DR. XINGTAN ZHANG (Orcid ID : 0000-0002-5207-0882)

Article type : Resource Article

Article, Discoveries section

Title: Chromosome-level reference genome assembly provides insights into aroma biosynthesis in passion fruit (*Passiflora edulis*)

Authors:

Dongna Ma[1,2], Shanshan Dong[3], Shengcheng Zhang[1], Xiuqing Wei[4], Qingjun Xie[5,6], Qiansu Ding[2], Rui Xia[5,7*], Xingtan Zhang[1*]

Affiliations:

[1]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China.

[2] Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, China

[3] Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Sciences, Shenzhen, China

[4] Fruit Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou, China

[5] State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources,

Guangzhou, China
[6] Guangdong Provincial Key Laboratory of Plant Molecular Breeding, Guangzhou, China
[7] Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in South China, Ministry of Agriculture, South China Agricultural University, Guangzhou, China
Corresponding authors*: Xingtan Zhang, Email: tanger_009@163.com;
Rui Xia, Email: rxia@scau.edu.cn;

Keywords: genome assembly, Pacbio HiFi sequencing, WGD, aroma biosynthesis

Abstract

Passion fruit, native to tropical America, is an agriculturally, economically and ornamentally important fruit plant that is well-known for its acid pulp, rich aroma and distinctive flavor. Here, we present a chromosome-level genome assembly of passion fruit by incorporating PacBio long HiFi reads and Hi-C technology. The assembled reference genome is 1.28 Gb size with a scaffold N50 of 126.4 Mb and 99.22% sequences anchored onto nine pseudochromosomes. This genome is highly repetitive, accounting for 86.61% of the assembled genome. A total of 39,309 protein-coding genes were predicted with 93.48% of those being functionally annotated in the public databases. Genome evolution analysis revealed a core eudicot-common γ whole-genome triplication event and a more recent whole-genome duplication event, possibly contributing to the expansion of certain gene families. The 33 rapidly expanded gene families were significantly enriched in the pathways of isoflavone biosynthesis, galactose metabolism, diterpene biosynthesis, and fatty acid metabolism, which might be responsible for the formation of featured flavors in the passion fruit. Transcriptome analysis revealed that genes related to ester and ethylene biosynthesis were significantly up-regulated in the mature fruit and the expression levels of those genes were consistent with the accumulation of volatile lipid compounds. The passion fruit genome analysis improves our understanding of the genome evolution of this species and shed new lights into the molecular mechanism of aroma biosynthesis in passion fruit.

1 Introduction

Cultivation of fruit trees attract broad attention from the horticulturist and agriculturist from all over the world. Development of high-quality reference genomes will provide the genetic bases for cultivar breeding, germplasm conservation, and scientific research for these economically and agriculturally important fruit plants. However, a large proportion of fruit trees have relatively high levels of heterozygosity and contain abundant repetitive sequences (Argout et al., 2011; Zhang, Liu, & Ming, 2014; Zhang et al., 2012), holding back our efforts to study these species. The newly developed long-read sequencing technology and high-throughput chromatin conformation capture

(Hi-C) (Berkum et al., 2010) scaffolding strategy have greatly facilitated chromosomal level assemblies for the fruit trees with complex genomes, laiding the foundation for subsequent in-depth study of the genetic basis of the important agronomic traits of fruit trees.

Whole-genome duplication (WGD) or polyploidization (Jiao et al., 2011) doubled the chromosomes initially and resulted in gene duplication followed by subsequent neo- or sub-functionalization (Sankoff, Zheng, & Zhu, 2010). Almost all fruit plants with sequenced genomes have experienced WGD events (Soltis, Bell, Kim, & Soltis, 2008). As the first sequenced fruit plant genome, the grape genome underwent an ancient whole-genome triplication (WGT) event (Jaillon et al., 2007). The subsequently sequenced apple genome revealed not only an ancient WGT event, but also two additional WGD events (Velasco et al., 2010). The pear genome analysis identified two WGD events (International Peach Genome et al., 2013), the durian and kiwi genomes revealed two (Teh et al., 2017; Wu et al., 2019). WGD events not only increase the size of the plant genome, but also broaden genetic variation and elevated the complexity of transcriptional regulatory, further resulting in increased species diversity.

Passiflora Linn. (Passifloraceae) is the largest genus in the passionflower family with 520 species identified (Araya et al., 2017). *Passiflora edulis* Sims. (Passion fruit), a perennial evergreen climbing vine with its origin in tropical South America, is the most widely planted passion fruit species in the genus *Passiflora* (López-Vargas, Fernández-López, Pérez-Álvarez, & Viuda-Martos, 2013). Its egg-shaped fruit contains yellow juice that resembles an egg yolk, but featured for pleasant and distinctive aroma, providing healthy fruits as well as popular raw materials for beverage industry. Due to the rich varieties of flavonoids, alkaloids and other bio-active ingredients in the passion fruit (Antognoni et al., 2007; Zeraik & Yariwake, 2010), the fruits have also some medicinal properties, such as anti-anxiety, anti-inflammatory, and hypoglycemia (Gupta, Kumar, Chaudhary, Maithani, & Singh, 2012; Sato et al., 2012). Passion fruit germplasm resources are relatively abundant, and the cultivars currently widely grown include purple passion fruit, yellow passion fruit, and hybrids of the both. Passion fruit is an

important crop species, and its research is of great significance for the development of the global agricultural economy.

The use of passion fruit in fruit industry is expanding due to the characterization of the active ingredients and health benefits, which might have driven the consumer market of passion fruit related products. Current research on passionflower is focused on cultivation and breeding, and beverage processing. Knowledge on the adaptation mechanisms, odor synthesis pathways, and genetic evolution is limited due to a lack of genomic data. Here, we present a chromosome-level genome assembly of the purple fruit passion that is the most widely planted passion fruit specie owning to the fruit quality and strong adaptability. This is the first genome available in the genus *Passiflora* and provides new insights into the evolutionary history of the specie and the genetic mechanism underlying the aroma synthesis.

2 Materials and methods

2.1 Sample collection

Passion fruit samples (fresh leaves) for genomic and transcriptomic analyses (stem, root, leaf, and fruits at four different developmental stages) were obtained from a plantation in Xiamen, Fujian Province, China (118°27'E, 24°64'N) and stored at -80 °C prior to DNA extraction.

2.2 Genome sequencing

Genomic DNA was isolated from plant leaves using the cetyltriethylammnonium bromide method. An Illumina genomic library was constructed according to Illumina's standard protocol and paired-end reads (2 × 150 bp) sequenced on an Illumina HiSeq X Ten platform. A 15Kb DNA SMRTbell library was constructed and sequenced on a PacBio Sequel2 platform for circular consensus sequencing (CCS). To ensure the validity of the cell fixation, we used tender leaves and checked the nuclei integrity by DAPI staining. The Hi-C library (Berkum et al., 2010) was constructed using HindIII enzyme according to the description of the BioMarker Technologies Company (Xie et al., 2015) and sequenced on an Illumina HiSeq X Ten platform. All of the

sequencing services were provided by Biomarker Technologies Co., Ltd. (Beijing, China)

2.3 Transcriptome sequencing

Four tissues of passion fruit (stem, root, leaf, and four fruits) were collected for RNA-seq analysis, and seven sequencing libraries were constructed from these tissues using an Illumina standard mRNA-seq prep kit with an average insert fragment size of ~250 bp. The libraries were sequenced on an Illumina Novaseq platform with paired-end model.

2.4 Genome survey and assembly

A total of 89.12 Gb of high-quality paired-end reads were obtained by Illumina genomic sequencing (~70.35X coverage, Table S1). The genome size, heterozygosity and repeat content were estimated based on k-mer distribution using 21-mers extracted from the Illumina short reads. The estimated genome size was further validated using flow cytometry. A total of 223.91 Gb raw PacBio subreads were filtered and corrected using pbccs pipeline with default parameters (https://github.com/PacificBiosciences/ccs). The resulted CCS reads were subjected to hifiasm for de novo assembly (https://github.com/chhylp123/hifiasm). We corrected the primary contigs by the Pilon (version1.18) (Utturkar, Klingeman, Hurt, & Brown, 2017) program using 89.12 Gb (70.35×) of Illumina paired-end reads. BWA (version0.7.10-r789) (Li, 2013) and SAMtools (version1.9) (Li et al., 2009) were used for reads alignment and SAM/BAM format conversion. BUSCO (version3.0) (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) program with embryophyta_odb10 database were used to assess the completeness of genome and gene annotation .

2.5 Chromosome assembly using Hi-C

Approximately 96.4 Gb of Hi-C data were generated. The raw data were filtered using perl script as implemented in the software LACHESIS (Burton et al., 2013). BWA software was used to map the Hi-C reads to the draft assembly and uniquely mapped reads were selected for further analysis. We further applied our newly developed ALLHiC (Zhang, Zhang, Zhao, Ming, & Tang,

2019) pipeline to link the contigs into nine pseudo-chromosomes. HiC-pro (version2.10.0) (Servant et al., 2015) program was used to calculate Hi-C mapping rate and evaluate the quality of Hi-C scaffolding.

2.6 Protein-coding gene prediction

Three approaches, as incorporated in MAKER (Cantarel et al., 2008) pipeline, were used to predict the high-quality protein-coding genes: *ab initio* gene predictions, transcript evidence, and homologous-based. For the homology-based prediction models, eight proteomes (including *Manihot esculenta*, *Ricinus communis*, *Salix purpurea*, *Populus trichocarpa*, *Linum usitatissimum*, *Citrus clementina*, *Arabidopsis thaliana*, and *Oryza sativa*) were downloaded from Phytozome database (https://phytozome.jgi.doe.gov/pz/portal.html). The protein sequences of these species were aligned to the *P. edulis* genome by the TBLASTN software, and then the exact gene structures were predicted using GeneWise (Birney & Durbin, 2000) software.

