, NC\_031163, KU851961, KX499859 and KX499861), any reads with a length longer than 20 kb, for which at least 80% of the total length were aligned and that showed an identity greater than 70% were used for the next assembly. The chloroplast genome was assembled with Canu<sup>51</sup> (v1.3) (genome size = 160k) and the mitochondrial genome was assembled with Falcon (v0.4), the following error correction was then performed with the above-described procedures using Quiver and Pilon.

**Annotation of repeats.** The repetitive sequences, including tandem repeats and TEs, in the HFTH1 and GDDH13 genomes were searched. First, we used Tandem Repeats Finder<sup>52</sup> (TRF, v4.09) to annotate the tandem repeats using the following parameters: `2 7 7 80 10 50 2000`. Then TEs were identified at both the DNA and protein levels using a combination of de novo and homology-based approaches. At the DNA level, LTR\_FINDER<sup>53</sup> (v1.0.6) was first used to identify LTR-RTs and RepeatModeler<sup>54</sup> (v1.0.5) was utilised to construct a de novo repeat library, which comprised a repeat consensus database with classification information. We employed RepeatMasker<sup>54</sup> (v4.0.6) to search for similar TEs in the known Repbase TE library<sup>23</sup>, MIPS Repeat Element Database<sup>55</sup> (v9.3) and de novo repeat library. At the protein level, RepeatProteinMask within the RepeatMasker package was used to search against the TE protein database using a WU-BLASTX engine.

Telomere sequences were identified by searching both ends of the pseudo-chromosomes for high copy number repeats with the repeat unit 5-TTTAGGG-3. Putative heterochromatin regions were identified by searching for 1 Mb windows (250 kb step) with >90% repeat elements and the adjacent centromere windows were merged.

**Gene prediction and annotation.** The MAKER pipeline<sup>56</sup> (v2.31.8), which incorporates ab initio prediction, homology-based prediction and RNA-Seq assisted prediction, was used to annotate gene models. The protein sequences used for homology-based prediction were from five sequenced plants, namely, *Arabidopsis thaliana* (<https://www.arabidopsis.org/index.jsp>), *Prunus persica* (<https://www.rosaceae.org/>), *Pyrus communis* (<https://www.rosaceae.org/>), *Malus domestica* (GDDH13, <https://iris.angers.inra.fr/gddh13/the-apple-genome-downloads.html>), and *Fragaria vesca* (<https://www.rosaceae.org/>) and a total of 1440 benchmarking universal single-copy orthologues of embryophyta within the BUSCO software (v3.0.1), were initially mapped onto the HFTH1 genome using tBlastn. Subsequently, Exonerate<sup>57</sup> (v2.2.0) were used to polish the BLAST hits and thereby acquire exact intron/exon positions. The transcriptional data, including three tissues of the HFTH1 and seven tissues from HFP, were assembled with Histat2<sup>58</sup> (v2.05) and StringTie<sup>59</sup> (v1.3.0), and the results were used to identify candidate exon regions, donor and acceptor sites. The repeat regions in the HFTH1 genome were first soft-masked, and MAKER was then run twice. First, we ran MAKER with only transcriptional data to generate imperfect gene models, which were used to train the parameters for SNAP<sup>60</sup> (V2006-07-28) and Augustus<sup>61</sup> (v3.2.2). All data and predictions were then used to produce a consensus gene set. We removed a gene model from the consensus gene set if it had a MAKER-defined annotation edit distance (AED) score higher than 0.5 and lacked transcript data or homologous protein support. To obtain a more complete gene set, we added the genes predicted by Augustus that had transcriptional data or homologous proteins support but were not included in the initial gene set, to the final gene set.

Gene functions were assigned according to the best match by aligning the protein sequences to the Swiss-Prot and TrEMBL databases<sup>62</sup> using Blastp (with a threshold of  $E$ -value  $\leq 1e^{-5}$ ). The motifs and domains were annotated using InterProScan<sup>63</sup> (v5.24) by searching against publicly available databases, including ProDom, PRINTS, Pfam, SMRT, PANTHER and PROSITE. The Gene Ontology (GO) IDs for each gene were assigned according to the corresponding InterPro entry.

