the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2021) doi: 10.1111/tpj.15394

The chromosome-scale reference genome of *Rubus chingii* Hu provides insight into the biosynthetic pathway of hydrolyzable tannins

Longji Wang[1,2,†], Ting Lei[1,2,†], Guomin Han[1,†], Junyang Yue[3,†], Xueru Zhang[4], Qi Yang[4], Haixiang Ruan[1], Chunyang Gu[1], Qiang Zhang[1], Tao Qian[1], Niuniu Zhang[1], Wei Qian[1], Qi Wang[1], Xiaojing Pang[1], Yue Shu[1], Liping Gao[1,2] and Yunsheng Wang[1,2,*]

[1]*Life Science College, Anhui Agricultural University, Hefei 230036, China,*

[2]*State Key Laboratory of Tea Plant Biology, Utilization, Anhui Agricultural University, Hefei 230036, China,*

[3]*Horticulture College, Anhui Agricultural University, Hefei 230036, China, and*

[4]*GrandOmics Biosciences, Wuhan 430073, China*

Received 10 March 2021; revised 5 June 2021; accepted 21 June 2021.

*For correspondence (e-mail wangyunsheng@ahau.edu.cn).

†These authors contributed equally to this article.

SUMMARY

***Rubus chingii* Hu (Fu-Pen-Zi), a perennial woody plant in the Rosaceae family, is a characteristic traditional Chinese medicinal plant because of its unique pharmacological effects. There are abundant hydrolyzable tannin (HT) components in *R. chingii* that provide health benefits. Here, an *R. chingii* chromosome-scale genome and related functional analysis provide insights into the biosynthetic pathway of HTs. In total, sequence data of 231.21 Mb (155 scaffolds with an N50 of 8.2 Mb) were assembled into seven chromosomes with an average length of 31.4 Mb, and 33 130 protein-coding genes were predicted, 89.28% of which were functionally annotated. Evolutionary analysis showed that *R. chingii* was most closely related to *Rubus occidentalis*, from which it was predicted to have diverged 22.46 million years ago (Table S8). Comparative genomic analysis showed that there was a tandem gene cluster of UGT, carboxylesterase (CXE) and SCPL genes on chromosome 02 of *R. chingii*, including 11 CXE, eight UGT, and six SCPL genes, which may be critical for the synthesis of HTs. *In vitro* enzyme assays indicated that the proteins encoded by the CXE (LG02.4273) and UGT (LG02.4102) genes have tannin hydrolase and gallic acid glycosyltransferase functions, respectively. The genomic sequence of *R. chingii* will be a valuable resource for comparative genomic analysis within the Rosaceae family and will be useful for understanding the biosynthesis of HTs.**

Keywords: *Rubus chingii* Hu, chromosome-scale genome, Rosaceae, evolutionary analysis, hydrolyzable tannins, tandem gene cluster.

INTRODUCTION

Rosaceae is a family with approximately 3000 species and an extraordinary spectrum of distinctive fruits, including apple (*Malus domestica*), pear (*Pyrus bretschneideri*), cherry (*Prunus avium*), strawberry (*Fragaria ananassa*), and raspberry (*Rubus*) (Xiang et al., 2017). The genus *Rubus*, one of the most diverse Rosaceae genera with 740 species, produces aggregate drupetum fruits that are acclaimed for their tender flesh and pleasant flavor (Janick and Moore, 1996; Jennings, 1988). Among them, red raspberry (*Rubus idaeus* L.), black raspberry (*Rubus occidentalis* L.), and blackberry (*Rubus fruticosus* L.) are the most popular varieties with the largest cultivated area. Raspberries are botanically classified in the *Rubus* subgenus *Idseobatus* (Hummer and Janick, 2007), whereas blackberries belong to the subgenus *Rubus* (Tutin et al., 1980).

Modern uses of *Rubus* plants include mainly consumption as fresh fruits and the production of juices and dairy products. These delicious fruits are rich in polyphenols, vitamins, sugars, and minerals and have a high antioxidant capacity. *Rubus* plants were not originally used as food according to ancient traditional literature; rather, they were considered to be medicinal, with a wide range of pharmacological uses. For instance, Hippocrates recommended blackberry stems and leaves soaked in white wine as an astringent poultice for difficulties during childbirth (Hummer and Janick, 2007). More recently, *Rubus* plants were found to be very high in secondary metabolites such as

hydrolyzable tannins (HTs), condensed tannins, flavonols, and anthocyanins, which function as antioxidants and provide other health benefits (Kaume et al., 2012; Moyer et al., 2002). China is regarded as the center of origin and distribution for the genus *Rubus*. There are 210 species of *Rubus* in China, particularly in the southeast. *Rubus chingii* is one of the "magical" raspberry species in China. The dried fruitlet fruits of *R. chingii*, referred to as "Fu-Pen-Zi" in Chinese, are used in traditional Chinese medicine. Fu-Pen-Zi means "overturned bowl" in Chinese and may refer to the shape of an upside-down raspberry fruit or to the diuretic effects of this herb. Its effect is recorded in the Compendium of Materia Medica, which states that Fu-Pen-Zi has sour and warm properties and is associated with the liver and kidney meridians. Modern medicine has shown that this herbal medicine can help to prevent frequent urination, reduce lower back soreness, improve eyesight, and prevent cancer (Yu et al., 2019; Zeng et al., 2018; Zhang et al., 2015). Tannins, particularly HTs, are reported to be its bioactive constituents (Han et al., 2012; Zhang et al., 2019). However, relatively little is understood about the genetic mechanisms that control the biosynthesis and accumulation of HTs, and few genetic resources are available for *R. chingii*.

In recent years, the whole genome sequencing of Rosaceae plants has proceeded rapidly with the development of sequencing technology. More than 20 Rosaceae plants have had their entire genomes sequenced, including apple (*Malus* × *domestica* Borkh.) (Velasco et al., 2010), pear (*Pyrus bretschneideri* Rehd.) (Wu et al., 2013), woodland strawberry (*Fragaria vesca*) (Shulaev et al., 2011), black raspberry (Jibran et al., 2018; VanBuren et al., 2016; VanBuren et al., 2018), and other Rosaceae crops. *Rubus chingii* is diploid ($2n = 2x = 14$) and belongs to the same subgenus (*Idaeobatus*) as red raspberry (*R. idaeus* L.), with which it can be crossed (Thompson and Zhao, 1993). Evolutionary analysis of Rosaceae genomes from different orders provides evidence for the phylogenetic position of *Rubus*. Here, we report a high-quality draft genome of *R. chingii* assembled by integrating Nanopore long-read sequencing, BioNano DLS optical mapping, and Hi-C mapping technologies. In addition, transcriptomics, phenolic metabolomics, and gene functional analyses were used to characterize an aggregation gene cluster associated with the biosynthesis of HTs. The *R. chingii* genomic resources provided here will be valuable for biological and agronomic research on *Rubus* species and will provide new tools for Rosaceae geneticists and breeders.

RESULTS

Botanical and horticultural characteristics of *Rubus chingii*

A hand-painted illustration of *R. chingii* was created for this paper based on the botanical and horticultural characteristics of the species (Figure 1). Stems of *R. chingii* are semi-woody and biennial with a few thorns, and its root system is perennial. Leaves tend to be divided into five or seven palmatilobate segments towards the base. The flowers of *R. chingii* have five white petals and multiple stamens. After the petals have fallen, the fruit develops as an aggregate of drupelets that change during ripening from green to red. Flowers and fruits hang like an upside-down bowl on the branches of the previous year's primocane.