For transcript evidence, the RNA-seq data from different tissues (stem, root, leaf, and fruit) were assembled using Scallop (version0.10.4) (Shao & Kingsford, 2017) software with default parameters. The resulting assembled transcripts were used for training in the SNAP (version2006-07-28) (Bromberg & Rost, 2007), GENEMARK (version4.48_3.60_lic) (Besemer, Lomsadze, & Borodovsky, 2001), and AUGUSTUS (version3.3.3) (Stanke, Steinkamp, Waack, & Morgenstern, 2004). Finally, these tiers of coding evidence were incorporated in the MAKER pipeline to make predictions for high-quality protein-coding genes.

2.7 Functional annotation

Functional annotations of protein-coding genes of *P. edulis* were performed using BLASTP against the public databases with an E-value cut-off of 1.0×10^{-5}: InterPro (Daly, Sutherland-Smith, & Penny, 2013), eggNOG (Huerta-Cepas et al., 2019), Gene Ontology (GO) (Ashburner, Ball, Blake, & Botstein, 2000), Cluster of Orthologous Groups of proteins (COG) (Tatusov, Galperin, Natale, & Koonin, 2000), Swiss-Prot (Bairoch & Apweiler, 2000) and Kyoto

Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2014). InterProScan (version4.8) (Jones et al., 2014) and HMMER (version3.3) (Klingenberg, Aßhauer, Lingner, & Meinicke, 2013) were used to query against the InterPro and Pfam (Bateman et al., 2004) databases to identify the protein domain. Then the GO terms were assigned based on the InterPro or Pfam entry. Additionally, protein coding genes were searched against the KEGG database (release53) to identify possible pathways in which those genes may be involved. GO enrichment and KEGG pathway analysis were performed using an online platform, OmicShare (https://www.omicshare.com/).

2.8 Identification of repetitive elements

In *de novo* prediction, we first customized the genome's repeat sequence library using the pipeline of RepeatModeler (Price, Jones, & Pevzner, 2005), which employed RECON and RepeatScout to get the consensus repeat library. RepeatMasker (Tarailo-Graovac & Chen, 2009) was further used to identify and cluster repeat elements through a homology-based algorithm. TEclass (Abrusan, Grundmann, DeMester, & Makalowski, 2009) software was used to classify the unknown type TEs. The Tandem Repeat Finder (TRF) (Benson, 1999) package was used to identify tandem repeat sequences in the genome. The centromeres and telomeres were predicted using the distribution of tandem repeat sequences on chromosomes based on the same methods that were used in *Oropetium thomaeum* genome (VanBuren et al., 2015).

2.9 Identification of non-coding RNA genes

We used tRNAscan-SE (version1.3.1) (Chan & Lowe, 2019) to identify tRNA genes with eukaryote parameters, RNAmmer (version1.2) (Lagesen et al., 2007) with default parameters to predict rRNA genes, and INFERNAL (version1.1.3) (Nawrocki & Eddy, 2013) with default parameters to annotate snRNA and miRNA genes.

2.10 Reconstruction of the phylogenetic tree

OrthoFinder (version2.4.0) (Emms & Kelly, 2015) was used to identify single-copy

homologous genes in the protein sequences of *P. edulis* and other eight angiosperm species, including *M. esculenta*, *R. communis*, *S. purpurea*, *Po. trichocarpa*, *L. usitatissimum*, *C. clementina*, *A. thaliana*, *O. sativa*. MAFFT (version7.407) (Katoh & Standley, 2013) was used to align each orthologous gene sequences with default parameters. The individual gene alignment was processed using in-house python scripts to extract the conserved regions. The conserved regions of individual genes were concatenated into a supermatrix dataset. Then Modelfinder, as implemented in IQ-TREE (Nguyen, Schmidt, Haeseler, & Minh, 2015) was used to estimate the best substitution models. Finally, RAxML (Stamatakis, 2014) was used to infer the maximum likelihood tree with best-fit substitution model and 500 bootstrap replicates. Divergence time estimates was calculated using r8s with two secondary calibration points obtained from TimeTree database (Kumar, Stecher, Suleski, & Hedges, 2017) (http://www.timetree.org/), ~160 and ~106 million years ago (mya) for the split time of *A. thaliana* and *O. sativa* and *A. thaliana* and *Po. trichocarpa*, respectively.

2.11 Expansion and contraction of gene families

All the deduced proteins were filtered with in-house python scripts to remove alternative splicing and redundant genes, retaining only the longest transcript. All-to-all BLASTP with an E-value cut-off of 1.0×10^{-5} was performed to identify gene family clusters using MCL (Enright, 2002) based on sequence similarity information in BLAST output. Expansion and contraction of the gene family were extrapolated using CAFE (Han, Thomas, Lugo-Martinez, & Hahn, 2013). The neutral mutation rate ρ or the constant number 6.38×10^{-9} was used to estimate the time of divergence.

2.12 Analysis of synteny and whole-genome duplication

The CIRCOS (version0.69-6) (Krzywinski et al., 2009) software was used to visualize gene density, GC content, repeat content and gene synteny on individual pseudochromosomes. To examine WGD in the *P. edulis* genome, all-to-all BLASTP search was used to identify homologous genes with an E-value cut-off of 1.0×10^{-8}. MCScanX (Haibao et al., 2008) software was used to identify collinear blocks. Then synonymous substitution rates (Ks) of the collinear

orthologous gene pairs were calculated using the python script synonymous_calc.py with Nei-Gojobori method (Nei, Gojobori, & Evolution, 1986).

2.13 Transcriptome analysis and identification of specifically expressed genes in fruit

Raw RNA-seq data from five tissues were filtered using Trimmomatic (Bolger, Lohse, & Usadel, 2014) program. This quality check (QC) process trimmed the first 10 and last 5 low-quality bases from each read and discarded reads that were shorter than 36 bp. The resulting clean reads were mapped against to coding sequences (CDS) predicted from passion fruit genome using bowtie (version1.3.0) (Langmead, Trapnell, Pop, & Salzberg, 2009). FPKM (fragments per killobase of exon per million fragments mapped) was further calculated in RSEM (version1.3.2) program (https://github.com/deweylab/RSEM), implemented in Trinity (Grabherr et al., 2011). The differentially expressed genes were identified if log|FC|>1 and *p-value*<0.05. We also identified the fruit specifically expressed genes if these genes were highly expressed in fruit sample (FPKM>30), while have relatively low levels of expression in other tested tissues (stem, root and leaf) with FPKM<4.

3 Results

3.1 Sequencing and assembly of *P. edulis* genome

The genome size of *P. edulis* was estimated to be ~1.27 Gb based on k-mer counting (Figure S1) and ~1.41 Gb by flow cytometry (Figure S2). The k-mer distribution analysis revealed a primary peak at 29× and a secondary peak at 58× of the sequencing depth, suggesting a moderate level of heterozygosity (0.75%) and highly repetitive sequence content (72.68%) in the genome. To obtain a reference genome for *P. edulis*, we generated 223.91 Gb of PacBio long reads using CCS model, which were subsequently corrected into 9.8 Gb high-fidelity (HiFi) reads. A total of ~89.12 Gb (70×) of short reads were also generated by Illumina HiSeq X Ten platform (Table S1). We initially assembled the genome using hifiasm, resulting in a contig level assembly of 1.40 Gb spanning 24,088 contigs. The heterozygous sequences were further identified and removed using Khaper program (https://github.com/lardo/khaper) based on a k-mer counting strategy. The

resulting assembly was 1.28 Gb with a contig N50 of 70 Kb and the longest contig of 6.87 Mb (Table 1). The assembly completeness was assessed using 1,375 plant conserved proteins (embryophyta_odb10) collected in BUSCO program (Simao et al., 2015). Among them, 88.1% of genes completely recalled, 66.8% were single-copy and 21.3% originated from duplication (Table S2). Moreover, Illumina short reads were mapped against our assembly, resulting in a mapping rate of 99.21% for the short reads and a mapping coverage of 99.18% for the assembled genome (Table S3). The genome assembly was further evaluated through a comparison with transcripts assembled from RNA-seq data. 99.16% of bases and 98.93% of sequences from these transcripts were successfully aligned to our assembly (Table S4).

Previous karyotype analysis revealed that passion fruit is a diploid organism with nine pairs of chromosomes (Melo & Guerra, 2003). The scaffold assembly was obtained using ALLHIC pipeline with 75.96 Gb Hi-C clean data. Eleven scaffolds were constructed with a scaffold N50 of 126.4 Mb and the longest scaffold of 281.91 Mb (Table 1). A total of 1.27 Gb sequences were anchored onto nine pseudochromosomes, accounting for 99.22% of the initial assembly (Table S5). In addition, Hi-C data were mapped against Hi-C scaffold assembly, showing 7.43% uniquely mapping rate and 59.31% valid rate of assembled sequences (Table S6). Genome-wide analysis of chromatin interaction showed a well-organized pattern of Hi-C signals along diagonals (Figure 1, Figure S3).

3.2 Genome annotation

We annotated 39,309 protein-coding genes using the MAKER pipeline by incorporating transcriptome, homology and *ab initio* prediction (Table S7). The average gene length was 3,650 bp with 6.88 exons on average. We functionally annotated these genes against published databases, including InterPro, eggNOG, and Swiss-Prot (Table 2), resulting in 93.35%, 88.63%, and 71.42% of the genes functionally assigned, respectively. We further annotated these genes using COG, GO and KEGG databases. Approximately 81.81% of genes have orthologous groups in COG, 70.25% have GO term classification, and 25.58% could be mapped to known plant biological pathways. The BUSCO analysis showed that 85.1% of plant conserved genes presented

in our annotation (Table S8).

