A new *de novo* assembly of sweet cherry (*Prunus avium*) improves genome coverage and completeness

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

Sweet cherry (*Prunus avium*) is one of the economically significant fruit species in the world. However, the available genomic resource for sweet cherry is limited, which has hindered sweet cherry molecular breeding. Here, we report a high-quality reference genome of the diploid sweet cherry (2n=2x=16) cv. 'Tieton' using the linked reads sequencing platform. Over 750 million clean reads representing 112.63 Gb of raw sequence data were generated. The Supernova genome assembler produced a highly ordered and more 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 79.63% of 352.90 Mb sweet cherry genome. Eight chromosome-scale pseudomolecules were constructed which covered 214 Mb sequence of the final scaffold assembly. A combination of de novo, homology-based, and RNA-seq methods predicted 30,975 protein-coding loci. 98.39% of core eukaryotic genes and 97.43% of single copy orthologues in embryo plants were captured as complete or partial, indicating the completeness of assembly. Our study reveals that the linked-read sequencing technology can be used to effectively construct high-quality reference genome of sweet cherry, which will benefit molecular breeding and cultivar identification in sweet cherry.

A new *de novo* assembly of sweet cherry (*Prunus avium*) improves genome coverage and completeness

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

Sweet cherry (*Prunus avium*) is one of the economically significant fruit species in the world. However, the available genomic resource for sweet cherry is limited, which has hindered sweet cherry molecular breeding. Here, we report a high-quality reference genome of the diploid sweet cherry (2n=2x=16) cv. 'Tieton' using the linked reads sequencing platform. Over 750 million clean reads representing 112.63 Gb of raw sequence data were generated. The Supernova genome assembler produced a highly ordered and more 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 79.63% of 352.90 Mb sweet cherry genome. Eight chromosome-scale pseudomolecules were constructed which covered 214 Mb sequence of the final scaffold assembly. A combination of de novo, homology-based, and RNA-seq methods predicted 30,975 protein-coding loci. 98.39% of core eukaryotic genes and 97.43% of single copy orthologues in embryo plants were captured as complete or partial, indicating the completeness of the assembly. Our study reveals that the linked-read sequencing technology can be used to effectively construct high-quality reference genome of sweet cherry, which will benefit molecular breeding and cultivar identification in sweet cherry.

KEYWORDS: sweet cherry; genome sequencing; genome assembly; 10x Genomics Chromium; linked reads

INTRODUCTION

Sweet cherry (*Prunus avium*), originated in Asia Minor near the Black Sea and the Caspian Sea, is one of the economically significant fruit species in the world. The sweet cherry production in China has experienced a dramatic increase over the last three decades. Meanwhile, dedicated breeding efforts have also been devoted. Marker-assisted breeding and genomic selection are currently major strategies to speed up the breeding cycle (Ru et al. 2015). However, they are limited by lacking of high-quality reference genome. Even though sweet cherry has a simple and compact genome (2n=2x=16), only one draft genome assembly has been reported (Shirasawa et al. 2017). Using the short-read sequencing technology, the previous assembly had smaller scaffolds with a N50 of 219.6 Kb and lower genome coverage of 77.8%. Recent advances in the linked reads sequencing pipeline developed by the 10x Genomics has been proved to assembling cost-effective and high-quality genome, because it utilizes barcoded sequencing library to generate long-range information (preferably >100 kb) and standard short-read sequencing to ensure massive throughput, high accuracy, and low cost (Pollard et al. 2018). It was primarily designed for human genome assembly, but has been proven in many other animal and plant species, such as wild dog, proso millet and pepper (Armstrong et al. 2018; Hulse-Kemp et al. 2018; Ott et al. 2018).

In current study, we confirmed that linked reads technology can effectively *de novo* assemble the genome of sweet cheery cv. 'Tieton', the most popular variety in China. The high-quality genome assembly as well as gene annotation and chromosome-scale pseudomolecules construction in this study provide a valuable resource for genetic marker development and gene mapping to speed up sweet cherry breeding. Additionally, our assembly platform will extend support for future *de novo* genome assemblies using linked reads in relative *Prunus* species.

MATERIALS AND METHODS

Sample and DNA extraction

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

Chromium library construction and sequencing

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

Genome size and heterozygosity estimation and de novo assembly

The sweet cherry genome size and heterozygosity were estimated based on k-mer frequency of the sequence data using the k-mer counting program Jellyfish (v.2.0.8) (Marcais & Kingsford 2011) and GenomeScope (Vurture et al. 2017). The sweet cherry genome was assembled using the Supernova assembler (v2.0, https://www.10x genomics.com/) with 40x, 50x, 60x, 65x, 68x, 70x, and 75x coverage of the estimated genome size.