Sequencing, assembly, and annotation of the *Rubus chingii* genome

A wild-collected specimen of *R. chingii*, cv. Wanfu 1, planted in the botanical garden of Anhui Agricultural University was selected for sequencing. First, an initial *k*-mer analysis showed that its genome size was about 239.4 Mb. Its heterozygosity was 0.80%, making it a highly heterozygous genome (Table S1). The count distribution of 17-mers followed a Poisson distribution, with the highest peak occurring at a depth of 32 (Figure S1). The genome had an average GC content of 36%; its unimodal GC content distribution and GC depth, as well as the sequencing depth of the genome assembly, suggested that there was no contamination from other species (Figures S2, S3). Second, integrated Oxford Nanopore technologies were used to assemble the genome. We obtained 24.8 Gb of data that comprised 1 328 395 clean reads with an average length of 18.7 kb, an N50 of 26.2 kb, and a maximum read length of 200.9 kb (Table S2). This version of the genome assembly was approximately 233.9 Mb in size with a scaffold N50 of 8.2 Mb (Table S3). Based on BUSCO predictions of the assembled sequence, about 97.1% of the complete gene elements were found in the *R. chingii* genome, indicating that most conserved genes were relatively complete and that the assembly was of high quality (Table S4). Finally, using Hi-C technology, 94.09% (220.05 Mb) of the assembled scaffold sequences (231.21 Mb) were anchored on to seven chromosomes with a maximum and average length of 37.2 and 31.4 Mb, respectively. The number of corresponding scaffold cut bins was 2383, accounting for 98.76% (Table 1, Tables S5, S6). A Hi-C interaction heatmap of the chromosomes was created based on the signal intensities of the interactions between the scaffold sequences that determined order and direction (Figure 2a). There was no obvious noise (strong interaction intensity) outside the diagonal, again confirming the quality of the assembly.

Simple sequence repeat (SSR) sequences in the genome were analyzed using MISA, and 1 817 604 such sequences were found in the genome (Tables S9, S10). In total, 36.47% of the *R. chingii* genome consisted of repetitive sequences (Table S11). This level of repetitive sequence is greater than that in some sequenced Rosaceae species, including strawberry (22.0%) (Shulaev et al., 2011) and peach (29.6%) (International Peach Genome et al., 2013),

Figure 1. Plant, dried fruitlet, and ripe fruits of Fu-Pen-Zi (*Rubus chingii* Hu).
Image was hand-painted by Rongqing Yang and the calligraphy was created by Yuncai Tang.

Table 1 Major indicators of the *Rubus chingii* Hu genome

Assembly feature	Statistics
Estimated genome size (by *k*-mer analysis) (Mb)	239.44
Number of scaffolds	155
Scaffold N50 (Mb)	8.2
Assembled genome size (Mb)	231.21
Assembly rate of genome (%)	96.56
Anchoring size on chromosomes (Mb)	220.05
Anchoring rate on chromosomes (%)	95.17
Number of chromosomes	7
Average length of chromosomes (Mb)	31.4
Repeat region of assembly (%)	36.5
Number of predicted protein-coding genes	33 130
Average coding sequence length (bp)	2803
Average exons per gene	4.7

but lower than that of apple (42.4%) (Velasco et al., 2010) and black raspberry (56.6%) (VanBuren et al., 2016). These differences may be partly due to the sensitivity and thoroughness of repeat identification and the integrity of the respective genome assemblies. Similar to most plant genomes, the predominant type of transposable elements (TEs) were long terminal repeat retrotransposons (LTRs), which accounted for 17.70% of the genome and included LTR/Gypsy (7.64%) and LTR/Copia retroelements (4.38%) (Table S12, Figure S4). DNA transposons comprised 6.05% of the genome, among which MULE-MuDR transposons were the most abundant, with 18 739 elements accounting for 1.57% of the genome. In addition, 618 non-coding RNAs, 488 transfer (t)RNAs, 90 ribosomal (r)RNAs, and 10 regulatory RNAs were also identified (Table S12).

Evolutionary analysis of *Rubus chingii*

To reveal the genomic foundation of species adaptation during evolution, we compared the *R. chingii* proteome with those of 15 representative plant species (Table S13). In total, 21 686 orthologous gene families containing 470 865 genes were obtained. Of these, 252 357 genes from 6976 families were shared among all 16 species, representing a core set of ancestral clusters (Figure 3a). On the other hand, 4736 genes in 19 different families were specific to *R. chingii*, suggesting that they may contribute to the unique biological and phytochemical properties of this sublineage.

Functional enrichment analysis based on Gene Ontology (GO) annotation revealed that the specific genes in *R. chingii* tended to be placed in the "cellular process" biological process and "catalytic activity" molecular function categories as summarized at level 2 (Figure 3b). Among the GO terms enriched in the *R. chingii*-specific *genes were terms related to the biosynthesis of major characteristic secondary metabolites (e.g., HTs). These enriched GO terms included "UDP-glucose: glycoprotein glucosyltransferase activity" (GO:0003980,* $P < 0.05$), "superoxide metabolic process" (GO:0006801, $P < 0.01$), "methyltransferase activity" (GO:0008168, $P < 0.001$), and "oxidoreductase activity, acting on the CH-CH group of donors" (GO:0016627, $P < 0.001$) (Table S14).

The expansion and/or contraction of gene families have been well documented as a crucial driving force in lineage splitting and functional diversification of flowering plants (Chen et al., 2013). Here, we characterized gene families that appeared to have undergone discernible changes in adaptive evolution on divergent branches, with a particular emphasis on gene families involved in traits and fruit qualities of *R. chingii*. A phylogenetic analysis was performed

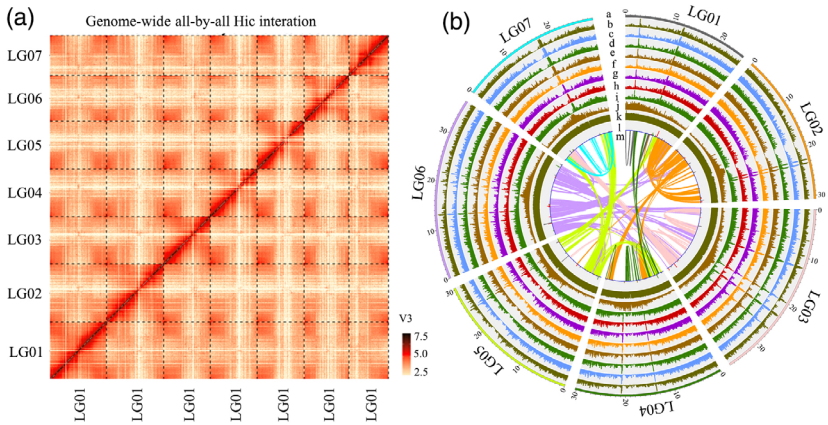

Figure 2. Landscape of the *Rubus chingii* genome.
(a) Genome-wide Hi-C map.
(b) Global view. a) Circular representation of the pseudomolecules. b–h) Gene expression supported by transcriptome data (tracks, from outside to inside: leaf, adultoid, ripe fruit, fruitlet, stem, flower, and root). i–k) Distributions of gene density, repeat density, and GC density, respectively, with densities calculated in 100-kb windows. l) Locations of genes mapped to the flavonoid (brown line), lignin (red line), and polyphenol (green line) metabolic pathways and to the UDP glycosyltransferase family (gray line). m) Syntenic blocks. Band width is proportional to the syntenic block size.

to investigate the evolutionary relationships among species, as well as their estimated divergence times (Figure 3c). Our results showed that, among the 21 620 gene families inferred to be present in the most recent common ancestor of the 16 representative plant species, 1594 families contracted in *R. chingii* genome, and 1538 families gained new gene copies (Figure 3c). The GO annotation of 7414 genes from the 1423 families with significant expansions ($P < 0.05$) demonstrated that they were mainly enriched in functional categories related to "oxidation-reduction process" (GO:0055114, $P < 0.001$), "chitin catabolic process" (GO:0006032, $P < 0.001$), "terpene synthase activity" (GO:0010333, $P < 0.001$), and "heme binding" (GO:0020037, $P < 0.001$) (Table S15).

Previous studies of multiple sequenced plant genomes have shown that polyploidization is a prominent feature in the evolutionary history of angiosperms and that whole genome duplication (WGD) events, in particular, have had profound effects on crop gene amplification and genome evolution (Salman-Minkov et al., 2016; Yue et al., 2020). Here, we identified 18 578 duplicated genes encompassing 56.1% of the putative protein-coding genes in the *R. chingii* genome (Table S16). We took advantage of these pairwise paralogs to calculate the age distribution of synonymous substitution rates (K_s) that peaked at approximately 1.48, providing clear evidence of one round of WGD in *R. chingii* (Figure 3d). We compared this result with corresponding K_s distribution values derived from paralogous pairs in three other representative dicot genomes (*Arabidopsis thaliana*, *F. vesca*, and *R. occidentalis*). Our results confirmed that this WGD event was shared by the Rosaceae family, as it was observed in *F. vesca* and *R. occidentalis* but not in *A. thaliana*.