Transcription factors (TFs) play important roles in plant development and its response to the environment. We predicted 1,722 transcription factors from the passion fruit genome using PlantTFDB (version3.0) (Jin, Zhang, Kong, Gao, & Luo, 2014). These TFs can be classified into 52 families with bHLH (146 genes), MYB (135 genes), ERF (121 genes), FAR1 (115 genes), and NAC (110 genes) being the top five largest TF superfamilies (Table S9).

We further identified miRNA, sRNA, and snRNA genes by mapping the genome sequences to the Rfam database using INFERNAL, and predicted tRNAs and rRNAs using tRNAscan-SE and RNAmmer, resulting in a prediction of 86 miRNAs, 28 sRNAs, 225 snRNAs, 939 tRNAs, and 784 rRNAs in *P. edulis* genome (Table S10).

P. edulis genome is highly repetitive with a total of 1104.86 Mb repetitive sequences annotated, accounting for 86.3% of the genome length (Figure S4, Table S11). Long terminal repeat (LTR) retrotransposon was the dominant repeat type, taking up 963.67 Mb (75.35%) of the genome sequences (Figure S5). We also performed an analysis the length of Pacbio reads compare to length of LTRs in the genome. The maximum and N50 length of pacbio reads are 291,409 bp and 15,170 bp, respectively (Table S12). The maximum and N50 length of LTRs are 45,317 bp and 2,567 bp, respectively (Table S12). It means that most PacBio reads are longer than LTRs, and are sufficient to span the LTRs. LTRs consist of two major types, Ty1/Copia and Ty3/Gypsy, representing 15.09% and 42.67% of the assembled genome, respectively. Non-LTR retrotransposons, including LINE and SINE, accounted for a small proportion of genome sequences, 4.26% and 0.1%, respectively. In addition, a total of 28,229 tandem repeats were identified, accounting for 53.94 Mb (4.22%) of genome sequences. The 33 putative centromeric fragments were detected on the passion fruit chromosomes (Table S13). All chromosomes are distributed except for the deletion of Chr7 (Table S13). We also identified 6 putative telomeric fragments, among which three were predicted on Chr1, while only one was detected in Chr2, Chr4

315 and Chr8 (Table S14).

316

317 **3.3 Evolutionary history and whole genome duplication**

318 To analyze co-linear relationships within the passion fruit genome, we identified homologous
319 proteins using BLASTP and syntenic blocks using MCScanX. 423 co-linear blocks with 7,776
320 gene pairs were found (Table S15). Syntenic relationship was shown in the CIRCOS plot along
321 with GC content, gene density, distribution of TEs and gene expressions (Figure 2).

322

323 To study the evolutionary history and divergence time of passion fruit, we performed
324 comparative genomic analysis of passion fruit with the genomes of the eight selected angiosperm
325 species, including five Malpighiales plants (*L. usitatissimum*, *M. esculenta, Po. trichocarpa*, *R.*
326 *communis*, *S. purpurea*), a *Citrus* plant (*C. clementina*), and two model organisms (*A. thaliana* and
327 *O. sativa*). A total of 40,345 gene families were identified from 417,083 genes (Table S16). In the
328 passion fruit genome, 13,972 gene families with 8,106 single-copy genes were identified from
329 39,309 genes. Compared to the other five Malpighiales plants, passion fruit has 760 unique gene
330 families (Figure 3). GO enrichment analysis showed that these unique gene families were related
331 with RNA modification, DNA integration, RNA-directed DNA polymerase activity, DNA
332 recombination, calcium-dependent phospholipid binding, unsaturated fatty acid biosynthetic
333 process and defense response (Table S17). KEGG analysis showed that most of these unique gene
334 families were clustered in the pathways of unsaturated fatty acids biosynthesis, fatty acid
335 metabolism, isoflavonoid biosynthesis, thiamine metabolism, flavonoid biosynthesis,
336 phenylpropanoid biosynthesis, and amino sugar and nucleotide sugar metabolism (Figure S6,
337 Table S18). The expression and regulation of these passion fruit unique gene families might likely
338 contribute to the special flavor of its fruit.

339

340 We identified 79 high-quality single-copy orthologous genes from the aforementioned nine
341 plant genomes. Maximum likelihood analysis of these genes recovered a close relationship of
342 passion fruit with *Po. trichocarpa* and *S. purpurea* (Figure 4a). Molecular dating using r8s with

fossil calibration indicated that *P. edulis* split from the ancestor of *Po. trichocarpa* and *S. purpurea* at ~64.41 mya. The Ks values of the orthologs among specie pairs (Table S15) revealed a peak Ks distribution of 0.81 for *P. edulis* - *Po. trichocarpa* and 0.87 for *P. edulis* - *S. purpurea*, corresponding to the divergence times of 63.48 mya and 68.18 mya, respectively (Figure 4b). The divergence times between passion fruit and other two Malpighiales species were 105.80 mya (Ks=1.35, *P. edulis* - *L. usitatissimum*) and 66.61 mya (Ks=0.85, *P. edulis* - *M. esculenta*), respectively.

WGD events were investigated in passion fruit genome based on the distribution of Ks values between orthologs (Figure 4b). Consistent with previously reported results (Blanc & Wolfe, 2004), *A. thaliana* genome revealed an ancient WGD event with a peak value of ~0.76. The *P. edulis* genome recovered two peaks for Ks distributions, one peak at 0.59 (Figure S7) representing a relatively recent WGD event that occurred at ~46.24 mya, and another at 1.59 representing a well-known whole genome triplication event (γ) shared by most dicotyledonous plants (Wu, Han, & Jiao, 2020). During the evolution of plant genomes, the frequency of WGD and polyploidization is higher than that in mammals, resulting in a large proportion of duplicated genes and repetitive sequences retained in plant genomes (Lockton & Gaut, 2005). The two WGD events contributed to the relatively high level of duplication (21.3%) in our BUSCO analysis (Table S2).

The expansion and contraction of gene families might play an essential role in the evolution of passion fruit, contributing to phenotypic diversification, the adaptation to the environment and even speciation. Comparative analysis of the gene family expansion and contraction in nine plants (Figure 4a) showed that 1,525 gene families have expanded and 5,239 contracted in the passion fruit genome. Among the 50 significant evolving (p-value < 0.01) gene families, 33 showed significant expansions, and 17 significant contractions. Significantly expanded gene families were searched against InterProScan (Pfam database), which identified several major functional domains, including Acyl-CoA dehydrogenase/oxidase, agglutinin domain, Cystatin domain, SAM-dependent carboxyl methyltransferase, Zinc finger (CCHC-type) domains, and Plant

self-incompatibility S1 (Table S19). KEGG enrichments showed that most of the rapidly expanded gene families were clustered in the pathways of isoflavonoid biosynthesis, galactose metabolism, diterpenoid biosynthesis, fatty acid degradation, and fatty acid metabolism (Table S20). These metabolic processes may be related to the fruit development and its flavor.

3.4 RNA-seq analysis reveals the genetic mechanism of aroma synthesis

Previous study revealed that esters and terpenes were the main volatile compositions underlying the pleasant aroma in passion fruit (Casimir, Kefford, & Whitfield, 1981). To investigate the genes that play important roles in aroma related biosynthesis pathways, we performed transcriptomic analyses of fruits of four different developmental stages (green, intermediate, lightly ripened fruit and ripe fruit) and other vegetative organs, including stem, root and leaf. We first identified 376 genes that were highly expressed in at least one fruit sample (FPKM>30) but down-regulated in other organs with FPKM less than 4 (Figure S8). GO (Table S21) and KEGG analyses (Table S22) of the 376 fruit specifically expressed genes showed that they were significantly enriched in flavonoid biosynthetic process, anthocyanin-containing compound biosynthetic process and leucocyanidin oxygenase activity. Anthocyanins are natural colorants that give fruits their reddish color as well as extra nutritional value.

We further collected 22 genes associated with fruit development and aroma biosynthesis based on previous publications (Table S23) (Gang & David, 2005; Miriam et al., 2009; Schwab, Davidovich-Rikanati, & Lewinsohn, 2008). Among these genes, 45 were highly expressed in the fruits of passion fruit (Figure 5a). We classified the 45 genes into three function categories, aroma formation, fruit ripening, and carbohydrate metabolism. Majority of the genes in the aroma formation group showed stage-specific expression pattern in fruit samples (Figure 5a). For instance, *HMGR*, encoding 3-hydroxy-3-methylglutaryl-CoA reductase, was specifically and highly expressed in fruit stage 1, demonstrating its important role in producing isoprenoid compounds in the early stage of passion fruit development (Figure 5b). Two TPS (trehalose-6-phosphate synthase) genes (*TPS1* and *TPS6*) were highly up-regulated in fruit stage 2,

likely underlying the accumulation of terpenoids in passion fruits. Two *ADH* (alcohol dehydrogenases) gene copies (ZX.01G0025650 and ZX.01G0084850) were specifically expressed in fruit stage 3, which might play important roles in converting hexenal to alcohols in passion fruit. In addition, the expression level of *AAT1*, *ADH3*, *LOX5* and *TPS10* gradually increased during fruit ripening process (from fruit stage 1 to 4), implying their important roles in accumulation of terpenoids and volatile ester compounds (Figure 5b) that eventually contribute to the pleasant aroma in passion fruit.

Genes related with fruit ripening and carbohydrate metabolism were also analyzed. SAMS2, which catalyzes cysteine to S-Adenosylmethionine (Figure 5b), showed a high expression level at the early stage of fruit development. Two *CYP71B34* copies and two *PMEI* were enriched at the stage 2 and 3, respectively. Meanwhile, three *ACO1* copies, one *BLX1* and *ACS* likely contributed to fruit ripening at the mature stage. Three genes (*PFK*, *NpAIdP1* and *MCSA1*) that are involved in carbohydrate metabolism were highly expressed at stage 3 or 4, illustrating their potentially important functions in passion fruit ripening.

