A *de novo* assembly of the sweet cherry (*Prunus avium* cv. Tieton) genome using linked-read sequencing technology

Jiawei Wang [Equal first author, 1], **Weizhen Liu** [Corresp., Equal first author, 2], **Dongzi Zhu** [1], **Xiang Zhou** [3], **Po Hong** [1], **Hongjun Zhao** [1], **Yue Tan** [1], **Xin Chen** [1], **Xiaojuan Zong** [1], **Li Xu** [1], **Lisi Zhang** [1], **Hairong Wei** [1], **Qingzhong Liu** [Corresp. 1]

[1] Scientific Observation and Experiment Station of Fruits in Huang-huai area, Ministry of Agriculture, Shandong Institute of Pomology, Taian, Shandong, China

[2] School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei, China

[3] Key Laboratory of Agricultural Animal Genetics, Breeding, and Reproduction of Ministry of Education & Key Laboratory of Swine Genetics and Breeding of Ministry of Agriculture, Huazhong Agricultural University, Wuhan, Hubei, China

Corresponding Authors: Weizhen Liu, Qingzhong Liu
Email address: liuweizhen@whut.edu.cn, qzliu001@126.com

The sweet cherry (*Prunus avium*) is one of the most economically important fruit species in the world. However, there is a limited amount of genetic information available for this species, which hinders breeding efforts at a molecular level. We were able to describe a high-quality reference genome assembly and annotation of the diploid sweet cherry (2n=2x=16) cv. Tieton using linked-read sequencing technology. We generated over 750 million clean reads, representing 112.63 Gb of raw sequencing data. The Supernova assembler produced a more highly-ordered and continuous genome sequence than the current *P. avium* draft genome, with a contig N50 of 63.65 Kb and a scaffold N50 of 2.48 Mb. The final scaffold assembly was 280.33 Mb in length, representing 82.12% of the estimated Tieton genome. Eight chromosome-scale pseudomolecules were constructed, completing a 214 Mb sequence of the final scaffold assembly. *De novo*, homology-based, and RNA-seq methods were used together to predict 30,975 protein-coding loci. 98.39% of core eukaryotic genes and 97.43% of single copy orthologues were identified in the embryo plant, indicating the completeness of the assembly. Linked-read sequencing technology was effective in constructing a high-quality reference genome of the sweet cherry, which will benefit the molecular breeding and cultivar identification in this species.

A *de novo* assembly of the sweet cherry (*Prunus avium* cv. Tieton) genome using linked-read sequencing technology

Jiawei Wang[1#], Weizhen Liu[2#*], Dongzi Zhu[1#], Xiang Zhou[3], Po Hong[1], Hongjun Zhao[1], Yue Tan[1], Xin Chen[1], Xiaojuan Zong[1], Li Xu[1], Lisi Zhang[1], Hairong Wei[1], and Qingzhong Liu[1*]

[1] Scientific Observation and Experiment Station of Fruits in Huang-huai area, Ministry of Agriculture, Shandong Institute of Pomology, Taian, Shandong, China
[2] School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei, China
[3] Key Laboratory of Agricultural Animal Genetics, Breeding, and Reproduction of Ministry of Education & Key Laboratory of Swine Genetics and Breeding of Ministry of Agriculture, Huazhong Agricultural University, Wuhan, China

[#] Jiawei Wang, Weizhen Liu, and Dongzi Zhu contributed equally to this work.

[*]Correspondence Authors:
Dr. Weizhen Liu, email: liuweizhen@whut.edu.cn
Dr. Qingzhong Liu, email: qzliu001@126.com

Abstract: The sweet cherry (*Prunus avium*) is one of the most economically important fruit species in the world. However, there is a limited amount of genetic information available for this species, which hinders breeding efforts at a molecular level. We were able to describe a high-quality reference genome assembly and annotation of the diploid sweet cherry (2n=2x=16) cv. Tieton using linked-read sequencing technology. We generated over 750 million clean reads, representing 112.63 Gb of raw sequencing data. The Supernova assembler produced a more highly-ordered and continuous genome sequence than the current *P. avium* draft genome, with a contig N50 of 63.65 Kb and a scaffold N50 of 2.48 Mb. The final scaffold assembly was 280.33 Mb in length, representing 82.12% of the estimated sweet cherry cv. Tieton genome. Eight chromosome-scale pseudomolecules were constructed, completing a 214 Mb sequence of the final scaffold assembly. *De novo*, homology-based, and RNA-seq methods were used together to predict 30,975 protein-coding loci. 98.39% of core eukaryotic genes and 97.43% of single copy orthologues were identified in the embryo plant, indicating the completeness of the assembly. Linked-read sequencing technology was effective in constructing a high-quality reference genome of the sweet cherry, which will benefit the molecular breeding and cultivar identification in this species.

1. Introduction

The sweet cherry (*Prunus avium*), originated in Asia Minor near the Black Sea and the Caspian Sea. It is known as one of the most economically significant fruit species in the world (Quero-García et al., 2017) and its production in China has increased dramatically over the last three decades with the expansion of acreage dedicated to its cultivation. Recent breeding efforts have focused on improving yield, fruit quality, tree architecture and biotic and abiotic resistance (Aranzana et al., 2019). Sweet cherry and other *Prunus* crops have a long juvenile period, which means that traditional breeding methods are slow to produce improvements (Quero-García et al., 2017). Marker-assisted breeding and genomic selection can speed up the breeding cycle, but these methods require a high-quality reference genome in order to obtain a sufficient amount of genetic variants and to identify the regulatory regions controlling the morphological and physiological characteristics of the plant (Aranzana et al., 2019; Ru et al., 2015). Only one draft genome assembly of sweet cherry cv. Satonishiki (Shirasawa et al., 2017)and one mitochondrial genome sequence of cv. Summit have been reported (Yan et al., 2019), despite the simple genome of the sweet cherry (2n=2x=16). The draft genome of sweet cherry cv. Satonishiki was sequenced using Illumina short-read sequencing technology, resulting in a fragmented assembly of 272.4 Mb with a scaffold N50 of 219.6 Kb (Shirasawa et al., 2017). The linked-read sequencing pipeline developed by 10x Genomics may result in more continuous genomes for the sweet cherry at a lower financial cost (Pollard et al., 2018; Zheng et al., 2016). This technology use a barcoded sequencing library to generate long-range information (preferably >100 kb) and standard short-read sequencing to ensure massive throughput and high accuracy. It was designed for human genome assembly, but has been used effectively in many other animal and plant species, including the wild dog, proso millet pepper and soybean (Armstrong et al., 2018; Hulse-Kemp et al., 2018; Liu et al., 2018; Ott et al., 2018).

We demonstrated that linked-read technology is effective in the *de novo* assembly of the genome of the sweet cheery cv. Tieton, which is the most popular cherry variety in China. The sweet cherry cv. Tieton genome assembly surpasses the cv. Satonishiki genome assembled using Illumina short-reads in continuity, with a tenfold improvement of scaffold N50 (Shirasawa et al., 2017). The high-quality genome assembly and annotation in this study are valuable for genetic marker development and gene mapping, which may improve sweet cherry breeding. Our assembly platform will support future *de novo* genome assemblies for other *Prunus* crops using the linked reads method.