**RNA sequencing and data analysis.** RNA was extracted from seven tissues of HFP and three tissues of HFTH1 using the Quick RNA Isolation Kit (Cat.No.046-50QK, Huayueyang Biotechnology Beijing Co.Ltd., <http://www.huayueyang.com/>) and then characterised by agarose gel electrophoresis and a Nano Drop ND1000 spectrophotometer (Nano Drop Technologies, Wilmington, DE, USA). The complementary DNA (cDNA) libraries were constructed as follows, the first step involves purifying the poly-A containing mRNA molecules using oligo-dT attached magnetic beads. Then, the mRNA is fragmented into small pieces using divalent cations. The cDNA were synthesised using the cleaved RNA fragments as templates. The cDNA fragments then go through an end repair process, the addition of a single A base, and then ligation of the adapters. The products are then purified and enriched with PCR to create the final cDNA library, and are sequenced on an Illumina HiSeq 4000. RNA reads were first mapped to the HFTH1 genome using HISAT2, and gene expression was then measured in fragments per kilobase of exon per million fragments mapped (FPKM) using StringTie. The counts of the mapped reads of LTR-RTs were calculated with BEDTOOLS<sup>64</sup> (v2.23.0).

**Phylogenetic analysis.** The sequences of protein-coding genes from HFTH1 and seven plants (*A. thaliana*, *F. vesca*, *R. occidentalis* (<https://www.rosaceae.org/>), *P. mume* (downloaded from GeneBank, GCA\_000346735.1), *P. persica*, *P. communis* and *M. domestica*) were used for a gene family clustering analysis. First, Blastp was used to generate pairwise protein sequence with an *E*-value cutoff of  $1e^{-5}$ . Second, OrthoMCL<sup>65</sup> (v2.0.9) was used to cluster genes with an inflation value of 1.5. The protein sequences from 1499 single-copy gene families found in more than eight species were extracted and aligned using MAFFT<sup>66</sup> (v7.058), and the alignment was then back-translated to the nucleotide alphabet using PAL2NAL<sup>67</sup> (v14). The poorly aligned positions and divergent regions of the alignment were eliminated using Gblocks (v0.91b, [molevol.cimima.csic.es/castresana/Gblocks.html](http://molevol.cimima.csic.es/castresana/Gblocks.html)). Phylogenetic analysis was performed using a maximum likelihood (ML) method implemented in RaxML<sup>68</sup> (v8.0.19) with the GTRGAMMA substitution model and 100 nonparametric Bootstrap replicates. *A. thaliana* was selected as the out-group. The divergence time for eight species was estimated based on fourfold degenerate sites from the filtered alignment. The Markov chain Monte Carlo algorithm for Bayes estimation was adopted to estimate the divergence time using MCMCTree within the PAML package<sup>69</sup> (v4.6). The calibration time for the divergence between *Rosaceae* and *A. thaliana* (97~109 Mya) was obtained from the TimeTree database (<http://www.timetree.org/>).

SynMap (CoGe, <http://www.genomeevolution.org>) was used to detect the conserved syntenic blocks using homologous gene pairs with the following parameters: Maximum distance between two matches (-D): 20; Minimum number of aligned pairs (-A): 10; Algorithm (Quota Align Merge) with maximum distance between two blocks (-Dm): 500. Circos<sup>70</sup> was used for visualisation.