Identification of HT compounds and related genes

Quadrupole time-of-flight liquid chromatography/mass spectroscopy (Q-TOF LC/MS) was used to identify qualitatively the polyphenols in different tissues of *R. chingii* based on data from the published literature and standard compounds (Chen et al., 2019; Regueiro et al., 2014; Sanz et al., 2010; Staszowska-Karkut and Materska, 2020). The relationship between the theoretical *m/z* values and the measured values was also used to identify individual components with high precision. In total, 61 polyphenols were identified (Figure 4a,b and Table S17): 29 HTs, 15 flavonols, 11 phenolic acids, five condensed tannins, and one anthocyanin. All substances except the anthocyanin were detected in negative ionization mode. HTs are the most abundant polyphenols in *R. chingii*; some are found throughout the plant, and others accumulate only in specific tissues. Up to 20 kinds of HTs were present in leaves, whereas only 10 different HTs were found in the stem (Figure 4c). In addition, tetra- or penta-galloyl glucose was only detected in leaves, whereas tri-galloyl HHDP glucose and methylated/acylated ellagic acids were only found in fruits and roots, respectively. The specificity of HT accumulation in *R. chingii* reflects the complexity of the underlying biosynthetic pathways.

Recent studies have shown that the carboxylesterase (CXE) (Dai et al., 2020), UDP glycosyltransferase (UGT) (Cui et al., 2016; Niemetz and Gross, 2005), serine carboxypeptidase-like protein (SCPL) (Liu et al., 2012), and polyphenol oxidase (PPO) (Grundhofer et al., 2001) families may be involved in pathways of HT synthesis or degradation. Based on previously published protein sequences of *CXE*, *UGT*, *SCPL*, and *PPO* genes from *Arabidopsis* and other plants, in total, 39

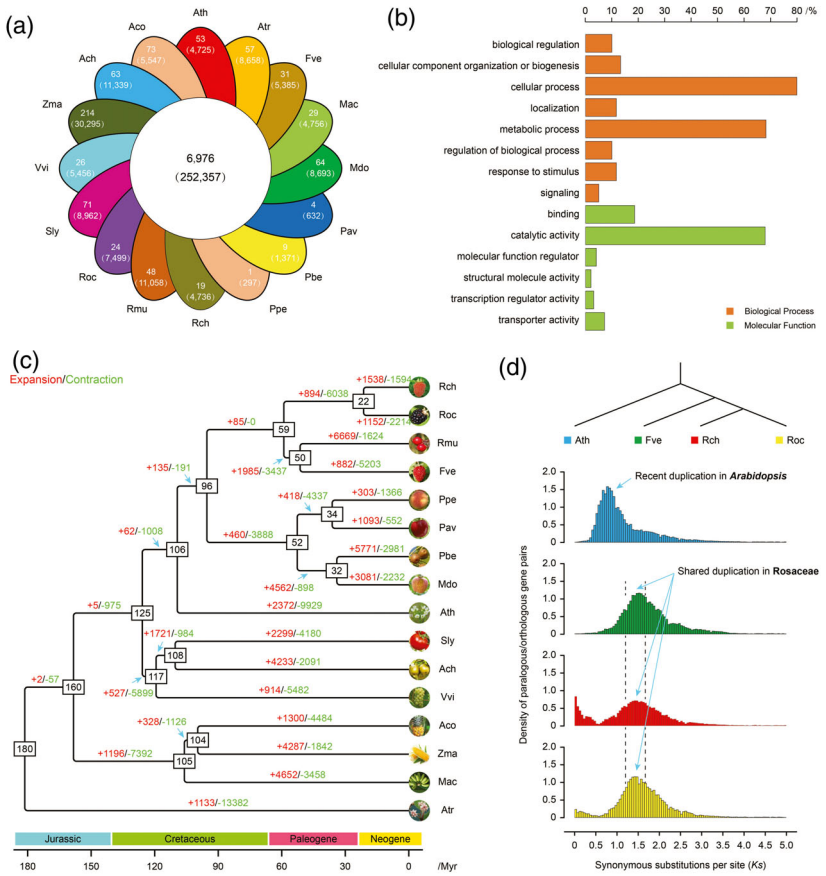

Figure 3. Comparative analysis of genome evolution and gene families in *Rubus chingii*.
(a) Venn diagram showing the shared and specific gene families in *R. chingii* and 15 representative plant species. Values in parentheses indicate the number of genes within the corresponding families. Three-letter acronyms are used as abbreviations for species names.
(b) Specific genes in *R. chingii* were assigned to biological process and molecular function Gene Ontology categories based on their Gene Ontology annotations.
(c) Expansion and contraction of gene families among the 16 plant species. A phylogenetic tree was constructed based on high-quality single-copy and multi-copy orthologous genes, using *Amborella trichopoda* as an outgroup. Numerical values in the box denote the estimated divergence times of each node (MYA, million years ago).
(d) Whole-genome duplication events detected in R. chingii, Arabidopsis thaliana, *Fragaria vesca*, and Rubus occidentalis.

*CXE*s, 139 *UGT*s, 56 *SCPL*s, and 57 *PPO*s were identified in the genome of *R. chingii* (Table S19). These genes are widely distributed on different chromosomes, and there are several *UGT* or *CXE* gene clusters on chromosomes 04 and 06 (Figure S5).

Interestingly, a tandem cluster consisting of CXE, UGT, and SCPL genes was found of *R. chingii*, which clustered between 27.30 and 29.22 Mb on chromosome 02. In terms of distribution, the three gene clusters were arranged in tandem, with 11 CXE genes in front, followed by eight UGT genes and six SCPL genes. Collinearity analysis revealed that the length of the homologous segment from *R. occidentalis* was 2.18 Mb; it may therefore have expanded relative to that of *R. chingii*. This segment of *R. occidentalis* consisted of seven CXE genes, 12 UGT genes, and six SCPL genes. Similarly, there were also homologous segments in the *F.* × *ananassa* (strawberry) and *Rosa rugosa* Thunb. (rose) genomes (Figure S6). However, in the *M.* × *domestica* (apple) genome, the related syntenic region has expanded to different chromosomes and now exists on chromosomes 5, 8, 10, and 15.

To investigate the functions of genes in this syntenic region, we analyzed relevant gene expression levels in different tissues and organs (root, stem, leaf, flower, fruitlet, and ripe fruit) (Figure 5b, Table S22). As shown in Figure 5b, 16 genes in this gene cluster had the highest expression level in ripe fruits, including six CXE genes, four UGT genes, and six SCPL genes. Meanwhile, four genes had the highest expression levels in the root (R), including three CXE genes and one UGT gene. Moreover, a low expression level was detected for three genes including, LG02.4108, 4182, and 4199, in different tissues and organs. To evaluate the relationship further between the content of HTs and expression levels of the genes in this segment, correlative matrix analyses were investigated (Figure S7, Table S21). The results showed that the expression of several CXE