4 Discussion

Passion fruit is a popular fruit tree species for its unique aroma and precious nutritional value of its fruit. Previous studies of this species mainly focused on hybrid breeding, cultivation, and beverage product processing, whereas the molecular basis of its aroma formation and the evolutionary background are limited for the lack of a high-quality reference genome. Therefore, we used PacBio long HiFi reads and Hi-C technology to produce a chromosomal level genome for passion fruit, and annotated 39,309 protein-coding genes using the MAKER pipeline. It is the first genome in the family Passifloraceae, and provides an important information base for the study of the genomics of fruit trees worldwide. Meanwhile, it facilitates the discovery, cloning, functional validation and evolutionary analysis of genes related to important traits in this important species.

The biosynthesis of volatile aromatic substances was mainly related to fatty acid metabolism,

phenyl propane metabolism, terpenoid and amino acid synthesis pathway (Gang, 2005). Here, we proposed a potential pathway that might contribute to aroma formation during passion fruit ripening process (Fig. 5c) based on the comparative genomics and transcriptomic studies. The main enzymes involved in the aroma synthesis pathway are alcohol acyltransferase (*AAT*), lipoxygenase (*LOX*), and terpenoid synthases (*TPS*). *AAT* is a critical gene in the fatty acid and amino acid synthesis that directly regulates lipid synthesis. As the fruit matures, the expression of *AAT1* increases rapidly, lipid content also increased significantly (Gonzalez-Aguero et al., 2009). *AAT1* could bind a variety of alcohols to generate lipid diversity. For example, peach fruit *AAT1* specifically binds to the corresponding alcohols to form hexyl acetate, trans-2-hexenyl acetate and cis-3-hexenyl acetate (Xi et al., 2012); Strawberry *AAT1* gene specifically binds to various substrates to form methyl hexanoate, hexyl butyrate, hexyl acetate and octyl acetate (Aharoni, Keizer, Bouwmeester, & Sun, 2000). *AAT1* is also the major player involved in the formation of the aroma of papaya (Balbontin et al., 2010). We identified three *AAT1* genes from passion fruit, whose expression level gradually increased as the fruit matured. *LOX* is a critical gene in the fatty acid synthesis pathway. It recognizes and combines the linoleic acid and linolenic acid, then converts those substrates to aldehydes and alcohols, hence providing raw materials for the synthesis of downstream lipids (Howe, Lee, Itoh, Li, & DeRocher, 2000). Four copies of *LOX* genes obtained from peach fruits revealed that *LOX1* and *LOX5* gene expression was up-regulated and lipid content was significantly increased when the fruit was fully mature (Eduardo et al., 2012; Zhang et al., 2006). We identified that *LOX1* and *LOX5* genes were up-regulated during fruit ripening as well in passion fruit. Terpenoids are the primary aromatic component of certain fruits, and that the aroma of wine is directly related to the *TPS* genes (Jaillon et al., 2007). Four *TPS* genes were identified from 48 candidate genes. The main volatile substances of passion fruit juice include esters, alcohols, ketones, and terpenes. Comparison of volatile components of fruits at different developmental stages indicated that the relative content of esters gradually increased as the fruit matured with the highest proportion of esters being observed in mature fruits (Guo, chan, Xia, Yang, & Zhang, 2017). As the distinctive aroma of passion fruit mainly resulted from the accumulation of lipid compounds, our study suggested that *ATT1*, *LOX1* and *TPS* genes might

play essential roles in the biosynthesis of the aroma substances in passion fruit.

Ethylene is one of the key hormones that regulates fruit ripening. The production of ethylene is a three-step reaction that begins with the transfer of the adenosyl group from ATP to L-methionine by *SAM* synthase. The *ACS* rate-limiting enzyme then converts l-methionine to 1-aminocyclopropane-1-carboxylic acid, and then *ACO* catalyzes the transition from *ACC* to ethylene (Rodrigues, Bianchetti, & Freschi, 2014). The biosynthesis of ethylene is positively correlated with the gene expression of *ACC* and *ACO*, which are efficiently expressed during fruit ripening in tomatoes, and bananas (Jaillon et al., 2007). Previous studies showed that *ACS* gene expression greatly improves the expression level of *AAT1* and *AAT2* genes while increasing the accumulation of volatile esters (Zhu, Rudell & Mattheis, 2008). Consistent with previous studies, our transcriptome results also revealed significant increase of *ACS* and *ACO* expression during fruit ripening. In addition, up-regulation of other ripening-related genes, including *PME* genes and *CYP71B34* were observed in our results. The *PME* genes, encoding pectin esterase, plays an important role in fruit softening in strawberries (Cristina, Ignacio, Pietro, Ángel, & Victoriano, 2004). Meanwhile, the cytochrome P450 *CYP71B34* contributed to banana fruit ripening (Asif et al., 2014). We also found that expression of three carbohydrate metabolism related genes (*PFK*, *ADP* and *MCSA*) were up-regulated significantly during fruit ripening. *PFK*, which is responsible for converting fructose-6-phosphate and ATP to fructose-1,6-bisphosphate and ADP, can be regulated by a variety of activators and inhibitors. *MCSA* belongs to the family of transferases and participates in cysteine metabolism, selenoamino acid metabolism, and sulfur metabolism. These genes can affect the amount of G3P in plant cells, thus regulating the downstream aroma synthesis and ethylene metabolism. Our study confirmed the important roles of these genes in fruit ripening process and allowed new insights into the aroma biosynthesis pathway in passion fruit (Fig. 5c).

Conclusion

In conclusion, we presented a chromosome-level genome assembly of passion fruit with the aid of PacBio long HiFi reads and Hi-C technology. The 1.28 Gb passion fruit genome encoded

39,309 protein-coding genes and 1.1 Gb repetitive sequences with LTRs being the most dominant transposable elements, accounting for 75.35% of the genome sequences. Two whole genome duplication events resulted in gene expansion and neofunctionalization, likely contributing to accumulation of lipid metabolism and special aroma. Comparative transcriptomic analyses identified 376 specifically expressed genes in fruit that might be related with fruit development and ripening, among which 45 candidate genes played important roles in the ethylene and aroma synthesis pathways in fruit development. The genomic data presented here will not only promote molecular breeding of the passion fruit cultivars, but also lay a foundation for future in-depth research of the genetic basis of the fruit aroma formation.

Author contributions

X.Z., and R.X. designed and coordinated the entire project. D.M., and X.Z., led and performed the entire project together. Q.D. performed the collection and processing of samples. D.M, X.Z Q.X. and S.Z. performed the analyses of genome assembly and annotation. D.M., X.W. and X.Z. performed the analyses of genome evolution and gene families. D.M, X.Z., S.D., R.X. and Q.X. participated in manuscript writing and revision. All authors read and approved the final manuscript.

Data availability statement

Genome sequences have been submitted to the National Genomics Data Center (NGDC). PacBio whole-genome sequencing data and Illumina data have been deposited in BioProject/GSA (https://bigd.big.ac.cn/gsa.) under accession codes CRA003002 and the whole-genome assembly and annotation data have been deposited in CNGBdb (https://db.cngb.org/) under accession codes CNA0017758.

Acknowledgements

This work was supported by a National Natural Science Foundation of China grant (No. 31701874 to X.Z.) and Science and Technology Innovation Team of Fujian Academy of Agricultural Sciences (STIT2017-2-4 to X.W.)

References

Abrusan, G., Grundmann, N., DeMester, L., & Makalowski, W. (2009). TEclass--a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics, 25*(10), 1329-1330. https://doi.org/10.1093/bioinformatics/btp084

Antognoni, F., Zheng, S., Pagnucco, C., Baraldi, R., Poli, F., & Biondi, S. (2007). Induction of flavonoid production by UV-B radiation in *Passiflora quadrangularis* callus cultures. *Fitoterapia, 78*(5), 345-352. https://doi.org/10.1016/j.fitote.2007.02.001

Araya, S., Martins, A. M., Junqueira, N. T. V., Costa, A. M., Faleiro, F. G., & Ferreira, M. E. (2017). Microsatellite marker development by partial sequencing of the sour passion fruit genome (*Passiflora edulis* Sims). *BMC Genomics, 18*(1), 549. https://doi.org/10.1186/s12864-017-3881-5

Argout, X., Salse, J., Aury, J. M., Guiltinan, M. J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S. N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J. F., Sabot, F., Kudrna, D., Ammiraju, J. S., Schuster, S. C., Carlson, J. E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Berard, A., Viot, C., Boccara, M., Risterucci, A. M., Guignon, V., Sabau, X., Axtell, M. J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tahi, M., Akaza, J. M., Pitollat, B., Gramacho, K., D'Hont, A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W. R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S., & Lanaud, C. (2011). The genome of *Theobroma cacao*. *Nature Genetics, 43*(2), 101-108. https://doi.org/10.1038/ng.736

Asaph Aharoni, Leopold C. P. Keizer, Harro J. Bouwmeester, & Sun, Z. (2000). Identification of the *SAAT* gene involved in strawberry flavor biogenesis by use of DNA microarrays. *The Plant Cell, 12*(5),

647-661. https://doi.org/10.2307/3870992

Ashburner, M., A. Ball, C., A. Blake, J., & Botstein, D. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics, 25*, 25-29. https://doi.org/10.1038/75556

Asif, M., Lakhwani, D., Pathak, S., Gupta, P., Bag, S. K., Nath, P., & Trivedi, P. J. B. P. B. (2014). Transcriptome analysis of ripe and unripe fruit tissue of banana identifies major metabolic networks involved in fruit ripening process. *BMC Plant Biology*, *14*(1), 316. https://doi.org/10.1186/s12870-014-0316-1

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research, 28*(1), 45-48. https://doi.org/10.1093/nar/28.1.45

Balbontin, C., Gaete-Eastman, C., Fuentes, L., Figueroa, C. R., Herrera, R., Manriquez, D., Latche, A., Pech, J. C., & Moya-Leon, M. A. (2010). *VpAAT1*, a gene encoding an alcohol acyltransferase, is involved in ester biosynthesis during ripening of mountain papaya fruit. *Journal of agricultural and food chemistry, 58*(8), 5114-5121. https://doi.org/10.1021/jf904296c

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Research,* *32*(suppl_1), 138-141. https://doi.org/10.1093/nar/gkh121

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research, 27*(2), 573-580. https://doi.org/10.1093/nar/27.2.573

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.