2. Materials and Methods

2.1 Sample and DNA extraction

Leaf samples were collected from the sweet cherry cv. Tieton grown in the experimental orchard of Shandong Institute of Pomology, Taian, Shandong Province, China, and frozen in liquid nitrogen. High-molecular-weight (HMW) genomic DNA (gDNA) was extracted from the frozen leaves using MagAttract HMW DNA Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. The gDNA was quantified using Implen NanoPhotometer P330 (Implen, Munich, Germany) and assessed using agarose gel electrophoresis.

2.2 Chromium library construction and sequencing

The single Chromium library was constructed by CapitalBio Technology Inc. (Beijing, China) using the purified HMW gDNA sample. The library was sequenced in one lane as 150nt-Chromium-linked paired-end reads on an Illumina HiSeq X Ten sequencer (Illumina, http://www.illumina.com/). We filtered out raw reads with >5% undetermined bases (Ns), >30% nucleotides quality score lower than 20, and the adapter sequence overlap > 5 bp.

2.3 *de novo* assembly and evaluation

We estimated the size of the sweet cherry genome based on the k-mer frequency of the sequence data using the k-mer counting program Jellyfish (v.2.0.8) (Marcais & Kingsford, 2011) and GenomeScope (v1.0.0)(Vurture et al., 2017). The genome was assembled and scaffolded using the Supernova assembler (v2.0, https://www.10x genomics.com/). This program links sequencing reads to the originating HMW DNA molecule using barcoded information and constructs phased, whole-genome *de novo* assemblies form the Chromium-prepared library (Weisenfeld et al., 2017). Chromium-linked reads of different sizes (40x, 50x, 60x, 65x, 68x, 70x, and 75x) were used as input data. The assembly, using 70x coverage of the reads, was selected for analysis based on superior quality, and higher contig N50 and scaffold N50. Default parameters were set and two pseudohap assemblies were generated; pseudohap1 was used for further analysis. 150 million reads were sampled and aligned to the assembled genome sequence; the quality of the sweet cherry cv. Tieton genome assembly was evaluated using the Burrows-Wheller Alignment tool (BWA, 0.7.17-r1188) (Li & Durbin, 2009). Core Eukaryotic Genes Mapping Approach (CEGMA, v2.5) (Parra et al., 2007) and Benchmarking Universal Single-

Copy Orthologs (BUSCO, v3.0, embryophyta_odb10) (Simao et al., 2015) were used to assess the completeness of the assembly.

2.4 Chromosome-scale pseudomolecule construction

Scaffolds were assembled using the Supernova assembler and were ordered and oriented using seven previously published sweet cherry genetic maps for the construction of the chromosome-scale pseudomolecules. Five of the seven maps were built by Shirasawa et al. (2017), Peace et al. (2012), Klagges et al. (2013), Calle et al. (2018), and Guajardo et al. (2015). The initials of the first author were used to name their respective maps and the maps are referred to as KS, CP, CK, AC, and VG. The other two maps, named JWF (the framework map of the WxL map) and JWF1 (the second round of the WxL map), were both reported by Wang et al. (2015). Genetic markers and/or flanking sequences for these maps were aligned to the current scaffolds using GMAP (v2018-07-04)(Wu & Watanabe, 2005) as described by Hulse-Kemp et al. (2018). Markers were manually filtered out if they were aligned to more than one scaffold or the same scaffold in different linkage groups. The alignment results of GMAP were fitted into ALLMAPS (v0.8.4)(Tang et al., 2015) to generate the final consensus map and chromosome-scale pseudomolecules. Different weight parameters were tried for the seven linkage maps and the optimal weight settings with the largest number of anchored and oriented scaffolds were : KS =2, CP =3, CK =1, AC =1, VG =1, JWF =1, and JWF1=1.

2.5 Identification of repetitive elements in sweet cherry genome

Homology-based and *de novo* methods were combined to identify repetitive and transposon elements in our final assembly using RepeatMasker (v.4.0.6) (Smit et al., 2016) and RepeatModeler (v.1.0.11) (http://www.repeatmasker.org/RepeatModeler.html).

2.6 RNA-Seq analysis

Total RNA was extracted from the young leaves of a single plant for genome sequencing. The cDNA library was constructed based on the description of Wei et al. (2015) and sequenced by CapitalBio Technology Inc. (Beijing, China) using the Illumina HiSeq 2000 platform. The adapters were trimmed and low-quality reads were removed before the remaining high quality reads were assembled by Trinity (v2.8.5)(Grabherr et al., 2011).

2.7 Non-coding RNA prediction, protein-coding gene prediction and functional annotation

INFERNAL (v1.1.2) (Nawrocki et al., 2009) was used to identify the non-coding RNAs (ncRNAs) in the sweet cherry cv. Tieton genome against the RFAM database (Griffiths-Jones et al., 2005). The tRNAs were identified by tRNAscan-SE (v2.0.5)(Lowe & Eddy, 1997). The rRNAs were identified using RNAmmer (v1.1.2)(Lagesen et al., 2007).

Homology-based, *de novo* and RNA-seq methods were combined to predict the protein-coding genes in sweet cherry cv. Tieton genome. Augustus (v3.3.2)(Keller et al., 2011) and SNAP (v2013-11-29)(Korf, 2004) were used in the *de novo* annotation to predict the protein-coding gene in repeat-masked genome sequences. The predicted genes were annotated by

Genewise (v2.4.1)(Birney et al., 2004) and Exonerate (v2.4.0)(Slater & Birney, 2005). The Program to Assemble Spliced Alignments (PASA, v2.4.1) pipeline (Haas et al., 2003) was used in transcriptome-assistant method with the unigenes assembled by the RNA-seq data. EVidenceModeler (EVM, v1.1.1)(Haas et al., 2008) and PASA were used to combine the predicted results.

2.8 Gene family analysis

OrthoFinder (v2.2.7) (Emms & Kelly, 2015) was used to identify the orthologous genes from 13 plant genomes of the sweet cherry cv. Tieton (*Prunus avium*, Pa), peach (*Prunus persica*, Pp), Chinese plum (*Prunus mume*, Pm), flowering cherry (*Prunus yedoensis*, Py), apple (*Malus x domestica*, Md), pear (*Pyrus bretschneideri*, Pb), black raspberry (*Rubus occidentalis*, Ro), strawberry (*Fragaria vesca*, Fv), rose (*Rosa chinensis*, Rc), orange (*Citrus sinensis*, Cs), grape (*Vitis vinifera*, Vv), tomato (*Solanum lyconpersicum*, Sl), and arabidopsis (*Arabidopsis thaliana*, At) (Tomato Genome 2012, Zhang, Chen et al. 2012, Wu, Wang et al. 2013, Xu, Chen et al. 2013, Canaguier, Grimplet et al. 2017, Daccord, Celton et al. 2017, Li, Wei et al. 2017, Verde, Jenkins et al. 2017, Baek, Choi et al. 2018, Raymond, Gouzy et al. 2018, Sloan, Wu et al. 2018, VanBuren, Wai et al. 2018). The protein sequences of each plant genome were generated from their most recently annotated versions and were used as input sequences for OrthoFinder. Table S1 shows the annotated version and reference of the other 12 plant genomes except for our sweet cherry cv. Tieton genome. CAFÉ (v4.2) (De Bie et al., 2006) was used to analyze the expansion and contraction of their gene families. The species tree was generated using STRIDE (Emms & Kelly, 2017), as part of OrthoFinder and used as the input phylogenetic tree for CAFÉ.