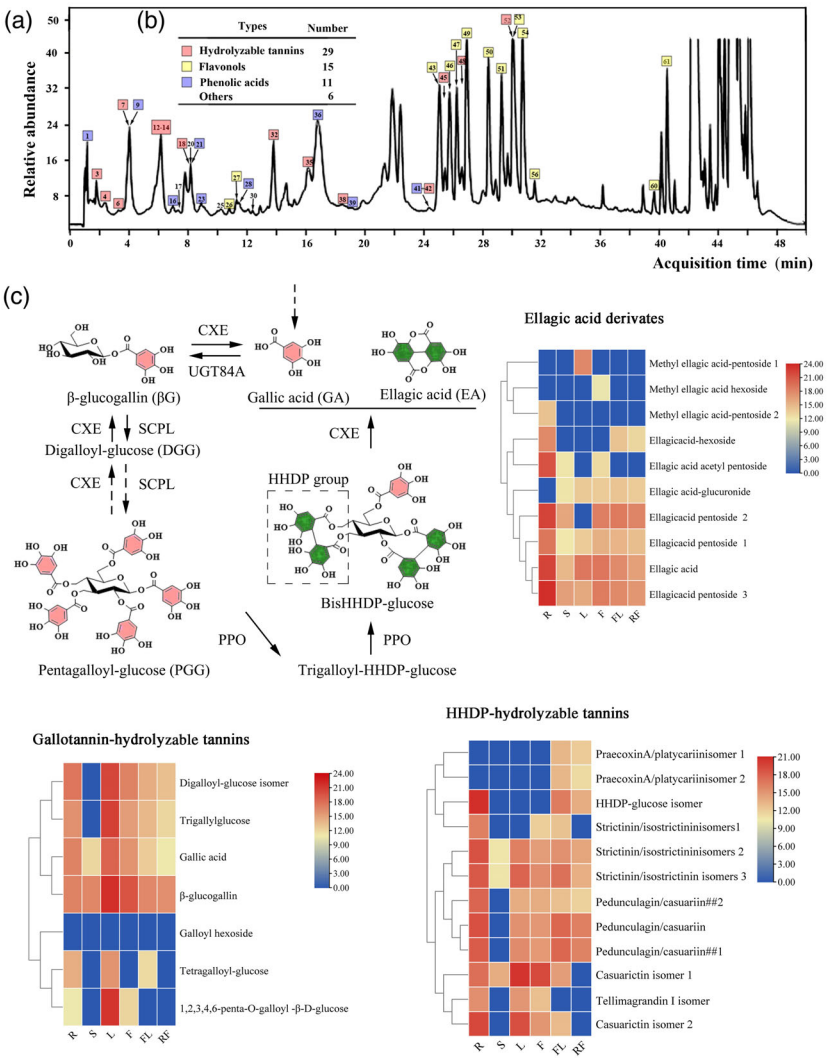

Figure 4. Identification of hydrolyzable tannins in *Rubus chingii*.
(a) Liquid chromatography–time-of-flight–mass spectroscopy chromatogram of phenolic compounds in *R. chingii* leaves. Pink, yellow, and blue boxes represent hydrolyzable tannins, flavonols, and phenolic acids, respectively.
(b) Numbers of phenolic compounds in *R. chingii*.
(c) Biosynthetic pathway and relative quantities of hydrolyzable tannins in different organs of *R. chingii*. All data reflect those in Table S18. CXE, carboxylesterase; F, flower; FL, fruitlet; L, leaf; PPO, polyphenol oxidase; R, root; RF, ripe fruit; S, stem; SCPL, serine carboxypeptidase-like protein; UGT, uridine diphosphate glycosyltransferase.

genes (LG2.4101, 4102, 4104, 4108, 4113, 4117, 4120, and 4123) and UGT genes (LG02.4207, 4211, and 4273) were highly or moderately positive correlation with the content of HTs ($1 > P > 0.8$ or $0.8 > P > 0.4$). Functional prediction analysis indicated that three CXE genes (LG02.4101, 4102, and 4108) and one UGT gene (LG02.4273) may be involved in the biosynthetic pathway of HTs (Figure S8). *In vitro* enzyme assays were performed on the protein products of the highly expressed genes LG02.4102 (a CXE gene) and LG02.4273 (a UGT gene) in *Escherichia coli*. The LG02.4273 recombinant protein exhibited glycosyltransferase activity and could catalyze the formation of β-glucogallin (penta-*O*-galloyl-β-D-glucose [βG]) from gallic acid. LG02.4102 exhibited tannin hydrolase activity and could catalyze the formation of several hydrolytic products from 1,2,3,4,6-*O*-pentagalloylglucose.

DISCUSSION

Rubus chingii (Fu-Pen-Zi) is a characteristic Rosaceae species in China. Its dried fruitlets are used in traditional Chinese medicines, and its ripe fruit provides valuable health benefits. Previous studies have indicated that HTs are key functional components of Rosaceae species (Chen et al., 2019; Zhang et al., 2015). In this study, 29 kinds of HTs were detected in different tissues of *R. chingii* using the Q-TOF LC/MS method. We also assembled a chromosome-

level reference genome sequence of *R. chingii* using multiple types of sequencing data and assembly technologies. The sequencing data enabled whole-genome analysis for the identification of key genes in the HT biosynthetic pathway of *R. chingii*.

The size of the assembled genome was 231.21 Mb, covering 96.6% of the estimated genome size (Table 1). Ninety-four percent of the assembled sequences were anchored on seven chromosomes (average length 31.4 Mb) using Hi-C technology. The estimated genome sizes of other diploid species in the Rosaceae are 243 Mb in black raspberry (VanBuren et al., 2016), 240 Mb in strawberry, 257 Mb in Japanese apricot (*Prunus yedoensis*) (Baek et al., 2018), 512 Mb in pear (Wu et al., 2013), and 651 Mb in apple (Daccord et al., 2017), all of which are greater than that of *R. chingii*. There were 33 130 predicted gene locations in the *R. chingii* genome, similar to other Rosaceae species such as strawberry (N = 33 387) (Shulaev et al., 2011) and black raspberry (N = 33 253) (Tables S7, S8) (VanBuren et al., 2016). In total, 85.31 Mb (36.5%) of repetitive sequences were identified, similar to the amount found in black raspberry (32.6%) (VanBuren et al., 2018) and lower than that in pear (53.1%), apple (57.3%), persimmon (65.0%) (Zhu et al., 2019), and cotton (69.83%) (Wang et al., 2019). Thus, the *R. chingii* genome can be used to understand better the genome evolution and the characteristic secondary metabolites of Rosaceae plants, given its small size and moderate level of repetitive content.

We studied the potential genetic basis for HT biosynthesis by comparative genome and metabolite analysis (Figure 4).