**Gap filling.** A modified protocol based on the method reported by Shi et al.<sup>24</sup> was used. Briefly, using BEDTOOLS, gap regions were extracted from the GDDH13 genome, and adjacent gaps (distance  $\leq 500$  bp) were merged as one gap. The 500 bp fragments located upstream and downstream of each gap region were extracted and aligned to the HFTH1 genome using BWA<sup>71</sup> with parameter -a. A gap was considered to be closed, if (i) both fragments aligned successfully (coverage  $\geq 80\%$ ) within 500 kb on the same scaffold or chromosome with the same orientation; (ii) the total alignment positions of both fragments were  $< 5$  (both fragments in repeat regions were avoided); (iii) the intervening sequence between aligned positions of both fragments did not contain any ambiguous (N) bases (only consider closed gaps); (iv) the intervening sequences can be aligned successfully (coverage  $\geq 80\%$ ) with any Golden Delicious genomes or  $> 90\%$  of the intervening sequence with effectively covered (depth  $> 3$ ) using 30-fold Illumina reads from GDDH13, the intervening sequences were polished using Pilon with the Illumina reads of GDDH13. The ambiguous (N) bases of gap regions in the GDDH13 genome were replaced with the polished intervening sequences. The gaps in the HFTH1 genome were filled with the GDDH13 genome using the same above-described method.

**Detection and analysis of genome variation.** SNP detection was performed using BWA and Genome Analysis Toolkit<sup>72</sup> (GATK, v3.8) with the following filtering options: Quality by depth (QD)  $> 2.0$ , Fisher strand (FS)  $> 60.0$ , RMS mapping quality (MQ)  $< 40.0$ , MQRankSum  $< -12.5$ , ReadPosRankSum  $< -8.0$ , Maximum depth (DP)  $> 360$ . An SNP was defined as Golden Delicious shared SNP if the SNP was detected in the Golden Delicious genomes of GDDH13, Velasco et al. and Li et al. but not detected in the HFP genome. Owing to the lack of Illumina reads in the genome reported by Velasco et al., we extracted 500 bp fragments located upstream and downstream of each SNP and aligned these to the genome reported by Velasco et al. to assess whether this SNP was shared. A functional analysis of SNPs was performed using the ANNOVAR software<sup>73</sup>.

Structural variants (SVs) were identified using BWA and Sniffles<sup>74</sup> (v1.0.7). First, we mapped the PacBio-corrected long reads (corrected with FALCON during the assembly step) of HFTH1 to the GDDH13 genome using BWA-MEM (using the -M and -x parameters), and Sniffles was then used to identify indels with length  $> 100$  bp and large inversions with length  $> 100$  kb. To reduce false-positive SV results as much as possible, we extracted sequences from 500 bp upstream to

500 bp downstream of each indel from the GDDH13 genome and aligned them to the HFTH1 genome using BWA-MEM with the parameter -a; if the fragment was aligned successfully and included a large insertion or clip (corresponding to a deletion of SV), and a large deletion or gap (corresponding to an insertion of SV), this SV was retained in the final result, and the border and length of this SV were recalculated based on the alignment. Long inversions were confirmed by whole-genome alignment using MUMmer<sup>75</sup> (v3.07).

Presence/absence variations (PAVs) were detected by whole-genome alignment using MUMmer with the parameter -maxmatch (using the GDDH13 genome as the reference). Using BEDTOOLS, we merge adjacent aligned blocks (distance  $\leq 50$  bp) and extracted unaligned regions with lengths longer than 100 bp. For some unaligned regions with ambiguous (N) bases, we removed or split these regions according to the positions of the N bases. These unaligned regions were further filtered using their average depth. First, BWA was used to map the Illumina short reads from GDDH13 and HFTH1 to the HFTH1 and GDDH13 genomes, respectively. The depth was calculated using SAMTOOLS<sup>76</sup> (v1.2). If an unaligned region had an average depth  $< 10\%$  of the average depth of the whole-genome, this region was regarded as a PV. The PVs in GDDH13 were treated as deletions, and the PVs in HFTH1 were treated as insertions to ensure consistency with indels for further analysis.