Nucleic Acids Research, 29(12), 2607-2618. https://doi.org/10.1093/nar/29.12.2607

Birney, E., & Durbin, R. (2000). Using GeneWise in the *Drosophila* annotation experiment. *Genome Research, 10*, 547-548. https://doi.org/10.1101/gr.10.4.547

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Bromberg, Y., & Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research, 35*(11), 3823-3835. https://doi.org/10.1093/nar/gkm238

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 Biotechnol, 31*(12), 1119-1125. https://doi.org/10.1038/nbt.2727

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research, 18*(1), 188-196. https://doi.org/10.1101/gr.6743907

Casimir, D. J., Kefford, J. F., & Whitfield, F. B.(1981). Technology and flavor chemistry of passion fruit juices and concentrates. *Advances in Food Research, 27*, 243-295. https://doi.org/10.1016/S0065-2628(08)60300-6

Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods in Molecular Biology, 1962*, 1-14. https://doi.org/10.1007/978-1-4939-9173-0_1

Cristina, C., Ignacio, d. l. F. J., Pietro, I., Ángel, B. M., & Victoriano, V. J. J. o. E. B. (2004). Pectin esterase gene family in strawberry fruit: study of FaPE1, a ripening-specific isoform*. *Journal of Experimental Botany*, *55*(398), 909-918. https://doi.org/10.1093/jxb/erh102

574 Daly, T. K., Sutherland-Smith, A. J., & Penny, D. (2013). In silico resurrection of the major vault protein
575 suggests it is ancestral in modern eukaryotes. *Genome Biology and Evolution, 5*(8), 1567-1583.
576 https://doi.org/10.1093/gbe/evt113
577 de Melo, N. F., & Guerra, M. (2003). Variability of the 5S and 45S rDNA sites in *Passiflora* L. species with
578 distinct base chromosome numbers. *Annals of Botany, 92*(2), 309-316.
579 https://doi.org/10.1093/aob/mcg138
580 Eduardo, I., Chietera, G., Pirona, R., Pacheco, I., Troggio, M., Banchi, E., Bassi, D., Rossini, L., Vecchietti,
581 A., & Pozzi, C. (2012). Genetic dissection of aroma volatile compounds from the essential oil of
582 peach fruit: QTL analysis and identification of candidate genes using dense SNP maps. *Tree*
583 *Genetics & Genomes, 9*(1), 189-204. https://doi.org/10.1007/s11295-012-0546-z
584 Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons
585 dramatically improves orthogroup inference accuracy. *Genome Biology, 16*, 157.
586 https://doi.org/10.1186/s13059-015-0721-2
587 Enright, A. J. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids*
588 *Research, 30*(7), 1575-1584. https://doi.org/10.1093/nar/30.7.1575
589 Blanc, G., & Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during
590 arabidopsis evolution. *The Plant Cell*, *16*(7), 1679-1691. https://doi.org/10.1105/tpc.021410
591 Gang, & David, R. (2005). Evolution of flavors and scents. *Annual Review of Plant Biology, 56*(1), 301-325.
592 https://doi.org/10.1146/annurev.arplant.56.032604.144128
593 Gonzalez-Aguero, M., Troncoso, S., Gudenschwager, O., Campos-Vargas, R., Moya-Leon, M. A., &
594 Defilippi, B. G. (2009). Differential expression levels of aroma-related genes during ripening of

apricot (*Prunus armeniaca* L.). *Plant Physiol Biochem, 47*(5), 435-440. https://doi.org/10.1016/j.plaphy.2009.01.002

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol, 29*(7), 644-652. https://doi.org/10.1038/nbt.1883

Gregg A. Howe, Gyu In Lee, Aya Itoh, Lei Li, & DeRocher, A. E. (2000). Cytochrome P450-dependent metabolism of oxylipins in tomato. Cloning and expression of allene oxide synthase and fatty acid hydroperoxide lyase. *Plant Physiology, 123*(2), 711-724. https://doi.org/10.1104/pp.123.2.711

Guo, Y. F., Hui chan, W. U., Xia, Y., Yang, D. P., & Zhang, Y. J. (2017). Volatiles in juice of passion fruitsat different developmental stages. *Fujian Journal of Agricultural Ences*. https://doi.org/10.19303/j.issn.1008-0384.2017.03.014

Gupta, R. K., Kumar, D., Chaudhary, A. K., Maithani, M., & Singh, R. (2012). Antidiabetic activity of *Passiflora incarnata* Linn. in streptozotocin-induced diabetes in mice. *Journal of Ethnopharmacology, 139*(3), 801-806. https://doi.org/10.1016/j.jep.2011.12.021

Haibao, Tang, John, Bowers, Xiyin, Wang, Ray, Ming, & Science, M. J. (2008). Synteny and collinearity in plant genomes. *Science, 320*(5875), 486-488. https://doi.org/10.1126/science.1153917

Han, M. V., Thomas, G. W., Lugo-Martinez, J., & Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution, 30*(8), 1987-1997. https://doi.org/10.1093/molbev/mst100

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research, 47*(1), 309-314. https://doi.org/10.1093/nar/gky1085

International Peach Genome, I., Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M. T., Grimwood, J., Cattonaro, F., Zuccolo, A., Rossini, L., Jenkins, J., Vendramin, E., Meisel, L. A., Decroocq, V., Sosinski, B., Prochnik, S., Mitros, T., Policriti, A., Cipriani, G., Dondini, L., Ficklin, S., Goodstein, D. M., Xuan, P., Del Fabbro, C., Aramini, V., Copetti, D., Gonzalez, S., Horner, D. S., Falchi, R., Lucas, S., Mica, E., Maldonado, J., Lazzari, B., Bielenberg, D., Pirona, R., Miculan, M., Barakat, A., Testolin, R., Stella, A., Tartarini, S., Tonutti, P., Arus, P., Orellana, A., Wells, C., Main, D., Vizzotto, G., Silva, H., Salamini, F., Schmutz, J., Morgante, M., & Rokhsar, D. S. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics, 45*(5), 487-494. https://doi.org/10.1038/ng.2586

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.

637 F., Weissenbach, J., Quetier, F., Wincker, P., & French-Italian Public Consortium for Grapevine
638 Genome, C. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major
639 angiosperm phyla. *Nature, 449*(7161), 463-467. https://doi.org/10.1038/nature06148
640 Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P.,
641 Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma,
642 H., Leebens-Mack, J., & dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and
643 angiosperms. *Nature, 473*(7345), 97-100. https://doi.org/10.1038/nature09916
644 Jin, J., Zhang, H., Kong, L., Gao, G., & Luo, J. (2014). PlantTFDB 3.0: a portal for the functional and
645 evolutionary study of plant transcription factors. *Nucleic Acids Research, 42*(Issue D1), 1182-1187.
646 https://doi.org/10.1093/nar/gkt1016
647 Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A.,
648 Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez,
649 R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification.
650 *Bioinformatics, 30*(9), 1236-1240. https://doi.org/10.1093/bioinformatics/btu031
651 Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information,
652 knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research, 42*(Issue D1),
653 199-205. https://doi.org/10.1093/nar/gkt1076
654 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7:
655 improvements in performance and usability. *Molecular Biology and Evolution, 30*(4), 772-780.
656 https://doi.org/10.1093/molbev/mst010
657 Klingenberg, H., Aßhauer, K. P., Lingner, T., & Meinicke, P. (2013). Protein signature-based estimation of

metagenomic abundances including all domains of life and viruses. *Bioinformatics, 29*(8), 973-980. https://doi.org/10.1093/bioinformatics/btt077

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. J. G. R. (2009). Circos: An information aesthetic for comparative genomics. *Genome Resrarch, 19*, 1639-1645. https://doi.org/10.1101/gr.092759.109

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution, 34*(7), 1812-1819. https://doi.org/10.1093/molbev/msx116

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*(9), 3100-3108. https://doi.org/10.1093/nar/gkm160

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology, 10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv, 1303*(3997).