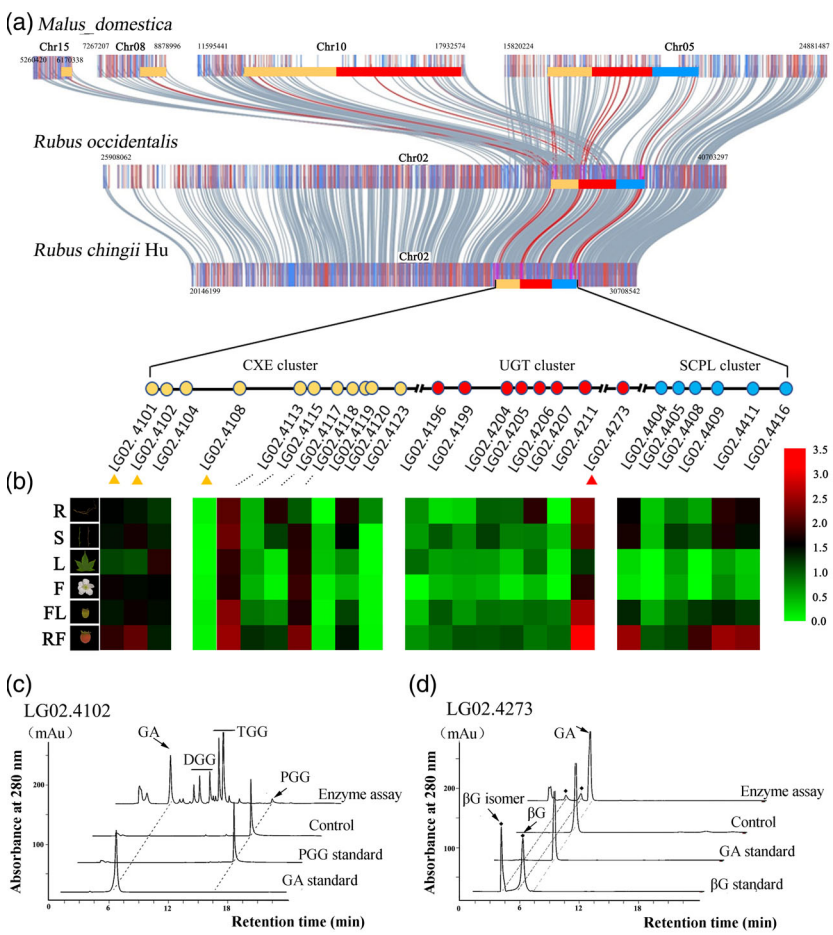

Figure 5. Identification of hydrolyzable tannin-related gene clusters in *Rubus chingii*.
(a) Collinearity analysis of hydrolyzable tannin-related gene clusters on pseudochromosome 02 of *R. chingii*. Yellow, red, and blue boxes or blocks represent *CXE*, *UGT*, and *SCPL* gene clusters, respectively.
(b) Expression patterns of *CXE*, *UGT*, and *SCPL* genes in the hydrolyzable tannin-related clusters. Gene information is given in Table S15. CXE, carboxylesterase; F, flower; FL, fruitlet; L, leaf; R, root; RF, ripe fruit; S, stem; SCPL, serine carboxypeptidase-like; UGT, uridine diphosphate glycosyltransferase.
(c,d) High-performance liquid chromatography chromatograms of the enzymatic products of LG02.4102 (CXE) and LG02.4273 (UGT) recombinant proteins obtained using PGG or GA as the substrate, respectively. βG, β-glucogallin; GA, gallic acid; DGG, di-*O*-β-D-glucose; PGG, penta-*O*-galloyl-β-D-glucose; TGG, tri-*O*-galloyl-β-D-glucose.

Glycosylation catalyzed by UGTs is an important process that influences the diverse functions of polyphenolic compounds in plants (Yoshida et al., 2000). βG is the galloyl acceptor in the biosynthesis of HTs, and *UGT84* subfamily genes are involved in plant βG synthesis (Cui et al., 2016; Khater et al., 2012). The SCPL family may participate in the acylation of HTs using βG as an acyl donor (Bontpart et al., 2015). Plant-specific tannin acyl-hydrolase (TA) belongs to the CXE family and is involved in the hydrolysis of HTs (Dai et al., 2020). Members of the *UGT*, *SCPL*, and *CXE* gene families were identified in the *R. chingii* genome; it contained 178 *UGT*, 55 *SCPL*, and 39 *CXE* genes. A tandem cluster of 11 *CXE*, eight *UGT*, and six *SCPL* genes was discovered on chromosome 02. Functional prediction analysis showed that there were three *TA* and one *UGT* homologs in this cluster. Related transcriptome analysis and recombinant protein activity assays further confirmed that this aggregation cluster was involved in the biosynthesis of HTs.

Collinearity analysis indicated that this homologous segment also existed in the genomes of black raspberry, strawberry, rose, apple, and pear. However, there was a comparatively large amount of expansion of these syntenic regions, particularly in the apple genome, in which the region was scattered across different chromosomes.

In plants, HTs play an important role in the resistance to biotic and abiotic stresses, and *TA* genes are important for the hydrolysis and release of tannins (Dai et al., 2020). However, the predicted key *TA* genes are missing from the genomes of several Rosaceae species such as apple and pear that have low HT concentrations in their fruit (Table S20). The expansion of the homologous HT-related segment, the loss of key TA genes, and the reduced fruit HT accumulation in apple and pear indicate that this aggregation cluster may be critical for HT synthesis. The pathway and regulation of HT metabolism in Rosaceae species will be studied further in our future research.

The *R. chingii* genome sequence provides an invaluable new resource for biological research on *Rubus*. In this study, a chromosome-scale genome and related transcriptome analysis provide insights into the biosynthesis of HTs in *R. chingii*. Additional genome-wide comparative studies will provide insight and advance our understanding of genome evolution in the Rosaceae. The availability of this genome sequence will also enable the continued study of comparative genomics among species, thereby shedding new light on the evolution of gene families.

EXPERIMENTAL PROCEDURES

Sampling and sequencing

All samples from different developmental stages and tissues of "Fu-Pen-Zi" (*R. chingii*) were collected at the Agricultural Innovation Industrial Park of Anhui Agricultural University, Hefei, Anhui Province, China (N31.94, E117.21). Leaf, stem, flower, root, fruitlet, and ripe fruit samples were collected, placed directly in liquid nitrogen and stored at −80°C for transcriptome sequencing and polyphenol compound analyses (three replicates per tissue).

Genomic DNA was extracted from leaves of a single plant using the Plant Genomic DNA kit (Qiagen, San Diego, CA, USA). The genomic DNA sample was further purified for Oxford Nanopore sequencing, and a genomic DNA library was constructed using the ONT 1D ligation sequencing kit (SQK-LSK108) according to the manufacturer's instructions. Single-molecule real-time sequencing of long reads was performed on a GridION X5 platform (Senol Cali et al., 2019) (Oxford Nanopore Technology, Oxford, UK). Compared with other sequencing platforms, the longer reads produced by the Nanopore platform offer many advantages. A separate paired-end (PE) DNA library with an insert size of 350 bp was constructed and sequenced on the HiSeq X Ten platform (Illumina, San Diego, CA, USA).

For transcriptome sequencing, total RNA was extracted from different samples of *R. chingii* using the QIAGEN RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). cDNA libraries were prepared using the TruSeq Sample Preparation Kit (Illumina), and PE sequencing was performed on the NovaSeq 6000 platform (Illumina).

Genome size estimation and *de novo* assembly

The genome size of *R. chingii* was estimated by the *k*-mer method using sequencing data from the Illumina DNA library. Quality-filtered reads were subjected to 17-mer frequency distribution analysis using the JELLYFISH program (Marcais and Kingsford, 2011). After removing adaptor contamination and filtering out low-quality reads, we obtained clean reads for assembly. First, NEXTDENOVO (https://github.com/Nextomics/Nextdenovo) was used for read error correction with default parameters. Next, SMART (Etherington et al., 2020) was used to assemble the corrected reads independently. Finally, the Illumina short-read data were compared with the Nanopolish-corrected genome using BWA with default parameters, and three iterations of Pilon were used to correct the assembly (Loman et al., 2015; Walker et al., 2014). A guanine-cytosine (GC) depth analysis was performed to assess potential sequencing contamination and assembly coverage. We then used Benchmarking Universal Single-Copy Orthologs (BUSCO) to search the annotated genes (Simao et al., 2015).

Chromosome assembly using Hi-C data

Hi-C technology is an efficient and low-cost strategy for clustering, ordering, and orienting sequences for pseudomolecule construction; it enables the generation of genome-wide 3D proximity maps (Burton et al., 2013). It has been successfully applied to recent complex genome projects, including the genomes of goat (Bickhart et al., 2017), Tartary buckwheat (Bickhart et al., 2017), and wild emmer (Avni et al., 2017). To obtain a chromosome-level assembly of the *R. chingii* genome, Hi-C fragment libraries were constructed following a previously published procedure (Belton et al., 2012) with modifications. In brief, sample cells were fixed with formaldehyde to crosslink DNA with proteins and/or proteins with proteins. The *Dpn*II restriction enzyme was used for chromatin digestion. After biotin labeling, blunt-end ligation, and DNA purification, Hi-C fragments were prepared and sampled for DNA quality testing. The Hi-C fragments were subjected to terminal biotin removal, ultrasonic interruption, terminal repair, and base A addition to form splice products. Polymerase chain reaction conditions were selected, and the sequences were amplified to obtain the library product. After quality control, the Hi-C libraries were quantified and sequenced on the Illumina HiSeq platform (Illumina). Quality control of the raw Hi-C data was performed as

follows: (i) low-quality sequences (quality scores <15), adaptor sequences, and sequences shorter than 30 bp were removed using FASTP (Chen et al., 2018); and (ii) clean PE reads were mapped to the draft genome assembly using BOWTIE 2 (Langmead and Salzberg, 2012) to obtain uniquely mapped PE reads. Finally, the LACHESIS *de novo* assembly pipeline was used to produce chromosome-level scaffolds.