**Detection and analysis of LTR-RTs.** The LTR\_retriever pipeline<sup>77</sup> was conducted to identify intact LTR-RTs of the HFTH1 genome from the outputs of LTRharvest<sup>78</sup> (parameters: -similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1) and LTR\_FINDER (Parameters: -D 15,000 -d 1000 -L 7000 -l 100 -p 20 -M 0.9) with default parameters. Owing to the high copy number of LTR-RTs, we extracted 500 bp fragments in the upstream and downstream of each LTR-RT. We first aligned both fragments back to the HFTH1 genome using BWA-MEM with parameter -a. If both fragments were uniquely aligned within 20 kb (the length of the longest LTR-RT) on the same scaffold and chromosome with the same orientation, this LTR-RT was retained (31 LTR-RTs were removed). We then aligned both fragments to the GDDH13 genome with the same parameters and filter criterion (1482 LTR-RTs were removed), and 5800 LTR-RTs were used for further analysis. The intervening sequences between the aligned positions of both fragments in the GDDH13 genome were aligned to the corresponding LTR-RT sequence in the HFTH1 genome using Blastn with parameter -F. Different types of LTR-RT were distinguished using the following criteria: if (i) the length of the intervening sequence was  $> 100$  bp and existed blast hits to the corresponding LTR-RT sequence in the HFTH1 genome, this LTR-RT was defined as Shared. The Shared type was further classified as high similarity (similarity  $\geq 99\%$ ) and low similarity (similarity  $< 99\%$ ) with the global similarity; (ii) if two copies of the TSDs flanked the LTR-RT in the HFTH1 genome while only one TSD (no intervening sequences) existed at the corresponding site of the GDDH13 genome, this LTR-RT was defined as Insertion; In contrast, (iii) if the intervening sequence had no blast hits to the corresponding LTR-RT sequence in the HFTH1 genome and two or no TSDs were found at the corresponding site of the GDDH13 genome, this LTR-RT was defined as Elimination; (iv) other remaining LTR-RTs were defined as Unknown.

The cumulative nucleotide substitution rate (*V*) of flanking sequences of each LTR-RT was calculated as  $V = (S_f - S_c) / T_{\text{itr}} + S_c / (2 \times T_{\text{div}})$ , where  $S_f$  is the cumulative nucleotide substitution frequency of flanking sequence, and it is equal to the number of cumulative nucleotide substitutions in the flanking sequence divided by the length of flanking sequence (500 bp);  $S_c$  is the cumulative nucleotide substitution frequency of the control region (500 bp) that was defined as a non-functional region (2 kb) with 9 kb away from the LTR-RT;  $T_{\text{itr}}$  is the divergence time of LTR-RT;  $T_{\text{div}}$  is the divergence time between Golden Delicious and Hanfu.

**Amplification of the partial redTE and 501 bp insertion.** The two primers (5'-GTCACCCAACCCACTGGGCCTTG-3' and 5'-CGGCCGCAATCGCAAGACGCAGA-3') were used for amplifying partial sequence of redTE, and the two primers (5'-GGATACATGCACTATTGATGCGCT-3' and 5'-GGGAGTGTGATATCCGACAGTGTGTCT-3') were used for amplifying 501 bp deletion sequence. Amplification was performed in a thermal cycler (Bio-Rad, C1000, USA), and the temperature programme consisted of an initial denaturation of 3 min at 95 °C followed by programmed for 32 cycles of denaturation at 98 °C for 10 s, annealing at 62 °C for 20 s and extension at 72 °C for 30 s, and final extension was for 2 min at 72 °C. The products were analysed by electrophoresis in 1.5% agarose gel containing ethidium bromide and photographed under a UV transilluminator (Azure C150, USA). The PCR product of 750 bp for redTE and 562 bp for 501 bp insertion were confirmed by Sanger sequencing.