2.9 Comparison between sweet cherry cv. Tieton genome and cv. Satonishiki genome

D-GENIES (v1.2.0) was used to compare the sweet cherry cv. Tieton genome with the cv. Satonishiki genome (Cabanettes & Klopp, 2018; Shirasawa et al., 2017). The whole sequence synteny analysis of the two assemblies were compared in both scaffold level and pseudochromosome level.

To compare the gene content between the two genome assemblies, we used three annotation versions that are the sweet cherry cv. Tieton genome annotation, the cv. Satonishiki genome annotation (Shirasawa et al., 2017), and an improved and re-annotated assembly of cv. Satonishiki genome released by NCBI Eukaryotic Genome Annotation Pipeline (NCBI Prunus avium Annotation Release 100, https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Prunus_avium/100/). OrthoFinder was used to compare the gene content among the three annotations (Emms & Kelly, 2015).

3. Results and Discussion

3.1 Sequencing summary

For sweet cherry cv. Tieton, a total of 121.61 Gb of raw sequencing data was generated with more than 810 million Chromium-linked paired-end reads. Table 1 shows the statistics of the

sequencing for the linked-read library. The low quality reads were filtered out and 750,890,534 clean reads were used for *de novo* assembly. The average Q20 was 95.52% and GC content was 40.8%. A cDNA library was constructed and sequenced to improve the precision of the genome annotation. As shown in Table S2, over 78 million 150-nt length paired-end reads were generated and assembled.

3.2 Determination of genome size and heterozygosity

The genome size of sweet cherry cv. Tieton was estimated to be 341.38 Mb based on 37-nt k-mer, which is very close to the genome size of 338 Mb estimated by flow cytometry (Arumuganathan & Earle, 1991). The k-mer distribution generated by GenomeScope was shown in Figure S1. The sweet cherry cv. Satonishiki genome estimated by k-mer method was 352.9 Mb (Shirasawa et al., 2017), larger than cv. Tieton genome. The genome size difference is probably due to the variety difference, but also may be caused by different library construction and sequencing methods. Heterozygosity of sweet cherry cv. Tieton genome was estimated to be 0.45%, and the repeat content was estimated to be 48.5% as shown in Figure S1.

3.3 Genome assembly and quality-assessment

The Supernova assembler (version 2.0) was used in *de novo* assembly and different sizes (40x, 50x, 60x, 65x, 68x, 70x, and 75x) of the Chromium-linked reads were attempted (Weisenfeld et al., 2017). Table S3 listed these assembly results, illustrating that the assembly using 70x coverage reads has the best assembly quality, and was selected for following analyses. GapCloser filled gaps in the raw sequencing data (Luo et al., 2012), resulting in the draft genome assembly of sweet cherry cv. Tieton of 280.33 Mb with contig N50 and scaffold N50 sizes of 63.65 kb and 2.48 Mb, respectively. Our sweet cherry cv. Tieton genome assembly had tenfold better contiguity than the cv. Satonishiki genome assembly (Shirasawa et al., 2017). The whole assembly increased in size from 272.36 to 280.33 Mb, whereas scaffold N50 increased from 219 kb to 2.48 Mb (Table 2).

150 million reads were sampled and 99.02% of the sampled reads were aligned to the sweet cherry cv. Tieton genome sequence using BWA (Li & Durbin, 2009), shown in Table S4. CEGMA (Parra et al., 2007) and BUSCO (Simao et al., 2015) were used to evaluate the completeness of the sweet cherry cv. Tieton genome and results were summarized in Table S5. Out of 248 core eukaryotic genes, 231 and 13 were found to be complete and partial genes in the CEGMA assessment, respectively. BUSCO analysis showed that our assembly captured 1,403 (97.43%) of the 1,440 single-copy orthologous genes of the embryo plant, of which 1,381 (95.9%) were complete (1,345 single-copy and 36 duplicated-copy), showing that the sweet cherry cv. Tieton genome assembly is well covered the gene space of the sweet cherry genome.

3.4 Chromosome-scale pseudomolecule construction

A consensus map was constructed from previously reported sweet cherry genetic maps for the chromosome-scale pseudomolecule construction (Calle et al., 2018; Guajardo et al., 2015; Klagges et al., 2013; Peace et al., 2012; Shirasawa et al., 2017; Wang et al., 2015). GMAP (Wu

& Watanabe, 2005) and ALLMAPS (Tang et al., 2015) were used to organize scaffolds onto eight chromosome-scale pseudomolecules (Hulse-Kemp et al., 2018). 494 scaffolds representing more than 214 Mb sequences, were anchored to eight chromosome-scale pseudomolecules of the sweet cherry cv. Tieton genome using 7,838 genetic markers (36.6 markers per Mb). 202.6 of the 214 Mb anchored sequences were oriented, the anchor rate and synteny of the maps were shown in Table S6 and Figure 1. This formation resulted in a higher contiguity than the sweet cherry cv. Satonishiki genome, consisting of 905 scaffolds spanning 191.7 Mb (Shirasawa et al., 2017).

3.5 Annotation of repetitive sequences

The Repbase library and repetitive motifs were searched and 32.71% (over 91 Mb) of the sweet cherry cv. Tieton genome assembly was found to be repetitive. Different repetitive elements were annotated in sweet cherry cv. Tieton genome, and their distribution were shown in Table 3. Long-terminal-repeat retrotransposons (6.39%) were predominant among the repetitive elements. The annotated repeat sequence length of the sweet cherry cv. Tieton genome was 28.4Mb shorter than the sweet cherry cv. Satonishiki genome (Shirasawa et al., 2017), which may explain why the k-mer method estimated a smaller genome size for cv. Tieton than cv. Satonishiki (299.17 versus 352.9 Mb).

3.6 cDNA assembly and noncoding RNA (ncRNA) annotation

Trinity was used to assembly the high quality cDNA reads (Grabherr et al., 2011). A total of 33,401 transcripts with a total length of 42.6 Mb were generated. The length of the assembled transcripts ranged from 201 to 15,591 nt, with a mean length of 1,276 nt. These assembled contigs were considered to be unigenes, and the distribution of their lengths is shown in Table S7.