Genome annotation

Homology-based, *de novo*, and transcript-based gene prediction methods were used to annotate protein-coding genes. For homology-based predictions, protein sequences from eight species (*A. thaliana*, *P. yedoensis*, *F. vesca*, *M.* × *domestica*, *R. chinensis*, *R. occidentalis*, *P. avium*, and *F.* × *ananassa*) were mapped to the *R. chingii* genome using GeMoMa (Keilwagen et al., 2016). For the *de novo* predictions, AUGUSTUS (Stanke et al., 2004) was used to predict genes based on a training set. *Rubus chingii* RNA-sequencing data were used for transcript-based gene prediction with TRANSDECODER (Haas et al., 2008). Finally, EVIDENCEMODELER (Haas et al., 2008) was used to integrate the predicted genes and generate a consensus gene set. Genes with TEs were identified and discarded using the TRANSPOSONPSI package (http://transposonpsi.sourceforge.net). Low-quality genes that encoded fewer than 50 amino acids or contained premature termination codons were also removed from the gene set. Functional annotation of the predicted genes was performed by BLASTP alignment of their protein sequences against public protein databases (Camacho et al., 2009): SwissProt (Boeckmann et al., 2003), GO (Ashburner et al., 2000), and KEGG (Kanehisa and Goto, 2000). The INTERPROSCAN package (Quevillon et al., 2005) was also used to annotate the predicted genes. After all the above predictions, BUSCO (Simao et al., 2015) was used to evaluate the integrity and completeness of the predicted gene set.

Homology-based non-coding RNA annotation was performed by mapping plant rRNA, tRNA, and small nuclear RNA genes from the Rfam database (Kalvari et al., 2018) to the *R. chingii* genome. tRNAscan-SE (Lowe and Eddy, 1997) was used for tRNA annotation, and RNAmmer (Lagesen et al., 2007) was used to predict rRNAs and their subunits.

The repetitive sequences in the genome consisted of SSRs, moderately repetitive sequences, and highly repetitive sequences. The microsatellite identification tool MISA (Thiel et al., 2003) was used with default parameters to search for SSR motifs in the *R. chingii* genome. REPEATMASKER (Tarailo-Graovac and Chen, 2009) was used to screen the assembled genome against Repbase (Bao et al., 2015) to identify known TEs.

Analysis of gene families and genome evolution

The ORTHOFINDER package (Emms and Kelly, 2015) was used to identify gene families in *R. chingii* and 15 representative plant species (Table S1). Species-specific genes and the families to which they belonged were determined based on the presence or absence in a given species. We investigated the dynamic evolution (i.e., expansion and contraction) of orthologous gene families using the latest version of Computational Analysis of gene Family Evolution (De Bie et al., 2006) with probabilistic graphical models. Evolutionary relationships among the 16 plant species were resolved using the Randomized Accelerated Maximum Likelihood (RAxML) package (Stamatakis, 2006) based on 75 single-copy and 400 multi-copy orthologous genes. The resulting phylogenetic trees were visualized using MEGA (Kumar et al., 2018), and estimated divergence times were retrieved directly from the online TimeTree database (Kumar et al., 2017).

We detected WGD events in a given species using paralogous gene pairs. In brief, we first identified paralogous gene pairs from results produced by the ORTHOFINDER package (Emms and Kelly, 2015), yielding, in total, 178 749, 85 766, 172 007, and 93 824 gene pairs in the proteomes of *R. chingii*, *A. thaliana*, *F. vesca*, and *R. occidentalis*, respectively. These represented approximately 56.1% (18 578 of 33 130), 58.5% (16 049 of 27 445), 56.4% (19 158 of 33 950), and 47.3% (15 733 of 33 286) of the total protein-coding genes in each species. We then calculated the synonymous substitutions per synonymous site (K_s) for these gene pairs using the NG (Nei & Gojoberi) method implemented in PAML (for Phylogenetic Analysis by Maximum Likelihood) (Yang, 2007). Finally, the K_s distribution for each species was visualized using the R statistical package (version 3.2.5).

Analysis of gene families involved in the HT pathway

DNA-Tool software (Callaway, 2013) was used to construct a database of *R. chingii* nucleic acids and proteins (https://www.ncbi.nlm.nih.gov/sra/PRJNA666516). The published protein sequences of *UGT*, *CXE*, and *SCPL* genes from Arabidopsis were used to identify homologous genes in the local *R. chingii* protein database with BLASTP. Corresponding proteins were also identified using the Pfam (Finn et al., 2014) and SMART (Letunic and Bork, 2018) databases. Finally, all the candidate protein sequences were analyzed using MEGA (Kumar et al., 2018), repetitive sequences were removed manually, and gene family branch classification was performed using related genes that shared a published function. MAPINSPECT (https://mapinspect.software.informer.com) was used to map the chromosomal locations of *UGT*, *SCPL*, *CXE*, and related genes in *R. chingii*.

The duplication patterns of *UGT*, *SCPL*, and *CXE* genes were analyzed using MCSCANX (Wang et al., 2012). Then whole-genome BLASTP analysis of *R. chingii* and other Rosaceae was performed using local BLAST. Collinearity analysis files of all protein-coding genes were imported into TBTOOLS software (https://github.com/CJ-Chen/TBtools) to identify syntenic blocks and duplication patterns.

Protein purification and enzymatic reactions

The *CXE* gene (LG02.4102) and the *UGT* gene (LG02.4273) were ligated into the PMAL-C2X vector, and prokaryotic expression was carried out in *E. coli*, followed by MBP Bind resin purification to obtain the recombinant proteins. The *CXE* reaction buffer included 100 mM phosphate-buffered saline (pH 7.5), 0.3 mM ascorbic acid, 0.2 mM βG, and 20 μg recombinant protein in 50 μl. The UGT buffer contained 0.1 M Tris-HCL (pH 8.0), 1.5 mM UDPG, 2 mM gallic acid, and 20 μg recombinant protein. The reaction mixtures were incubated at 35°C and 30°C for 1 h, and the same amount of methanol was added to terminate the reaction. The supernatants from the two reactions were obtained by centrifugation and used for high-performance liquid chromatography analysis, and ultraviolet detection wavelength at 280 nm.

A Thermo Scientific UltiMate 3000 system with a Phenomenex Poroshell HPH-C18 column (2.6 μm, 100 mm × 4.6 mm) was used for high-performance liquid chromatography detection. The column oven temperature was set at 25°C, and the mobile phase consisted of 1% acetic acid (A) and 100% acetonitrile (B). The elution gradient increased linearly from 0 to 10% B (v/v) at 5 min and to 35% B at 20 min, then decreased from 35% to 10% B at 21 min and 1% B at 23 min; it was then maintained at 1% B until 25 min. The flow rate was 0.4 ml min^{-1}.

Extraction and quantitative analysis of polyphenol compounds from *Rubus chingii*

Samples were ground to a fine powder in liquid nitrogen. The powdered samples (0.1 g) were extracted with 1 ml of extraction

solution (80% methanol, 19% water, and 1% hydrochloric acid) in an Eppendorf tube. The supernatant was collected following centrifugation at 4000 ***g*** for 15 min. The residues were re-extracted twice by this method, and the combined extracts were diluted 10-fold for quantitative analysis using Q-TOF LC/MS.

An Agilent Infinity Lab instrument (Agilent Technologies, Palo Alto, CA, USA) with a Poroshell HPH-C18 column (2.7 μm, 200 mm × 4.6 mm) was used in this study. The column temperature was 25°C. The mobile phase consisted of 0.4% acetic acid (A) and 100% acetonitrile (B). The elution gradient increased linearly from 5% to 10% B (v/v) at 10 min, to 12.5% B at 22 min, to 30% B at 42 min, and to 60% B at 45 min. It was maintained at 5% B to 47 min and then 5% B at 50 min. The flow rate was 0.4 ml min^{-1}. Polyphenol compounds were identified based on the data obtained from standard substances or published literature.

ACCESSION CODES AND GENOME LINKS

Sequencing data from this study have been deposited at the National Center for Biotechnology Information. The genome sequence data are available under NCBI BioProject number PRJNA666516 (*Rubus chingii* Hu raw sequence reads, https://www.ncbi.nlm.nih.gov/sra/PRJNA666516). We have uploaded the gene annotation, functional annotation and repeat sequence to the GDR database in time (Sook et al., 2004). The detailed links are as follows: the accession number was tfGDR1051, the access link is https://www.rosaceae.org/publication_datasets and the genome page below has links to various pages was https://www.rosaceae.org/Analysis/11326199.