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Lockton, S., & Gaut, B. S. J. T. i. G. (2005). Plant conserved non-coding sequences and paralogue evolution. *TRENDS in Genetics, 21*(1), 60-65. https://doi.org/10.1016/j.tig.2004.11.013

679 López-Vargas, J. H., Fernández-López, J., Pérez-Álvarez, J. A., & Viuda-Martos, M. (2013). Chemical,
680 physico-chemical, technological, antibacterial and antioxidant properties of dietary fiber powder
681 obtained from yellow passion fruit (*Passiflora edulis* var. *flavicarpa*) co-products. *Food Research*
682 *International, 51*(2), 756-763. https://doi.org/10.1016/j.foodres.2013.01.055
683 Miriam, González, Carlos, Gaete, E., Mónika, Valdenegro, Carlos, R, Figueroa, & Lida. (2009). Aroma
684 development during ripening of *Fragaria chiloensis* fruit and participation of an alcohol
685 acyltransferase (*FcAAT1*) gene. *Journal of Agricultural & Food Chemistry*. *57*(19), 9123-9132.
686 https://doi.org/ 10.1021/jf901693j
687 Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics,*
688 *29*(22), 2933-2935. https://doi.org/10.1093/bioinformatics/btt509
689 Nei, M., Gojobori, T. J. M. B., & Evolution. (1986). Simple methods for estimating the numbers of
690 synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5),
691 418-426. https://doi.org/10.1093/oxfordjournals.molbev.a040410
692 Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective
693 stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and*
694 *Evolution, 32*(1), 268-274. https://doi.org/10.1093/molbev/msu300
695 Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). *De novo* identification of repeat families in large
696 genomes. *Bioinformatics, 21*(1), 351-358. https://doi.org/10.1093/bioinformatics/bti1018
697 Rodrigues, M. A., Bianchetti, R. E., & Freschi, L. (2014). Shedding light on ethylene metabolism in higher
698 plants. *Front Plant Science, 5*(665), 1-16. https://doi.org/10.3389/fpls.2014.00665
699 Sankoff, D., Zheng, C., & Zhu, Q. (2010). The collapse of gene complement following whole genome

duplication. *BMC Genomics, 11*(1), 313. https://doi.org/10.1186/1471-2164-11-313

Sato, A. C., Andrade, S. A., Brito, M. V., Miranda, A., Sampaio, M. U., de Abreu Maffei, F. H., & Oliva, M. L. (2012). Effects of compounds from *Passiflora edulis* Sims f. Flavicarpa juice on blood coagulation and on proteolytic enzymes. *Protein and Peptide Letters, 19*(5), 501-508. https://doi.org/10.2174/092986612800191053

Schwab, W., Davidovich-Rikanati, R., & Lewinsohn, E. (2008). Biosynthesis of plant-derived flavor compounds. *Plant Journal for Cell & Molecular Biology, 54*(4), 712-732. https://doi.org/10.1111/j.1365-313X.2008.03446.x

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology, 16*, 259. https://doi.org/10.1186/s13059-015-0831-x

Shao, M., & Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnol, 35*(12), 1167-1169. https://doi.org/10.1038/nbt.4020

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*(19), 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

Soltis, D. E., Bell, C. D., Kim, S., & Soltis, P. S. (2008). Origin and early evolution of angiosperms. *Annals of the New York Academy of Sciences, 1133*, 3-25. https://doi.org/10.1196/annals.1438.005

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene

finding in eukaryotes. *Nucleic Acids Research, 32*(suppl_2), 309-312. https://doi.org/10.1093/nar/gkh379

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *5*(1), 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research, 28*(1), 33-36. https://doi.org/10.1093/nar/28.1.33

Teh, B. T., Lim, K., Yong, C. H., Ng, C. C. Y., Rao, S. R., Rajasegaran, V., Lim, W. K., Ong, C. K., Chan, K., Cheng, V. K. Y., Soh, P. S., Swarup, S., Rozen, S. G., Nagarajan, N., & Tan, P. (2017). The draft genome of tropical fruit durian (*Durio zibethinus*). *Nature Genetics, 49*(11), 1633-1641. https://doi.org/10.1038/ng.3972

Utturkar, S. M., Klingeman, D. M., Hurt, R. A., Jr., & Brown, S. D. (2017). A case study into microbial genome assembly gap sequences and finishing strategies. *Front Microbiol, 8*(1272), 1-11. https://doi.org/10.3389/fmicb.2017.01272

vanBerkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., & Lander, E. S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments*(39). https://doi.org/10.3791/1869

VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., Mockler, T. C. (2015). Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature, 527*(7579), 508-511. https://doi:10.1038/nature15714

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L. M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J. T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V. T., King, S. T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M. M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A. C., Bus, V., Chagne, D., Crowhurst, R. N., Gleave, A. P., Lavezzo, E., Fawcett, J. A., Proost, S., Rouze, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R. P., Durel, C. E., Gutin, A., Bumgarner, R. E., Gardiner, S. E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F., & Viola, R. (2010). The genome of the domesticated apple (Malus x domestica Borkh.). *Nature Genetics, 42*(10), 833-839. https://doi.org/10.1038/ng.654

Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., Tang, B., Dong, L., Ding, N., Zhang, Q., Bai, Z., Dong, X., Chen, H., Sun, M., Zhai, S., Sun, Y., Yu, L., Lan, L., Xiao, J., Fang, X., Lei, H., Zhang, Z., & Zhao, W. (2017). GSA: genome sequence archive. *Genomics Proteomics Bioinformatics, 15*(1), 14-18. https://doi.org/10.1016/j.gpb.2017.01.001

Wu, H., Ma, T., Kang, M., Ai, F., Zhang, J., Dong, G., & Liu, J. (2019). A high-quality *Actinidia chinensis* (kiwifruit) genome. *Horticulture Research, 6*(1), 117. https://doi.org/10.1038/s41438-019-0202-y

Wu, S., Han, B., & Jiao, Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Molecular Plant, 13*(1), 59-71. https://doi.org/10.1016/j.molp.2019.10.012

763 Xi, W.-p., Zhang, B., Shen, J.-y., Sun, C.-d., Xu, C.-j., & Chen, K.-s. (2012). Intermittent warming alleviated
764 the loss of peach fruit aroma-related esters by regulation of *AAT* during cold storage. *Postharvest*
765 *Biology and Technology, 74*, 42-48. https://doi.org/10.1016/j.postharvbio.2012.07.003
766 Xie, T., Zheng, J. F., Liu, S., Peng, C., Zhou, Y. M., Yang, Q. Y., & Zhang, H. Y. (2015). *De novo* plant
767 genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*.
768 *Molecular Plant, 8*(3), 489-492. https://doi.org/10.1016/j.molp.2014.12.015
769 Zeraik, M. L., & Yariwake, J. H. (2010). Quantification of isoorientin and total flavonoids in *Passiflora edulis*
770 fruit pulp by HPLC-UV/DAD. *Microchemical Journal, 96*(1), 86-91.
771 https://doi.org/10.1016/j.microc.2010.02.003
772 Zhang, B., Chen, K., Bowen, J., Allan, A., Espley, R., Karunairetnam, S., & Botany, a. I. F. J. J. o. E. (2006).
773 Differential expression within the *LOX* gene family in ripening kiwifruit. *Journal of Experimental*
774 *Botany* , *57*(14), 3825-3836. https://doi.org/10.1093/jxb/erl151
775 Zhang, J., Liu, J., & Ming, R. (2014). Genomic analyses of the CAM plant pineapple. *Journal of*
776 *Experimental Botany, 65*(13), 3395-3404. https://doi.org/10.1093/jxb/eru101
777 Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B., Yang, W., Tao, Y., Wang, J., Yuan, Z., Fan, G., Xing, Z.,
778 Han, C., Pan, H., Zhong, X., Shi, W., Liang, X., Du, D., Sun, F., Xu, Z., Hao, R., Lv, T., Lv, Y.,
779 Zheng, Z., Sun, M., Luo, L., Cai, M., Gao, Y., Wang, J., Yin, Y., Xu, X., Cheng, T., & Wang, J.
780 (2012). The genome of *Prunus mume*. *Nature Communication, 3*(1), 1318.
781 https://doi.org/10.1038/ncomms2290
782 Zhang, X., Zhang, S., Zhao, Q., Ming, R., & Tang, H. (2019). Assembly of allele-aware, chromosomal-scale
783 autopolyploid genomes based on Hi-C data. *Nature Plants, 5*(8), 833-845.

784 https://doi.org/10.1038/s41477-019-0487-8

785 Zhu, Y., Rudell, D. R., Mattheis, J. P. J. P. B., & Technology. (2008). Characterization of cultivar differences

786 in alcohol acyltransferase and 1-aminocyclopropane-1-carboxylate synthase gene expression and

787 volatile ester emission during apple fruit maturation and ripening. *Postharvest Biology and*

788 *Technology 49*(3), 330-339. https://doi.org/10.1016/j.postharvbio.2008.03.015