Noncoding RNA includes miRNA, rRNA, snoRNA, tRNA, and the tRNA pseudogene. A total of 109,277 ncRNAs were generated, with a total length of 7.35 Mb, representing 2.63% of the sweet cherry cv. Tieton genome. As summarized in Table 4, our annotation predicted fewer tRNAs and rRNAs, compared to the annotation in of sweet cherry cv. Satonishiki genome (Shirasawa et al., 2017).

3.7 Protein-coding gene prediction and functional annotation

In total, 30,439 genes coding for 30,975 proteins were predicted in the sweet cherry cv. Tieton genome assembly. A summary of the predicted results using different methods was shown in Table 5. The *de novo* methods predicted 47,866 gene models, but the average gene length was shorter than other methods. After correcting with the transcript evidence, more than 16,000 genes were filtered out.

The predicted 30,975 proteins were blasted against non-redundant protein sequences (NR, https://blast.ncbi.nlm.nih.gov), Uniprot (The UniProt 2017), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2014), and InterPro (Finn et al., 2017) by using BLASTP (v2.9.0)(Camacho et al., 2009). As shown in Table 6, 30,973 of 30,975 proteins (99.99%) were annotated in at least one database.

3.8 Gene family analysis compared with other plant species

OrthoFinder (Emms & Kelly, 2015) identified the potential orthologous genes between the sweet cherry cv. Tieton genome and the other 12 plant genomes. The results of gene orthologous analysis were shown in Table S8. Gene family clustering identified 23,129 common orthogroups consisting of 375,493 genes (81.1% of the total genes) in these genomes. 8,465 orthogroups were present in all species, and 246 were single-copy genes. In the sweet cherry cv. Tieton genome, 46 orthogroups (124 genes) were unique and 2,062 orphan genes were identified that could not be clustered with any genes in the thirteen genomes. A species tree was constructed using STRIDE (Emms & Kelly, 2017), as part of OrthoFinder. As shown in Figure 2, sweet cherry (*Prunus avium*) exhibits a closer relationship with cherry flowers (*Prunus yedoensis*) than peach (*Prunus persica*) and Chinese plum (*Prunus mume*). A comparison was conducted to evaluate the expansion or contraction of these gene families using CAFÉ (version 4.2) (De Bie et al., 2006), and the results were shown in Figure 2. 1,012 gene families expanded and 3,642 gene families contracted in the sweet cherry cv. Tieton genome compared to the other 12 plant genomes (Figure 2).

3.9 Comparison between sweet cherry cv. Tieton genome and cv. Satonishiki genome

According to Figure 3A, genomic analysis using D-GENIES showed a high scaffold-level synteny of the sweet cherry cv. Tieton genome compared to sweet cherry cv. Satonishiki genome. High chromosome-level synteny was also detected in the two sets of pseudomolecules, except at the end of chromosomes 1, 4, 5, and 6 (Figure 3B). Based on Figure 3A, the sweet cherry cv. Tieton genome assembly had a better contig contiguity, whereas the sweet cherry cv. Satonishiki genome was more fragmented.

The original annotation of sweet cherry cv. Satonishiki genome (Shirasawa et al., 2017) and the re-annotated version of cv. Satonishiki genome released by the NCBI Eukaryotic Genome Annotation Pipeline were used to compare the gene content with our annotation of sweet cherry cv. Tieton genome. OrthoFinder analysis showed that the originally annotated version of cv. Satonishiki had 48 species-specific orthogroups represented 349 genes from our cv. Tieton genome annotation and the NCBI annotation of cv. Satonishiki genome (Table 7). The original version of sweet cherry cv. Satonishiki assembly annotated 41% more genes than our cv. Tieton genome annotation, however, the re-annotated version of cv. Satonishiki genome annotated a similar number of genes with our cv. Tieton genome. The increased gene numbers in the original annotation of sweet cherry cv. Satonishiki genome can be attributed to the fragmentation of genes onto multiple individual contigs. The re-annotated version of sweet cherry cv. Satonishiki genome adopted RNA-seq to improve the quality of the gene annotation by connecting genes fragmented in the assembly process (Denton et al., 2014). This method was also used in our sweet cherry cv. Tieton genome annotation process.

4. Conclusion

We successfully assembled a high-quality reference genome of sweet cherry cv. Tieton using linked reads sequencing technology. The assembly will be a valuable resource for future breeding efforts, gene function characterization and cultivar identification in the sweet cherry, as well as for comparative genomic analysis with other *Prunus* species.

Data availability

Raw sequencing reads are available in GenBank under Bioproject ID PRJNA503752. Genome assembly, annotation and chromosome-scale pseudomolecule construction data are available on Figshare with DOI: 10.6084/m9.figshare.9810236.

Figures

Figure_1.

Title: Pseudomolecule construction of sweet cherry (*Prunus avium*) by assigning scaffolds to seven genetic maps.

Legend: Chr 1 to 8 represents constructed pseudomolecules by merging seven genetic maps. AC, VG, CK, CP, KS, JWF, and JWF1 denote the sweet cherry genetic maps reported in Calle et al. 2018 ; Guajardo et al. 2015 ; Klagges et al. 2013 ; Peace et al. 2012 ; Shirasawa et al. 2017 ; Wang et al. 2015, respectively.

Figure_2.

Title: Species tree and gene family expansion analysis of 13 plant species.

Figure_3.

Title: Synteny analysis between sweet cherry (*Prunus avium*) cv. Tieton genome and cv. Satonishiki genome. (A) Scaffold level synteny dot plot. (B) Chromosome-scale synteny dot plot.

Legend: Sequence identity is indicated by colors.

Figure_S1.

Title: Genome size estimation of sweet cherry (*Prunus avium*) cv. Tieton based on k-mer (37-mer) analysis.

References

Aranzana MJ, Decroocq V, Dirlewanger E, Eduardo I, Gao ZS, Gasic K, Iezzoni A, Jung S, Peace C, Prieto H, Tao R, Verde I, Abbott AG, and Arus P. 2019. Prunus genetics and applications after de novo genome sequencing: achievements and prospects. *Hortic Res* 6:58. 10.1038/s41438-019-0140-8

Armstrong EE, Taylor RW, Prost S, Blinston P, van der Meer E, Madzikanda H, Mufute O,

Mandisodza-Chikerema R, Stuelpnagel J, Sillero-Zubiri C, and Petrov D. 2018. Cost-effective assembly of the African wild dog (Lycaon pictus) genome using linked reads. *Gigascience*. 10.1093/gigascience/giy124

Arumuganathan K, and Earle ED. 1991. Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* 9:208-218. 10.1007/BF02672069 %U https://doi.org/10.1007/BF02672069

Baek, S., K. Choi, G. B. Kim, H. J. Yu, A. Cho, H. Jang, C. Kim, H. J. Kim, K. S. Chang, J. H. Kim and J. H. Mun (2018). "Draft genome sequence of wild Prunus yedoensis reveals massive inter-specific hybridization between sympatric flowering cherries." Genome Biol 19(1): 127.