ACKNOWLEDGEMENTS

We thank all the teachers and students of Xia Tao's research group of Anhui Agricultural University for the helpful discussion and technical support of genome analysis, and thank also to Rongqing Yang and Yuncai Tang for drawing the structure of *Rubus chingii* Hu in this article. This work involved the use of the Key Laboratory of Tea at Anhui Agricultural University Technology Facility. This work was supported by Key Research and Development Projects of Anhui Province, China (Project 202004a06020009).

AUTHOR CONTRIBUTIONS

WYS conceived and designed the experiments; WLJ collected samples and performed the experiments; LT, HGM, GLP, YJY, ZXR, and YQ analyzed the data; RHX, GCY, ZQ, and QT collected data; ZNN, QW, WQ, PXJ, and SY analyzed the statistical data; WYS, WLJ, and LT wrote the article.

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1 Frequency distribution of the 17-mer graph analysis used to estimate the size of the *R. chingii* Hu genome.

Figure S2 GC content distribution of the *R. chingii* genome. The GC content was established using 500 bp sliding windows.

Figure S3 The GC depth distribution of the *R. chingii* genome.

Figure S4 Repeat Masker analysis of the *R. chingii* genome assembly.

Figure S5 Chromosomal mapping of 4 gene families. Colors represent gene density.

Figure S6 Distribution of co-linear genes in hydrolyzed tannin biosynthesis gene clusters in other species of Rosaceae.The red line is collinearity. (A) The HTB gene cluster in *R. chingii* is Synteny block in *R.occidentalis* (B) The HTB gene cluster in *R. chingii* is Synteny block in black raspberry The HTB gene cluster in *R. chingii* is collinear block in *F.vesca* (C) The HTB gene cluster in *R. chingii* is Synteny block in *R.chinensis* (D) The HTB gene cluster in *R. chingii* is collinear block in black raspberry The HTB gene cluster in raspberry is Synteny block in *M.domestica*

Figure S7 Correlation matrix analysis of Genes expression and Hydrolyzed tannins content. Pearson correlation coefficient is calculated by R language (version 4.0.3).

Figure S8 Phylogenetic tree of Tannase gene and UGT gene in *R. chingii* chromosome 2. (A) Phylogenetic tree of tannase gene in *R. chingii* chromosome 2 (B) Phylogenetic tree of UGT gene in *R. chingii* chromosome 2. MEGA7 (https://www.megasoftware.net/) used to construct phylogenetic trees, bootstrap values from 1000.

Table S1 Estimation of genome size of the *Rubus chingii* Hu genome based on 17-mer statistics.

Table S2 Sequencing statistics from the nanopore sequencing.

Table S3 Summary of the genome final assembly after polish in the *R. chingii* genome assembly.

Table S4 Summary of BUSCO analysis results according to the *R. chingii* genome assembly.

Table S5 Contig cluster of seven pseudo-chromosomes length of *R. chingii*.

Table S6 Scaffold clustering statistics in the *R. chingii* genome assembly.

Table S7 Gene annotation statistics of the *R. chingii* genome assembly.

Table S8 Transcriptome support statistics in the *R. chingii* genome assembly.

Table S9 SSR distribution statistics in the *R. chingii* genome assembly.

Table S10 Summary of the SSR searched in the *R. chingii* genome assembly.

Table S11 Repeat annotation of the *R. chingii* genome assembly.

Table S12 Summary of non-protein-coding gene annotations in the *R. chingii* genome assembly.

Table S13 Representative plant sources were selected for gene family identification and evolutionary analysis.

Table S14 The Enriched GO terms for the specific genes in *R. chingii*.

Table S15 The Enriched GO terms for the expanded genes in *R. chingii*.

Table S16 In total, 18 578 duplicated genes were identified in the genome of *R. chingii*.

Table S17 Identification of phenolic compounds in *R. chingii* by Q-TOF-MS.

Table S18 The Relative content of hydrolyzed tannins in *R. chingii* different tissues was evaluated by Q-TOF.

Table S19 The gene family of carboxylesterase, Serine carboxypeptidase-like, UDP-glucosyltransferase and polyphenol oxidase.

Table S20 Species of polyphenols in some Rosaceae plants.

Table S21 A correlation matrix analysis was performed to investigate the correlation between hydrolyzable tannin content and gene expression.

Table S22 Expression patterns of hydrolyzed tannin-related synthesis gene clusters with chromosome 02 in different tissues.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at the detailed links are as follows: the accession number was tfGDR1051, the access link is https://www.rosaceae.org/publication_datasets and the genome page below has links to various pages was https://www.rosaceae.org/Analysis/11326199. *Rubus chingii* Hu raw sequence reads, https://www.ncbi.nlm.nih.gov/sra/PRJNA666516.

DATA AVAILABILITY STATEMENT

All relevant data supporting the findings of this work are available within the manuscript and the supporting materials.

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S.O., Gundlach, H. *et al.* (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, **357**, 93–97.

Baek, S., Choi, K., Kim, G.B., Yu, H.J., Cho, A., Jang, H. *et al.* (2018) Draft genome sequence of wild Prunus yedoensis reveals massive inter-specific hybridization between sympatric flowering cherries. *Genome Biology*, **19**, 127.

Bao, W., Kojima, K.K. & Kohany, O. (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. & Dekker, J. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.

Bickhart, D.M., Rosen, B.D., Koren, S., Sayre, B.L., Hastie, A.R., Chan, S. *et al.* (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, **49**(4), 643–650.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365–370.

Bontpart, T., Cheynier, V., Ageorges, A. & Terrier, N. (2015) BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds. *New Phytologist*, **208**, 695–707.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. & Shendure, J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, **31**, 1119–1125.

Callaway, E. (2013) DNA tool kit goes live online. *Nature*, **495**, 150–151.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. *et al.* (2009) BLAST plus: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chen, S.D., Krinsky, B.H. & Long, M.Y. (2013) New genes as drivers of phenotypic evolution. *Nature Reviews Genetics*, **14**, 645–660.

Chen, S.F., Zhou, Y.Q., Chen, Y.R. & Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, 884–890.

Chen, Y., Chen, Z., Guo, Q., Gao, X., Ma, Q., Xue, Z. *et al.* (2019) Identification of Ellagitannins in the unripe fruit of Rubus Chingii Hu and evaluation of its potential antidiabetic activity. *Journal of Agriculture and Food Chemistry*, **67**, 7025–7039.

Cui, L., Yao, S., Dai, X., Yin, Q., Liu, Y., Jiang, X. *et al.* (2016) Identification of UDP-glycosyltransferases involved in the biosynthesis of astringent taste compounds in tea (Camellia sinensis). *Journal of Experimental Botany*, **67**, 2285–2297.

Daccord, N., Celton, J.M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E. *et al.* (2017) High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*, **49**, 1099–1106.

Dai, X., Liu, Y., Zhuang, J., Yao, S., Liu, L., Jiang, X. *et al.* (2020) Discovery and characterization of tannase genes in plants: roles in hydrolysis of tannins. *New Phytologist*, **226**, 1104–1116.

De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.

Emms, D.M. & Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, **16**, 157.

Etherington, G.J., Heavens, D., Baker, D., Lister, A., McNelly, R., Garcia, G. *et al.* (2020) Sequencing smart: De novo sequencing and assembly approaches for a non-model mammal. *Gigascience*, **9**, giaa045.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Research*, **42**, D222–D230.

Grundhofer, P., Niemetz, R., Schilling, G. & Gross, G.G. (2001) Biosynthesis and subcellular distribution of hydrolyzable tannins. *Phytochemistry*, **57**, 915–927.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidence-Modeler and the program to assemble spliced alignments. *Genome Biology*, **9**, R7.

Han, N., Gu, Y., Ye, C., Cao, Y., Liu, Z. & Yin, J. (2012) Antithrombotic activity of fractions and components obtained from raspberry leaves (Rubus chingii). *Food Chemistry*, **132**, 181–185.

Hummer, K.E. & Janick, J. (2007) Rubus iconography: antiquity to the Renaissance. *Acta Horticulture*, **759**, 89–106.