**Luciferase reporter assays in *Nicotiana benthamiana* leaves.** The two primers (5'-GGTACCTTATATATGTGTGGACCCTAGAAACT-3' and 5'-GGAAGCTTACGAGCCGAAGCTCAA-3') were used for amplifying redTE, and it was cloned upstream of the 35S minimal promoter at the *KpnI*-*HindIII* sites in pGreenII 0800-LUC vector, generate the reporter construct redTE: minimal 35S:LUC. A reporter construct containing the cauliflower mosaic virus (CaMV) 35S minimal promoter driving expression of the firefly luciferase gene was used to test the control region segments. Two reporter constructs were transformed into

*Agrobacterium tumefaciens* strain GV3101. Bacterial suspensions were infiltrated into young leaves of the 8-week-old *N. benthamiana* plants using a needleless syringe. After infiltration, plants were grown first under dark for 12 h and then with 16 h light/8 h dark cycle for 60 h at 25 °C. The leaves were sprayed with 100 mM luciferin and maintained under dark condition for 2 min. The LUC images were captured in a low-light cooled CCD imaging apparatus (Tanon 5200Multi, China). The experiments were repeated independently at least three times with similar results.

**McrBC-based methylation assay.** A McrBC-PCR method was used to analyse the methylation degree of in the MdMYB1 promoter and redTE regions. Briefly, one microgram of genomic DNA (gDNA) isolated from Hanfu and HanM peel samples were digested overnight with McrBC (New England Biolabs), with three biological replicates. The digested gDNA and its respective control were used as the template for semi-quantitative PCR analysis. The MdMYB1 promoter and the redTE sequences were divided into seven and five fragments, respectively, amplified with their corresponding primers (Supplementary Table 12, MR1-MR7 are quoted from reference<sup>34</sup>) and visualised under the UV Transilluminator (Azure C150, USA) after 1.5% gel electrophoresis.

**Electrophoretic mobility shift assays.** The coding region of MdCBF2 (MD06G1072200) was cloned into the pGEX4T-1 vector, and its recombinant vector was transformed into Rosetta (DE3) for induction expression. The electrophoretic mobility shift assays (EMSA) were performed using the LightShift Chemiluminescent EMSA Kit (#89880; Thermo Scientific), according to the manufacturer's protocol. The unlabelled probes, biotin-labelled probes and biotin-labelled mutant probes at the 3' end were synthesised by Genewiz Co., Ltd (Supplementary Table 13). The protein-DNA samples were separated on 6.5% acrylamide gels, and transferred to a nylon membrane, and signals were captured using ChemiDoc MP Imaging System (BIO-RAD).

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

## Code availability

The repeat annotation and gap filling scripts are available through [<https://github.com/moold/Genome-data-of-Hanfu-apple>].

## Data availability

Data generated during the study are deposited in the NCBI under study [PRJNA482033](https://www.ncbi.nlm.nih.gov/study/PRJNA482033). Raw data (PacBio and Illumina reads) have been deposited in the Sequence Read Archive (SRA) under study accession number [SRX4557792](https://www.ncbi.nlm.nih.gov/sra/SRX4557792), [SRX4557793](https://www.ncbi.nlm.nih.gov/sra/SRX4557793) and [SRX4557794](https://www.ncbi.nlm.nih.gov/sra/SRX4557794). RNA-seq data of ten tested samples from Hanfu are available under the SRA accession numbers [SRX4557795](https://www.ncbi.nlm.nih.gov/sra/SRX4557795), [SRX4557802](https://www.ncbi.nlm.nih.gov/sra/SRX4557802), [SRX4557790](https://www.ncbi.nlm.nih.gov/sra/SRX4557790), and [SRX4557791](https://www.ncbi.nlm.nih.gov/sra/SRX4557791). Genome assembly and annotation data has been deposited at DDBJ/ENA/GenBank under the accession [RDQH00000000](https://www.ncbi.nlm.nih.gov/nuccore/RDQH00000000). The version described in this paper is version RDQH01000000. FASTA files of chromosomes and genes, as well as gff files for gene models can also be downloaded from [<https://github.com/moold/Genome-data-of-Hanfu-apple>]. Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this Article is available as a Supplementary Information file. The datasets generated and analysed during the current study are available from the corresponding author on reasonable request. The source data underlying Figs. 2, 3a, 3b, 3c, 3e, 4b and 4d, as well as Supplementary Figs. 3, 6, 11b, 12, 14, 15, 16 and 17 are provided in Source Data file 1. The source data underlying Figs. 3d, 3f, and 3g are provided in Source Data file 2.