Birney E, Clamp M, and Durbin R. 2004. GeneWise and Genomewise. *Genome Research* 14:988-995. 10.1101/gr.1865504

Cabanettes F, and Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 6:e4958. 10.7717/peerj.4958

Calle A, Cai L, Iezzoni A, and Wünsch A. 2018. High-density linkage maps constructed in sweet cherry (Prunus avium L.) using cross- and self-pollination populations reveal chromosomal homozygosity in inbred families and non-syntenic regions with the peach genome. *Tree Genetics & Genomes* 14. 10.1007/s11295-018-1252-2

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. 10.1186/1471-2105-10-421

Canaguier, A., J. Grimplet, G. Di Gaspero, S. Scalabrin, E. Duchene, N. Choisne, N. Mohellibi, C. Guichard, S. Rombauts, I. Le Clainche, A. Berard, A. Chauveau, R. Bounon, C. Rustenholz, M. Morgante, M. C. Le Paslier, D. Brunel and A. F. Adam-Blondon (2017). "A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3)." Genom Data 14: 56-62.

Daccord, N., J. M. Celton, G. Linsmith, C. Becker, N. Choisne, E. Schijlen, H. van de Geest, L. Bianco, D. Micheletti, R. Velasco, E. A. Di Pierro, J. Gouzy, D. J. G. Rees, P. Guerif, H. Muranty, C. E. Durel, F. Laurens, Y. Lespinasse, S. Gaillard, S. Aubourg, H. Quesneville, D. Weigel, E. van de Weg, M. Troggio and E. Bucher (2017). "High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development." Nat Genet 49(7): 1099-1106.

De Bie T, Cristianini N, Demuth JP, and Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22:1269-1271. 10.1093/bioinformatics/btl097

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, and Hahn MW. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Computational Biology* 10:e1003998. 10.1371/journal.pcbi.1003998

Emms DM, and Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157. 10.1186/s13059-015-0721-2

Emms DM, and Kelly S. 2017. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution* 34:3267-3278. 10.1093/molbev/msx259

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC, Wu CH, Xenarios I, Yeh LS, Young SY, and Mitchell AL. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research* 45:D190-D199. 10.1093/nar/gkw1107

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, and Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29:644. 10.1038/nbt.1883

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, and Bateman A. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* 33:D121-124. 10.1093/nar/gki081

Guajardo V, Solis S, Sagredo B, Gainza F, Munoz C, Gasic K, and Hinrichsen P. 2015. Construction of High Density Sweet Cherry (Prunus avium L.) Linkage Maps Using Microsatellite Markers and SNPs Detected by Genotyping-by-Sequencing (GBS). *PloS One* 10:e0127750. 10.1371/journal.pone.0127750

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, and White O. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* 31:5654-5666. 10.1093/nar/gkg770

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, and Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* 9:R7. 10.1186/gb-2008-9-1-r7 %U https://doi.org/10.1186/gb-2008-9-1-r7

Hulse-Kemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, Williams SR, Weisenfeld N, Ramakrishnan S, Kumar V, Shah P, Schatz MC, Church DM, and Van Deynze A. 2018. Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. *Hortic Res* 5:4. 10.1038/s41438-017-0011-0

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, and Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* 42:D199-205. 10.1093/nar/gkt1076

Keller O, Kollmar M, Stanke M, and Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 27:757-763. 10.1093/bioinformatics/btr010

Klagges C, Campoy JA, Quero-Garcia J, Guzman A, Mansur L, Gratacos E, Silva H, Rosyara UR, Iezzoni A, Meisel LA, and Dirlewanger E. 2013. Construction and comparative analyses of highly dense linkage maps of two sweet cherry intra-specific progenies of commercial

cultivars. *PloS One* 8:e54743. 10.1371/journal.pone.0054743
Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5:59. 10.1186/1471-2105-5-59
Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, and Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* 35:3100-3108. 10.1093/nar/gkm160
Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760. 10.1093/bioinformatics/btp324
Li, Y., W. Wei, J. Feng, H. Luo, M. Pi, Z. Liu and C. Kang (2017). "Genome re-annotation of the wild strawberry Fragaria vesca using extensive Illumina- and SMRT-based RNA-seq datasets." DNA Res.
Liu Q, Chang S, Hartman GL, and Domier LL. 2018. Assembly and annotation of a draft genome sequence for Glycine latifolia, a perennial wild relative of soybean. *Plant Journal* 95:71-85. 10.1111/tpj.13931
Lowe TM, and Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25:955-964.
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, and Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18. 10.1186/2047-217X-1-18
Marcais G, and Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764-770. 10.1093/bioinformatics/btr011
Nawrocki EP, Kolbe DL, and Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25:1335-1337. 10.1093/bioinformatics/btp157
Ott A, Schnable JC, Yeh CT, Wu L, Liu C, Hu HC, Dalgard CL, Sarkar S, and Schnable PS. 2018. Linked read technology for assembling large complex and polyploid genomes. *BMC Genomics* 19:651. 10.1186/s12864-018-5040-z
Parra G, Bradnam K, and Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061-1067. 10.1093/bioinformatics/btm071
Peace C, Bassil N, Main D, Ficklin S, Rosyara UR, Stegmeir T, Sebolt A, Gilmore B, Lawley C, Mockler TC, Bryant DW, Wilhelm L, and Iezzoni A. 2012. Development and evaluation of a genome-wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. *PloS One* 7:e48305. 10.1371/journal.pone.0048305
Pollard MO, Gurdasani D, Mentzer AJ, Porter T, and Sandhu MS. 2018. Long reads: their purpose and place. *Human Molecular Genetics* 27:R234-R241. 10.1093/hmg/ddy177
Quero-García J, Iezzoni A, Puławska J, and Lang G. 2017. *Cherries: Botany, Production and Uses*. Boston, MA: CABI PUBLISHING.
Raymond, O., J. Gouzy, J. Just, H. Badouin, M. Verdenaud, A. Lemainque, P. Vergne, S. Moja, N. Choisne, C. Pont, S. Carrere, J. C. Caissard, A. Couloux, L. Cottret, J. M. Aury, J. Szecsi, D. Latrasse, M. A. Madoui, L. Francois, X. Fu, S. H. Yang, A. Dubois, F. Piola, A. Larrieu, M. Perez, K. Labadie, L. Perrier, B. Govetto, Y. Labrousse, P. Villand, C. Bardoux, V. Boltz, C.

Lopez-Roques, P. Heitzler, T. Vernoux, M. Vandenbussche, H. Quesneville, A. Boualem, A. Bendahmane, C. Liu, M. Le Bris, J. Salse, S. Baudino, M. Benhamed, P. Wincker and M. Bendahmane (2018). "The Rosa genome provides new insights into the domestication of modern roses." Nat Genet 50(6): 772-777.