Figures and legends

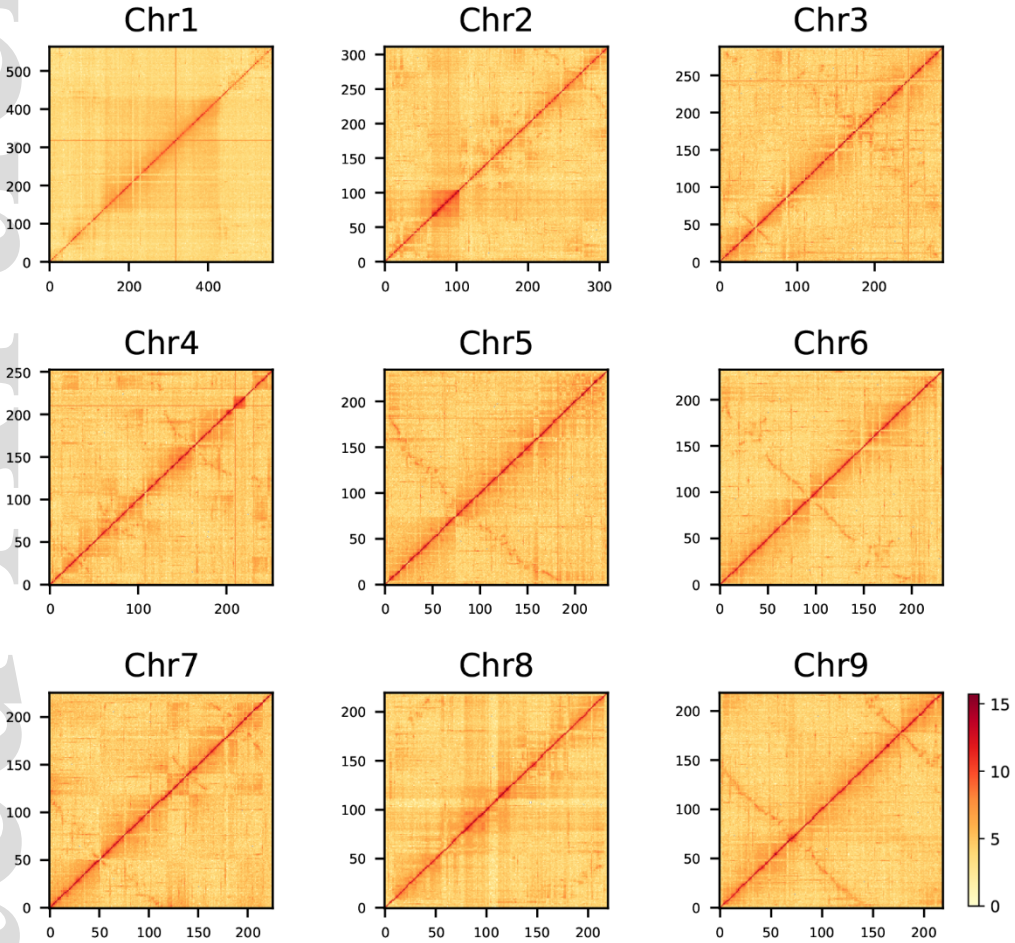


Figure 1. Genome-wide analysis of chromatin interactions at 500-kb resolution in *P. edulis* genome.

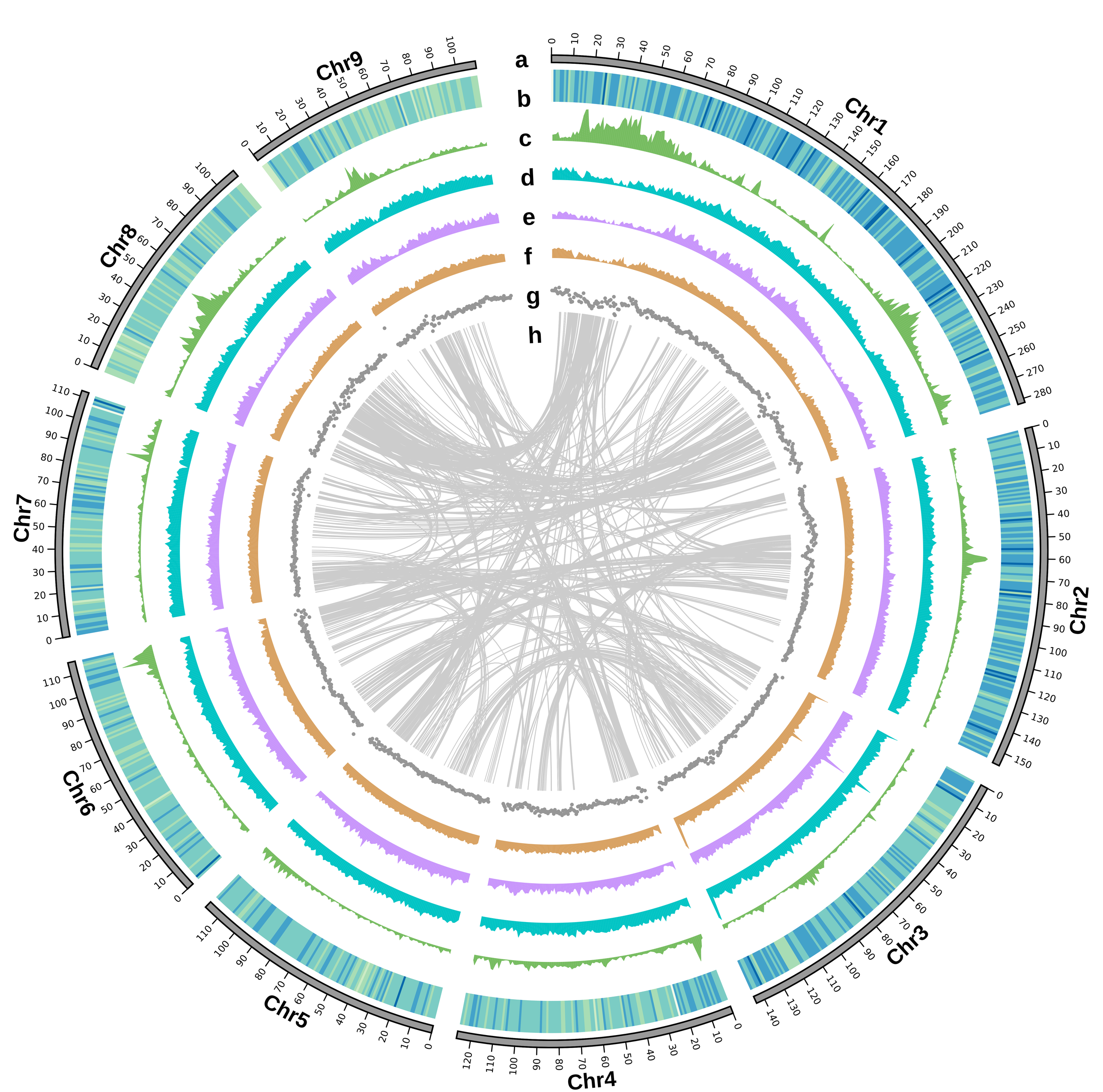

a
b
c
d
e
f
g
h
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9

Figure 2. Distribution of the elements on the chromosomes of *P. edulis*. (a) Chromosomes karayotype. (b) Gene expression analyzed using RNA-seq. (c) Gene density. (d) DNA transposable elements. (e) LTR/Copia transposable elements. (f) LTR/Gypsy transposable elements. (g) Distribution of GC content in the genome. (h) Schematic presentation of major inter-chromosomal relationships in the *P. edulis* genome.

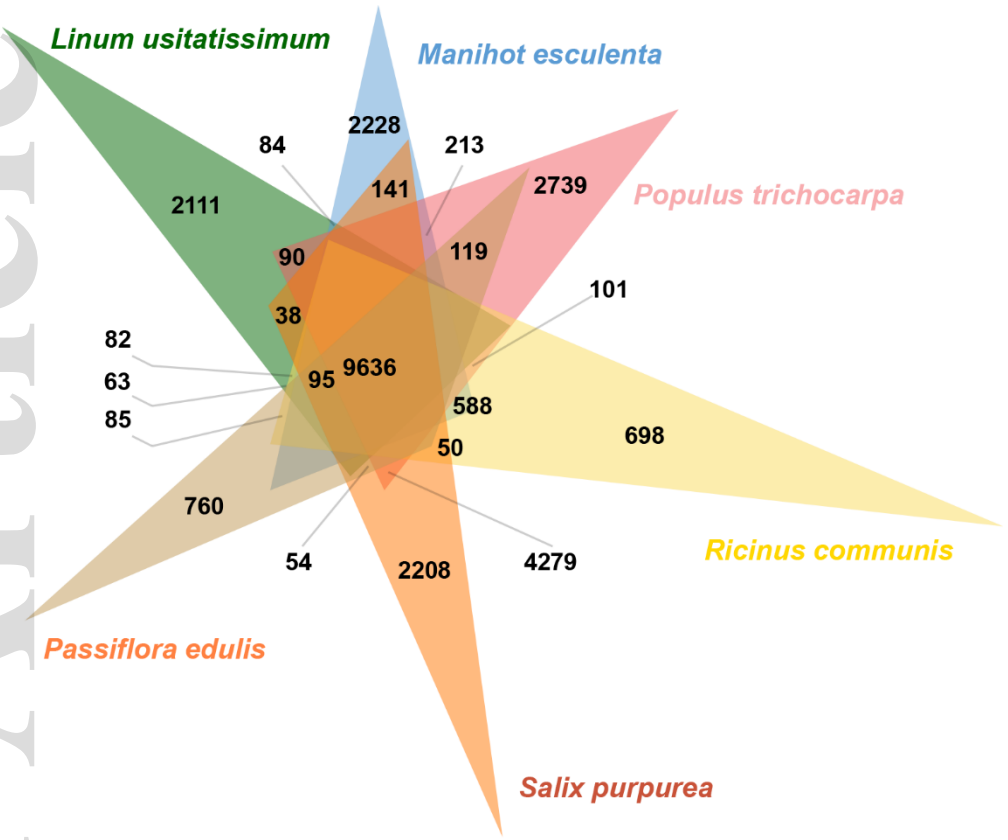
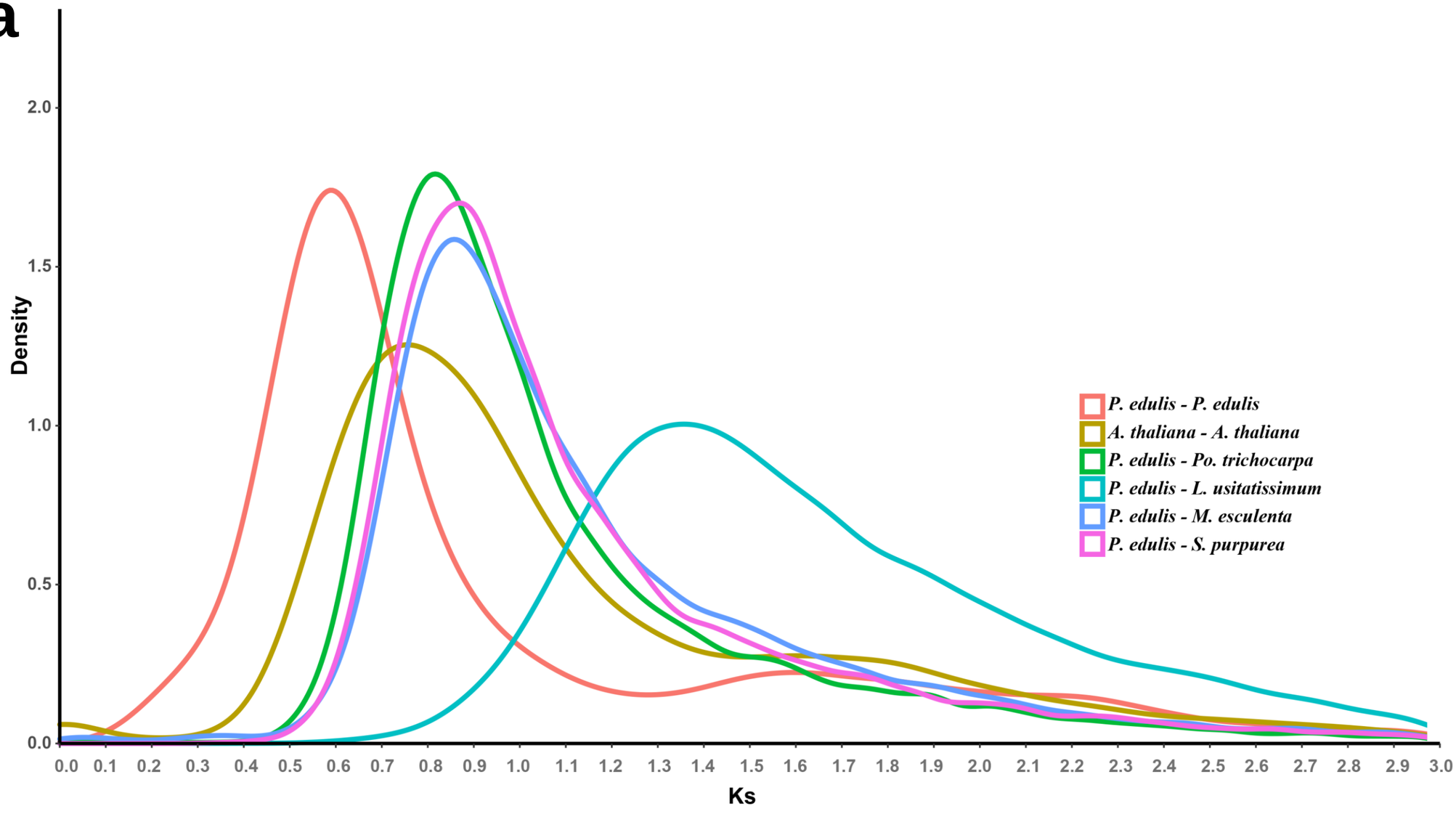