Received: 9 August 2018 Accepted: 13 March 2019

Published online: 02 April 2019

## References

- Peace, C. P. DNA-informed breeding of rosaceous crops: promises, progress and prospects. *Hortic. Res.* **4**, 17006 (2017).
- Evans, K. The apple genome-harbinger of innovation for sustainable apple production, in *Achieving Sustainable Cultivation of Apples* (Burleigh Dodds Science Publishing Limited, Cambridge, 2017).
- Daccord, N. et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
- Velasco, R. et al. The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
- Li, X. et al. Improved hybrid de novo genome assembly of domesticated apple (*Malus × domestica*). *GigaScience* **5**, 35 (2016).
- Duan, N. et al. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nat. Commun.* **8**, 249 (2017).
- Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2012).
- Espley, R. V. et al. Multiple repeats of a promoter segment causes transcription factor autoregulation in red apples. *Plant Cell* **21**, 168–183 (2009).
- Zhang, J. et al. Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl Acad. Sci.* **113**, E5163–E5171 (2016).
- Chakraborty, M. et al. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat. Genet.* **50**, 20 (2018).
- Takos, A. M. et al. Light-induced expression of a MYB gene regulates anthocyanin biosynthesis in red apples. *Plant Physiol.* **142**, 1216–1232 (2006).
- Telias, A. et al. Apple skin patterning is associated with differential expression of MYB10. *BMC Plant Biol.* **11**, 93 (2011).
- Ban, Y. et al. Isolation and functional analysis of a MYB transcription factor gene that is a key regulator for the development of red coloration in apple skin. *Plant Cell Physiol.* **48**, 958–970 (2007).
- Jaakola, L. New insights into the regulation of anthocyanin biosynthesis in fruits. *Trends Plant. Sci.* **18**, 477–483 (2013).
- Zhao, D. et al. Evaluation of the specific characters of Hanfu apple cultivar. *J. Fruit. Sci.* **26**, 6–12 (2009).
- Bickhart, D. M. et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).
- Gordon, D. et al. Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344–aae0344 (2016).
- 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).
- Kawahara, Y. et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4 (2013).
- Berardini, T. Z. et al. The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *Genesis* **53**, 474–485 (2015).
- Hibrand Saint-Oyant, L. et al. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nat. Plants* **4**, 473–484 (2018).
- Melters, D. P. et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
- Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
- Shi, L. et al. Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016).
- Miura, K. & Furumoto, T. Cold signaling and cold response in plants. *Int. J. Mol. Sci.* **14**, 5312–5337 (2013).
- Chagné, D. et al. The draft genome sequence of European pear (*Pyrus communis* L. 'Bartlett'). *PLoS ONE* **9**, e92644 (2014).
- Hollister, J. D., Ross-Ibarra, J. & Gaut, B. S. Indel-associated mutation rate varies with mating system in flowering plants. *Mol. Biol. Evol.* **27**, 409–416 (2010).
- Tian, D. et al. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**, 105–108 (2008).
- Espley, R. V. et al. Red colouration in apple fruit is due to the activity of the MYB transcription factor, MdMYB10. *Plant J.* **49**, 414–427 (2007).
- Butelli, E. et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* **24**, 1242–1255 (2012).
- van't Hof, A. E. et al. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102 (2016).
- Wei, L. & Cao, X. The effect of transposable elements on phenotypic variation: insights from plants to humans. *Sci. China Life Sci.* **59**, 24–37 (2016).
- Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.* **43**, 1160 (2011).
- El-Sharkawy, I., Liang, D. & Xu, K. Transcriptome analysis of an apple (*Malus × domestica*) yellow fruit somatic mutation identifies a gene network module highly associated with anthocyanin and epigenetic regulation. *J. Exp. Bot.* **66**, 7359–7376 (2015).
- Song, X. & Cao, X. Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. *Curr. Opin. Plant Biol.* **36**, 111–118 (2017).
- Martin, A. et al. A transposon-induced epigenetic change leads to sex determination in melon. *Nature* **461**, 1135 (2009).
- Maruyama, K. et al. Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. *Plant J.* **38**, 982–993 (2004).