Ru S, Main D, Evans K, and Peace C. 2015. Current applications, challenges, and perspectives of marker-assisted seedling selection in Rosaceae tree fruit breeding. *Tree Genetics & Genomes* 11:8. 10.1007/s11295-015-0834-5

Shirasawa K, Isuzugawa K, Ikenaga M, Saito Y, Yamamoto T, Hirakawa H, and Isobe S. 2017. The genome sequence of sweet cherry (Prunus avium) for use in genomics-assisted breeding. *DNA Research* 24:499-508. 10.1093/dnares/dsx020

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, and Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210-3212. 10.1093/bioinformatics/btv351

Slater GS, and Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31. 10.1186/1471-2105-6-31

Smit AFA, Hubley R, and Green P. 2016. RepeatMasker Open-4.0.6

Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, and Lu J. 2015. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology* 16:3. 10.1186/s13059-014-0573-1

Sloan, D. B., Z. Wu and J. Sharbrough (2018). "Correction of Persistent Errors in Arabidopsis Reference Mitochondrial Genomes." Plant Cell 30(3): 525-527.

Tomato Genome, C. (2012). "The tomato genome sequence provides insights into fleshy fruit evolution." Nature 485(7400): 635-641.

VanBuren, R., C. M. Wai, M. Colle, J. Wang, S. Sullivan, J. M. Bushakra, I. Liachko, K. J. Vining, M. Dossett, C. E. Finn, R. Jibran, D. Chagne, K. Childs, P. P. Edger, T. C. Mockler and N. V. Bassil (2018). "A near complete, chromosome-scale assembly of the black raspberry (Rubus occidentalis) genome." Gigascience 7(8).

Verde, I., J. Jenkins, L. Dondini, S. Micali, G. Pagliarani, E. Vendramin, R. Paris, V. Aramini, L. Gazza, L. Rossini, D. Bassi, M. Troggio, S. Shu, J. Grimwood, S. Tartarini, M. T. Dettori and J. Schmutz (2017). "The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity." BMC Genomics 18(1): 225.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, and Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33:2202-2204. 10.1093/bioinformatics/btx153

Wang J, Zhang K, Zhang X, Yan G, Zhou Y, Feng L, Ni Y, and Duan X. 2015. Construction of Commercial Sweet Cherry Linkage Maps and QTL Analysis for Trunk Diameter. *PloS One* 10:e0141261. 10.1371/journal.pone.0141261

Wei H, Chen X, Zong X, Shu H, Gao D, and Liu Q. 2015. Comparative transcriptome analysis of genes involved in anthocyanin biosynthesis in the red and yellow fruits of sweet cherry (Prunus avium L.). *PloS One* 10:e0121164. 10.1371/journal.pone.0121164

532 Weisenfeld NI, Kumar V, Shah P, Church DM, and Jaffe DB. 2017. Direct determination of diploid
533 genome sequences. *Genome Research* 27:757-767. 10.1101/gr.214874.116
534 Wu, J., Z. Wang, Z. Shi, S. Zhang, R. Ming, S. Zhu, M. A. Khan, S. Tao, S. S. Korban, H. Wang, N. J.
535 Chen, T. Nishio, X. Xu, L. Cong, K. Qi, X. Huang, Y. Wang, X. Zhao, J. Wu, C. Deng, C. Gou,
536 W. Zhou, H. Yin, G. Qin, Y. Sha, Y. Tao, H. Chen, Y. Yang, Y. Song, D. Zhan, J. Wang, L. Li, M.
537 Dai, C. Gu, Y. Wang, D. Shi, X. Wang, H. Zhang, L. Zeng, D. Zheng, C. Wang, M. Chen, G.
538 Wang, L. Xie, V. Sovero, S. Sha, W. Huang, S. Zhang, M. Zhang, J. Sun, L. Xu, Y. Li, X. Liu, Q.
539 Li, J. Shen, J. Wang, R. E. Paull, J. L. Bennetzen, J. Wang and S. Zhang (2013). "The genome
540 of the pear (Pyrus bretschneideri Rehd.)." Genome Res 23(2): 396-408.
541 Wu TD, and Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA
542 and EST sequences. *Bioinformatics* 21:1859-1875. 10.1093/bioinformatics/bti310
543 Xu, Q., L. L. Chen, X. Ruan, D. Chen, A. Zhu, C. Chen, D. Bertrand, W. B. Jiao, B. H. Hao, M. P. Lyon,
544 J. Chen, S. Gao, F. Xing, H. Lan, J. W. Chang, X. Ge, Y. Lei, Q. Hu, Y. Miao, L. Wang, S. Xiao,
545 M. K. Biswas, W. Zeng, F. Guo, H. Cao, X. Yang, X. W. Xu, Y. J. Cheng, J. Xu, J. H. Liu, O. J.
546 Luo, Z. Tang, W. W. Guo, H. Kuang, H. Y. Zhang, M. L. Roose, N. Nagarajan, X. X. Deng and
547 Y. Ruan (2013). "The draft genome of sweet orange (Citrus sinensis)." Nat Genet 45(1):
548 59-66.
549 Yan M, Zhang X, Zhao X, and Yuan Z. 2019. The complete mitochondrial genome sequence of
550 sweet cherry (Prunus avium cv. 'summit'). *Mitochondrial DNA Part B* 4:1996-1997.
551 10.1080/23802359.2019.1617082
552 Zhang, Q., W. Chen, L. Sun, F. Zhao, B. Huang, W. Yang, Y. Tao, J. Wang, Z. Yuan, G. Fan, Z. Xing,
553 C. Han, H. Pan, X. Zhong, W. Shi, X. Liang, D. Du, F. Sun, Z. Xu, R. Hao, T. Lv, Y. Lv, Z. Zheng,
554 M. Sun, L. Luo, M. Cai, Y. Gao, J. Wang, Y. Yin, X. Xu, T. Cheng and J. Wang (2012). "The
555 genome of Prunus mume." Nat Commun 3: 1318.
556 Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-
557 Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, Mudivarti PA, Wyatt PW,
558 Bharadwaj R, Makarewicz AJ, Li Y, Belgrader P, Price AD, Lowe AJ, Marks P, Vurens GM,
559 Hardenbol P, Montesclaros L, Luo M, Greenfield L, Wong A, Birch DE, Short SW, Bjornson
560 KP, Patel P, Hopmans ES, Wood C, Kaur S, Lockwood GK, Stafford D, Delaney JP, Wu I,
561 Ordonez HS, Grimes SM, Greer S, Lee JY, Belhocine K, Giorda KM, Heaton WH, McDermott
562 GP, Bent ZW, Meschi F, Kondov NO, Wilson R, Bernate JA, Gauby S, Kindwall A, Bermejo
563 C, Fehr AN, Chan A, Saxonov S, Ness KD, Hindson BJ, and Ji HP. 2016. Haplotyping germline
564 and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*
565 34:303-311. 10.1038/nbt.3432
566

Table 1(on next page)

Raw data and valid data statistics of sequencing for linked-read libraries of sweet cherry (*Prunus avium*) cv. Tieton

Table 1. Raw data and valid data statistics of sequencing for linked-read libraries of sweet cherry (*Prunus avium*) cv. Tieton