International Peach Genome & Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S., Marroni, F. *et al.* (2013) The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, **45**, 487–494.

Janick and Moore, 1996**Janick, J. & Moore, J.** (1996) Fruit breeding. In: *Vine and small fruit crops*, Vol. **II**. New York: Wiley, pp. 109–190.

Jennings, D.L. (1988) *Raspberries and blackberries: their breeding, diseases and growth*. Academic Press.

Jibran, R., Dzierzon, H., Bassil, N., Bushakra, J.M., Edger, P.P., Sullivan, S. *et al.* (2018) Chromosome-scale scaffolding of the black raspberry (Rubus occidentalis L.) genome based on chromatin interaction data. *Horticulture Research*, **5**, 8.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R. *et al.* (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, **46**, D335–D342.

Kanehisa, M. & Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**, 27–30.

Kaume, L., Howard, L.R. & Devareddy, L. (2012) The blackberry fruit: a review on its composition and chemistry, metabolism and bioavailability, and health benefits. *Journal of Agriculture and Food Chemistry*, **60**, 5716–5727.

Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J. & Hartung, F. (2016) Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, **44**, e89.

Khater, F., Fournand, D., Vialet, S., Meudec, E., Cheynier, V. & Terrier, N. (2012) Identification and functional characterization of cDNAs coding for hydroxybenzoate/hydroxycinnamate glucosyltransferases co-expressed with genes related to proanthocyanidin biosynthesis. *Journal of Experimental Botany*, **63**, 1201–1214.

Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, **35**, 1547–1549.

Kumar, S., Stecher, G., Suleski, M. & Hedges, S.B. (2017) TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, **34**, 1812–1819.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T. & Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, **35**, 3100–3108.

Langmead, B. & Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Letunic, I. & Bork, P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research*, **46**, D493–D496.

Liu, Y., Gao, L., Liu, L., Yang, Q., Lu, Z., Nie, Z. *et al.* (2012) Purification and characterization of a novel galloyltransferase involved in catechin galloylation in the tea plant (Camellia sinensis). *The Journal of biological chemistry*, **287**, 44406–44417.

Loman, N.J., Quick, J. & Simpson, J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, **12**, 733–735.

Lowe, T.M. & Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**, 955–964.

Marcais, G. & Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Moyer, R.A., Hummer, K.E., Finn, C.E., Frei, B. & Wrolstad, R.E. (2002) Anthocyanins, phenolics, and antioxidant capacity in diverse small fruits: vaccinium, rubus, and ribes. *Journal of Agriculture and Food Chemistry*, **50**, 519–525.

Niemetz, R. & Gross, G.G. (2005) Enzymology of gallotannin and ellagitannin biosynthesis. *Phytochemistry*, **66**, 2001–2011.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. *et al.* (2005) InterProScan: protein domains identifier. *Nucleic Acids Research*, **33**, W116–W120.

Regueiro, J., Sanchez-Gonzalez, C., Vallverdu-Queralt, A., Simal-Gandara, J., Lamuela-Raventos, R. & Izquierdo-Pulido, M. (2014) Comprehensive identification of walnut polyphenols by liquid chromatography coupled to linear ion trap-Orbitrap mass spectrometry. *Food Chemistry*, **152**, 340–348.

Salman-Minkov, A., Sabath, N. & Mayrose, I. (2016) Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, **2**, 16115.

Sanz, M., Cadahia, E., Esteruelas, E., Munoz, A.M., Fernandez De Simon, B., Hernandez, T. *et al.* (2010) Phenolic compounds in cherry (Prunus avium) heartwood with a view to their use in cooperage. *Journal of Agriculture and Food Chemistry*, **58**, 4907–4914.

Senol Cali, D., Kim, J.S., Ghose, S., Alkan, C. & Mutlu, O. (2019) Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Briefings in Bioinformatics*, **20**, 1542–1559.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L. *et al.* (2011) The genome of woodland strawberry (Fragaria vesca). *Nature Genetics*, **43**, 109–116.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Sook, J., Christopher, J., Margaret, S., Zhidian, D., Stephen, F., Ilhyung, C. *et al.* (2004) GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. *BMC Bioinformatics*, **5**, 130.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, **32**, W309–312.

Staszowska-Karkut, M. & Materska, M. (2020) Phenolic composition, mineral content, and beneficial bioactivities of leaf extracts from black currant (Ribes nigrum L.), Raspberry (Rubus idaeus), and Aronia (Aronia melanocarpa). *Nutrients*, **12**, 463.

Tarailo-Graovac, M. & Chen, N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, **4**, Unit 4.10.

Thiel, T., Michalek, W., Varshney, R.K. & Graner, A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics*, **106**, 411–422.

Thompson, M. & Zhao, C. (1993) Chromosome numbers of Rubus species in Southwest China. *Acta Horticulture*, **352**, 493–502.

Tutin, T., Heywood, V., Burges, N., Moore, D., Valentine, D., Walters, S. *et al.* (1980) *Flora Europea*: Cambridge University Press.

VanBuren, R., Bryant, D., Bushakra, J.M., Vining, K.J., Edger, P.P., Rowley, E.R. *et al.* (2016) The genome of black raspberry (Rubus occidentalis). *The Plant Journal*, **87**, 535–547.

VanBuren, R., Wai, C.M., Colle, M., Wang, J., Sullivan, S., Bushakra, J.M. *et al.* (2018) A near complete, chromosome-scale assembly of the black raspberry (Rubus occidentalis) genome. *Gigascience*, **7**, giy094.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A. *et al.* (2010) The genome of the domesticated apple (Malus x domestica Borkh.). *Nature Genetics*, **42**, 833–839.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.

Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li J. *et al.* (2019) Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. *Nature Genetics*, **51**, 224–229.

Wang, Y.P., Tang, H.B., DeBarry, J.D., Tan, X., Li, J.P., Wang, X.Y. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, **40**, e49.

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S. *et al.* (2013) The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Research*, **23**, 396–408.

Xiang, Y., Huang, C.H., Hu, Y., Wen, J., Li, S., Yi, T. *et al.* (2017) Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution*, **34**, 262–281.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yoshida, K., Toyama, Y., Kameda, K. & Kondo, T. (2000) Contribution of each caffeoyl residue of the pigment molecule of gentiodelphin to blue color development. *Phytochemistry*, **54**, 85–92.

Yu, G., Luo, Z., Wang, W., Li, Y., Zhou, Y. & Shi, Y. (2019) Rubus chingii Hu: a review of the phytochemistry and pharmacology. *Frontiers in Pharmacology*, **10**, 799.

Yue, J., Wang, R., Ma, X., Liu, J., Lu, X., Balaso Thakar, S. *et al.* (2020) Full-length transcriptome sequencing provides insights into the evolution of apocarotenoid biosynthesis in Crocus sativus. *Computational and Structural Biotechnology Journal*, **18**, 774–783.

Zeng, H.J., Liu, Z., Wang, Y.P., Yang, D., Yang, R. & Qu, L.B. (2018) Studies on the anti-aging activity of a glycoprotein isolated from Fupenzi (Rubus chingii Hu.) and its regulation on klotho gene expression in mice kidney. *International Journal of Biological Macromolecules*, **119**, 470–476.

Zhang, T.T., Lu, C.L., Jiang, J.G., Wang, M., Wang, D.M. & Zhu, W. (2015) Bioactivities and extraction optimization of crude polysaccharides from the fruits and leaves of Rubus chingii Hu. *Carbohydrate Polymers*, **130**, 307–315.

Zhang, X.Y., Li, W., Wang, J., Li, N., Cheng, M.S. & Koike, K. (2019) Protein tyrosine phosphatase 1B inhibitory activities of ursane-type triterpenes from Chinese raspberry, fruits of Rubus chingii. *Chinese Journal of Natural Medicines*, **17**, 15–21.

Zhu, Q.G., Xu, Y., Yang, Y., Guan, C.F., Zhang, Q.Y., Huang, J.W. *et al.* (2019) The persimmon (*Diospyros oleifera* Cheng) genome provides new insights into the inheritance of astringency and ancestral evolution. *Horticulture Research*, **6**, 138.