Figure 3. Venn diagram of gene families in *P. edulis* and five other Malpighiales plants.

a


Density
2.0
1.5
1.0
0.5
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0
Ks
P. edulis - P. edulis
A. thaliana - A. thaliana
P. edulis - Po. trichocarpa
P. edulis - L. usitatissimum
P. edulis - M. esculenta
P. edulis - S. purpurea

b

Gene families
Expansion / Contraction

Manihot esculenta
(+2670/-1729)

Ricinus communis
(+1714/-2634)

Salix purpurea
(+1052/-2473)

Populus trichocarpa
(+3383/-520)

Passiflora edulis
(+1525/-5239)

Linum usitatissimum
(+5289/-3285)

Citrus clementina
(+740/-6144)

Arabidopsis thaliana
(+1454/-5294)

Oryza sativa
(+2544/-4735)

(+434/-1723)
(+377/-526)
(+2901/-1361)
(+49/-21)
(+21/-4)
(+2901/-1361)
(+149/-787)

Million years ago
175 150 125 100 75 50 25 0

Figure 4. Evolutionary and comparative genomic analyses. (a) Density distribution of Ks values between syntenic genes of compared genomes. (b) Phylogenetic relationship of *Passiflora edulis*, *Linum usitatissimum, Manihot esculenta, Populus trichocarpa, Ricinus communis, Salix purpurea, Citrus clementina, Arabidopsis thaliana and Oryza sativa*. The divergence times among different plant species are labeled in the bottom. The numbers on each branch represent expansion (red) and contraction (blue) of gene families. Stars in green represent potential whole-genome duplication events occurred in the passion fruit genome.

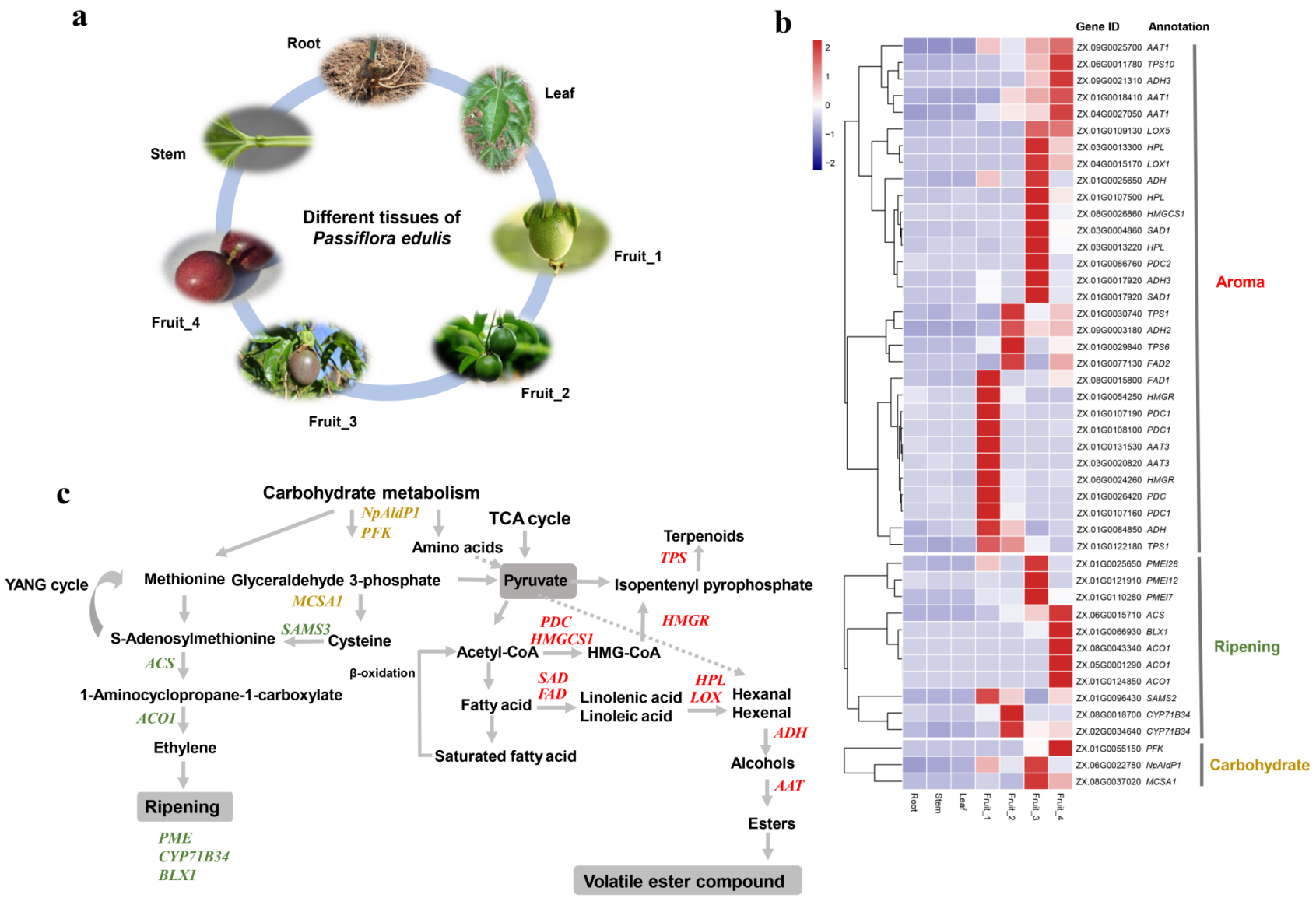


Figure 5. Putative biosynthetic pathways of ethylene and volatile lipid compounds in *P. edulis*. (a) Transcriptome sequencing tissues that were used in this study. (b) Highly expressed genes in *P. edulis* mature fruit compared to other tissues. (c) The main biosynthetic pathway of lipid volatiles. *AAT*, *ADH*, *LOX* and *TPS* represent genes encoding alcohol acyltransferase, alcohol dehydrogenase, lipoxygenase, and terpenoid synthases, respectively. The different colors of fonts represent genes in the aroma (red), ripening (green) and carbohydrate (yellow) metabolism synthesis pathways.

Table 1. Global statistics for assembly of *P. edulis* genome

Assembly feature	Statistic
PacBio sequencing assembly	
Number of contigs	20,971
Contig N50	70,158 bp
Contig N90	31,682 bp
Longest contig	6.87 Mb
Average length	60,948 bp
Total contig length	1.29 Gb
Hi-C scaffolding assembly	
Number of scaffolds	11
Scaffold N50	126.40 Mb

Scaffold N90	109.88 Mb
Longest scaffold	281.91 Mb
Average length	116.39 Mb
Total scaffold length	1.28 Gb

Table 2. Annotation statistics for the *P. edulis* genome

Annotation statistics for the genome	**Number**	**Percent (%)**
Total protein	39,309	100
InterPro	36,694	93.35
eggNOG	34,840	88.63
Swiss-Prot	28,076	71.42
GO	27,614	70.25
COG	32,159	81.81
KEGG	10,057	25.58
In all databases	9,778	24.87
In at least one database	36,745	93.48