Parameter	Value	Parameter	Value
Raw bases (Gb)	121.61	**Clean bases (Gb)**	112.63
Q20 (%)	97.52	**Clean reads**	750,890,534
Q30 (%)	94.24	**Clean ratio (%)**	92.62
GC content (%)	40.8	**Low ratio (%)**	5.51
N ratio (%)	0.01	**Adapter ratio (%)**	1.86

Table 2(on next page)

Comparison of sweet cherry (*Prunus avium*) genome assemblies of cv. Tieton and cv. Satonishiki

Table 2. Comparison of sweet cherry (*Prunus avium*) genome assemblies of cv. Tieton and cv. Satonishiki

Assembly parameters	cv. Tieton	cv. Satonishiki
Assembled genome size (Mb)	280.33	272.36
Scaffold N50 (Mb)	2.48	0.22
Number of scaffold	14,344	10,148
Longest of scaffold (Mb)	17.96	1.46
Contig N50 (kb)	63.65	28.779
Number of contig	19,420	32,301
Longest of contig (kb)	670.29	19.97
Total contig length (Mb)	237.92	246.8
GC content (%)	37.86	37.7
Ns (%)	15.12	9.34

Mb: Megabase, kb: Kilobase, GC: Guanine-cytosine, Ns: Ambiguous bases

Table 3(on next page)

Summary of detected repeat elements of sweet cherry (*Prunus avium*) cv. Tieton genome

Table 3. Summary of detected repeat elements of sweet cherry (*Prunus avium*) cv. Tieton genome

Repeat type	Number	Total length (bp)	Percent (%)
LTR	22,244	17,899,535	6.39
DNA elements	11,927	7,198,678	2.57
LINE	4,700	1,900,833	0.68
SINE	1	84	0
Simple repeat	6,266	4,736,127	1.69
Low complexity	141	23,252	0.01
Unknown	228,932	59,943,002	21.38
Total	274,211	91,701,511	32.71

LTR: Long terminal retrotransposon; SINE: Short interspersed nuclear elements; LINE: Long interspersed nuclear elements.

Table 4(on next page)

Summary of noncoding-RNAs prediction in sweet cherry (*Prunus avium*) cv. Tieton genome

Table 4. Summary of noncoding-RNAs prediction in sweet cherry (*Prunus avium*) cv. Tieton genome

Non-coding RNA type	Non-coding RNA number	Total length (bp)	Percentage (%)
miRNA	21,673	1,703,848	0.61
rRNA	35	51,780	0.02
snoRNA	86,993	5,560,365	1.98
tRNA	521	39,227	0.01
tRNA-pseudogene	48	3,585	0
Total	109,277	7,358,805	2.63

miRNA: micro-RNA; rRNA: ribosomal RNA; snoRNA: small nucleolar RNA; tRNA: transfer RNA

$$\text{Percentage (\%)} = \frac{\text{the total length of corresponding non} - \text{coding RNA type}}{\text{whole genome size of cv. Tieton}}$$

Table 5(on next page)

Statistics for protein-coding gene prediction of sweet cherry (*Prunus avium*) cv. Tieton genome

Table 5. Statistics for protein-coding gene prediction of sweet cherry (*Prunus avium*) cv. Tieton genome

Prediction method or software	Number of genes	mRNA number	Average RNA length	Exon number	Average exon length	Intron number	Average intron length
De novo	47,866	47,866	2118.8	179,067	302.9	131,201	359.5
RNA-seq	16,512	16,512	4032.3	91,646	228.5	75,134	344.6
EVM	30,455	30,455	2433.3	139,225	275.8	108,770	328.3
PASA	30,439	30,975	2720.6	140,185	277	109,210	329.2

EVM: EVidenceModeler; PASA: Program to Assemble Spliced Alignments.

Table 6(on next page)

Statistics of functional annotated genes of sweet cherry (*Prunus avium*) cv. Tieton genome

Table 6. Statistics of functional annotated genes of sweet cherry (*Prunus avium*) cv. Tieton genome

Functional database	Number of annotated genes	Percentage (%)
InterPro	30300	97.8
NR	30882	99.7
GO	16433	53.05
Uniprot	29444	95.05
KEGG	9202	29.7
Total	30973	99.99

NR: NCBI Non-redundant protein, GO: Gene ontology, KEGG: Kyoto Encyclopedia of Genes and Genomes

Table 7 (on next page)

Statistics of orthogroups analysis between sweet cherry (*Prunus avium*) cv. Tieton and cv. Satonishiki genome annotations.

Table 7. Statistics of orthogroups analysis between sweet cherry (*Prunus avium*) cv. Tieton and cv. Satonishiki genome annotations

Annotation summary	cv. Tieton	cv. Satonishiki	
		NCBI version	Original version
Number of genes	30975	35009	43673
Number of genes in orthogroups	26730	31314	25388
Number of unassigned genes	4245	3695	18285
Percentage of genes in orthogroups	86.3%	89.4%	58.1%
Percentage of unassigned genes	13.7%	10.6%	41.9%
Number of orthogroups containing species	21511	21258	20738
Percentage of orthogroups containing species	92.4%	91.3%	89%
Number of species-specific orthogroups	14	1	48
Number of genes in species-specific orthogroups	67	2	349

NCBI version is the improved assembly annotation of sweet cherry cv. Satonishiki released by National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Prunus_avium/100/).

Original version is the assembly annotation of sweet cherry cv. Satonishiki genome documented in (Shirasawa et al., 2017).

Figure 1

Title: Pseudomolecule construction of sweet cherry (*Prunus avium*) by assigning scaffolds to seven genetic maps.

Legend: Chr 1 to 8 represents constructed pseudomolecules by merging seven genetic maps. AC, VG, CK, CP, KS, JWF, and JWF1 denote the sweet cherry genetic maps reported in Calle et al. 2018 ; Guajardo et al. 2015 ; Klagges et al. 2013 ; Peace et al. 2012 ; Shirasawa et al. 2017 ; Wang et al. 2015, respectively.