38. Antonius-Klemola, K., Kalendar, R. & Schulman, A. H. TRIM retrotransposons occur in apple and are polymorphic between varieties but not sports. *Theor. Appl. Genet.* **112**, 999–1008 (2006).
39. Noiton, D. A. & Alspach, P. A. Founding clones, inbreeding, coancestry, and status number of modern apple cultivars. *J. Am. Soc. Hortic. Sci.* **121**, 773–782 (1996).
40. Bannier, H. -J. Modern apple breeding: genetic narrowing and inbreeding tendencies *Erwerbs-Obstbau* **52**, 85–110 (2011).
41. Li, Y. An investigation and studies on the origin and evolution of *Malus domestica* Borkh. in the World. *Acta Hortic. Sin.* **26**, 213–220 (1999).
42. Harris, S. A., Robinson, J. P. & Juniper, B. E. Genetic clues to the origin of the apple. *Trends Genet.* **18**, 426–430 (2002).
43. Chagné, D. et al. A functional genetic marker for apple red skin coloration across different environments. *Tree Genet. Genomes* **12**, 67 (2016).
44. Zhang, Q. et al. A single-nucleotide polymorphism in the promoter of a hairpin RNA contributes to *Alternaria alternata* leaf spot resistance in apple (*Malus domestica*). *Plant Cell* **30**, 1924–1942 (2018).
45. Chin, C. -S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050 (2016).
46. Chin, C. -S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563 (2013).
47. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
48. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
50. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119 (2013).
51. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
52. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res.* **27**, 573 (1999).
53. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucl. Acids Res.* **35**, W265–W268 (2007).
54. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **4**, 4.10 (2009).
55. Nussbaumer, T. et al. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucl. Acids Res.* **41**, D1144–D1151 (2012).
56. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
57. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinforma.* **6**, 31 (2005).
58. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357 (2015).
59. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290 (2015).
60. Korf, I. Gene finding in novel genomes. *BMC Bioinforma.* **5**, 59 (2004).
61. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucl. Acids Res.* **34**, W435–W439 (2006).
62. Consortium, U. UniProt: a hub for protein information. *Nucl. Acids Res.* gku989 (2014).
63. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
64. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
65. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
67. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucl. Acids Res.* **34**, W609–W612 (2006).
68. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
69. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
70. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
71. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
72. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
73. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl. Acids Res.* **38**, e164–e164 (2010).
74. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single molecule sequencing. *bioRxiv* 169557 (2017).
75. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
76. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
77. Ou, S. & Jiang, N. LTR\_retriever: a highly accurate and sensitive program for identification of long terminal-repeat retrotransposons. *Plant Physiol.* 01310.2017 (2017).
78. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinforma.* **9**, 18 (2008).

## Acknowledgements

This work was supported by the Agricultural Science and Technology Innovation Programme & Cooperation and Innovation Mission (No.CAAS-ASTIP-2016-RIP-02 and CAAS-XTCX2016), the earmarked fund for the China Agriculture Research System (No. CARS-27) and Fundamental Research Funds for Central Non-profit Scientific Institution (No.1610182016020 and Y2019XK09).

## Author contributions

L.Z., C.Z., De.W. and P.C. designed and managed the project. L.Z. and J.H. performed biological experiments for DNA sequencing. J.H., L.Z., J.L., Yu.T., C.Y. and G.L. performed data analyses. X.H. and G.Y. performed q-PCR. Yi.T. provided the samples of HFTH1, G.K. and Y.W. provided the samples of Hanfu and HanfuM, Y.G., Da.W., K.W. and H.Z. collected apple accessions samples and extracted DNA. J.H. and M.M. performed Hi-c data analyses. L.Z. and J.H. wrote the manuscript. C.M.R., H.G. and P.C. revised the manuscript. All authors read and approved to publish the manuscript.

## Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-09518-x>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Journal peer review information:** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019