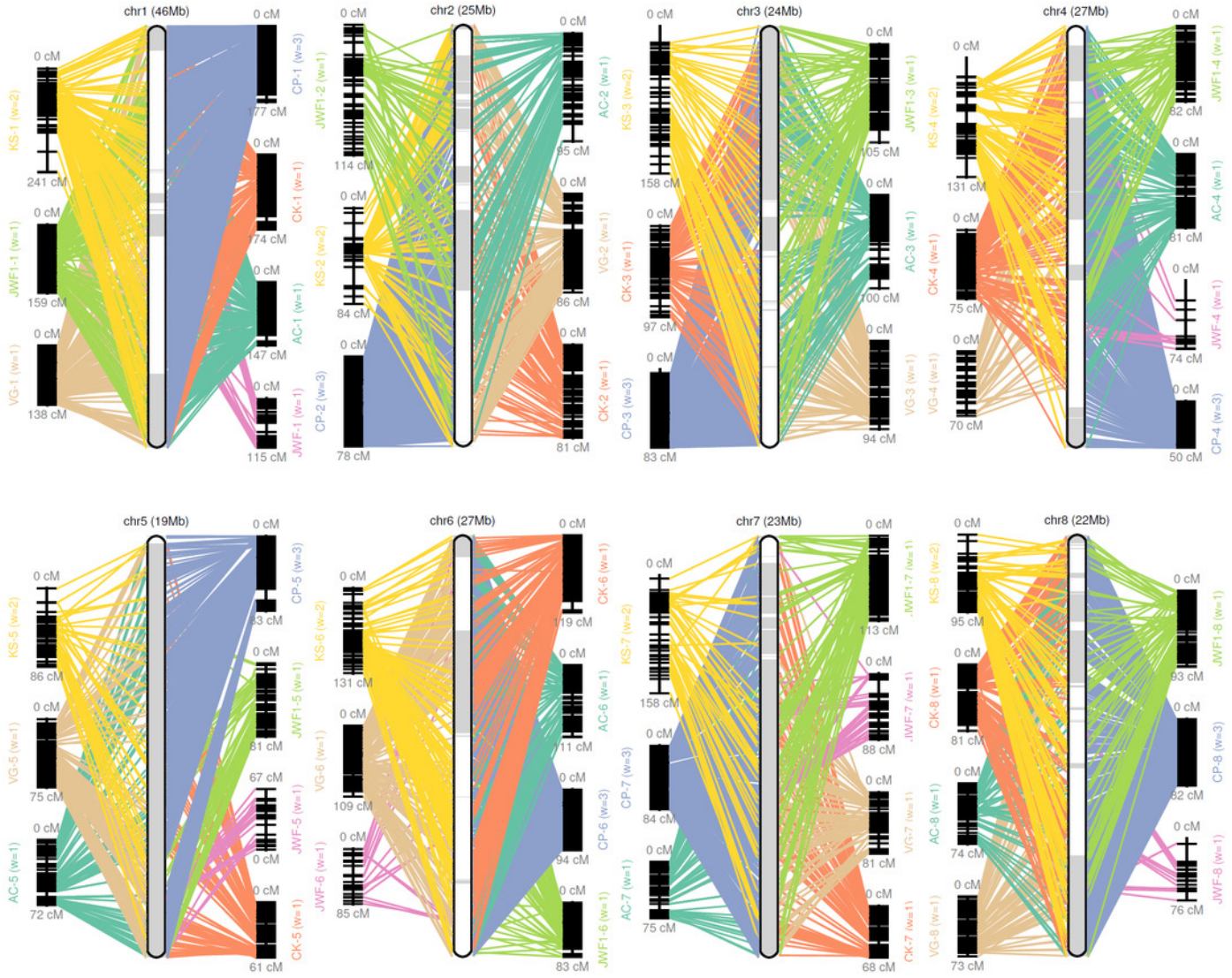
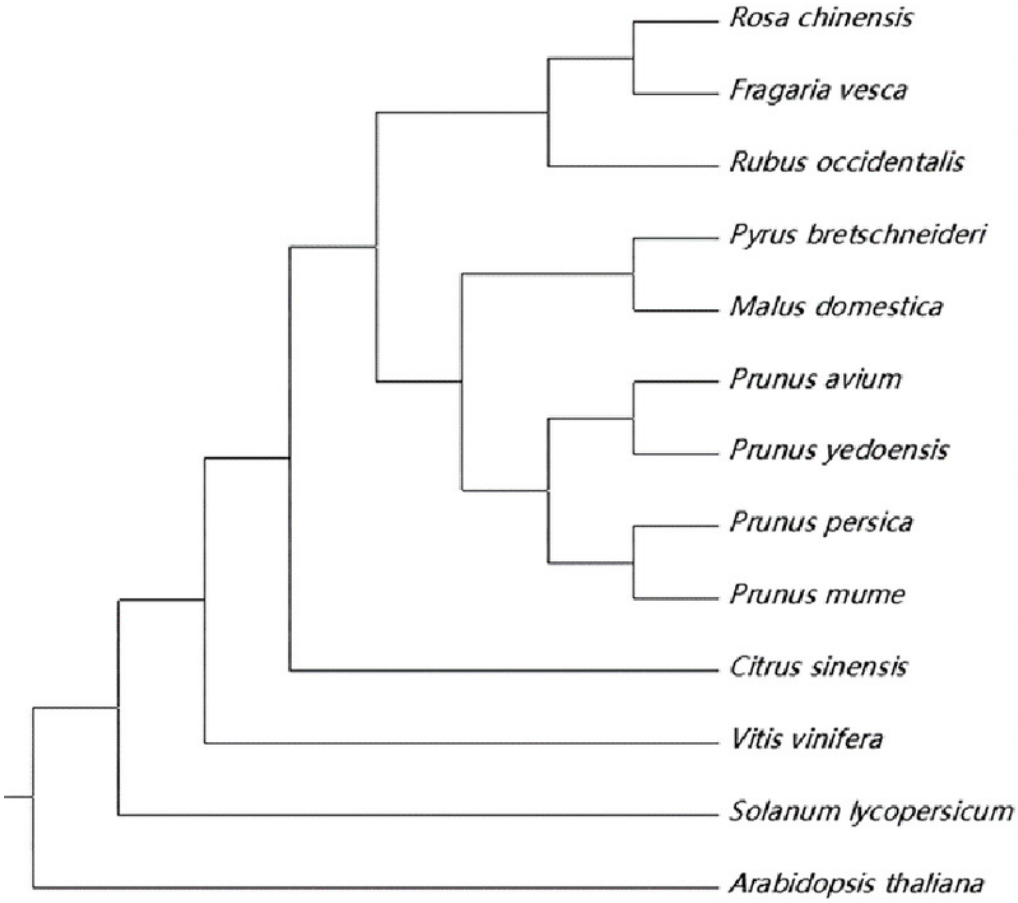


Figure 2

Title: Species tree and gene family expansion analysis of 13 plant species.

Legend: The species tree were constructed using STRIDE. Gene family expansions are indicated in red, and gene family contractions are indicated in green.



	Gene family	
	Expansions	Contractions
Rosa chinensis	2,606	1,112
Fragaria vesca	1,246	3,052
Rubus occidentalis	1,337	4,254
Pyrus bretschneideri	7,963	3,052
Malus domestica	6,028	1,052
Prunus avium	1,012	3,642
Prunus yedoensis	6,026	2,505
Prunus persica	687	2,051
Prunus mume	3,201	3,201
Citrus sinensis	1,344	10,218
Vitis vinifera	1,289	10,857
Solanum lycopersicum	2,835	9,505
Arabidopsis thaliana	2,437	10,805

Figure 3

Title: Synteny analysis between sweet cherry (*Prunus avium*) cv. Tieton genome and cv. Satonishiki genome. (A) Scaffold level synteny dot plot. (B) Chromosome-scale synteny dot plot.

Legend: Sequence identity is indicated by colors.