Assessment of genome assembly

To evaluate the quality of our sweet cherry assembly, 150 million reads were sampled and aligned to the assembled genome sequence using Burrows-Wheller Alignment tool (BWA) (Li & Durbin 2009). The completeness of the Supernova assembly were assessed by Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra et al. 2007) and Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao et al. 2015).

Chromosome-scale pseudomolecule construction

Seven previously published sweet cherry genetic maps were used for chromosome-scale pseudomolecule construction. Five maps were built by Shirasawa et al. (Shirasawa et al. 2017), Peace et al. (Peace et al. 2012), Klagges et al. (Klagges et al. 2013), Calle et al. (Calle et al. 2018), and Guajardo et al. (Guajardo et al. 2015), respectively. We used the initials of the first author to name these maps. Hereafter, each map will be referred to as KS, CP, CK, AC, and VG map. The other two maps, which were named as JWF (the framework map of 'WxL' map) and JWF1 (the second round map of 'WxL' map), were both reported in Wang et al (Wang et al. 2015). Genetic makers and/or flanking sequences for these maps were aligned to the current scaffolds by using GMAP (Wu & Watanabe 2005) as described by Hulse-Kemp (Hulse-Kemp et al. 2018). Markers were filtered out if they were aligned to more than one scaffold or aligned to the same scaffold but assigned in different linkage groups. Then, alignment results of GMAP were fitted into ALLMAPS (Tang et al. 2015) for pseudomolecule construction. Equal and unequal weights parameters for seven linkage maps were attempted. The optimal weight settings that generated largest number of anchored and oriented scaffolds were as follows: KS =2, CP =3, CK =1, AC =1, VG =1, JWF =1, and JWF1=1.

Repeat annotation

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

cDNA library preparation, sequencing and de novo assembly

Total RNA was extracted from young leaves of the same plant for genome sequencing. cDNA library was constructed as described by Wei et al. (Wei et al. 2015) and sequenced by CapitalBio Technology Inc. (Beijing, China) using the Illumina HiSeq 2000 platform. More than 78 million paired-end reads were generated with the length of 150 nt. After trimming the adapters and removing the low-quality reads, 77,258,972 clean reads was obtained. These high quality reads were assembled by Trinity (Grabherr et al. 2011).

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

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

We combined *de novo*, homology-based, and RNA-seq methods to predict protein-coding genes in the sweet cherry genome. For the *de novo* annotation, Augustus (Keller et al. 2011) and SNAP (Korf 2004) were used to perform protein-coding gene prediction on repeat-masked genome sequences. The predicted genes were annotated by Genewise (Birney et al. 2004) and Exonerate (Slater & Birney 2005). Simultaneously, Program to Assemble Spliced Alignments (PASA) pipeline (Haas et al. 2003) was used in transcriptome-assistant method with the unigenes assembled by the RNA-seq data. EVidenceModeler (Haas et al. 2008) and PASA were used to combine three predicted results.

Gene family analysis

OrthoFinder (version 2.2.7) was used to identify orthologous genes among thirteen plant genomes (Emms & Kelly 2015), which are sweet cherry (*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). Protein sequences of each plant genome were generated from their latest annotation versions and used as the input sequences (Table S1). CAFÉ (version 4.2) was used to analyze the expansion and contraction of gene families (De Bie et al. 2006). The species tree generated by STRIDE (Emms & Kelly 2017), as part of OrthoFinder, was used as the input phylogenetic tree for CAFÉ.

RESULTS AND DISCUSSION

Sequencing summary

A total of 121.61 Gb raw sequencing data was generated, consisting of more than 810 million Chromium-linked paired-end reads. After filtering the low quality reads, clean reads

were used for *de novo* assembly (Table 1). To improve the precision of the genome annotation, one cDNA library was constructed and sequenced. More than 78 million 150-nt length paired-end reads were generated and assembled.

Determination of genome size and heterozygosity

The genome size was estimated to be 299.17 Mb based on 17 nt k-mer (Figure S1), which is smaller than the genome size of 338 Mb estimated by using the flow cytometry (Arumuganathan & Earle 1991) and 352.9 Mb estimated using the k-mer method in previously assembly cv 'Satonishiki' (Shirasawa et al. 2017). The underestimation of the sweet cherry genome size may be caused by the missing of ~38 Mb genome sequence during the Chromium library construction or failure assembly of ~38 Mb repeat sequences. According to GenomeScope, the heterozygosity was estimated to be 0.49%, and repeat content was estimated to be 57.50% (Figure S1).

Genome assembly and quality assessment

The assembly using 70x coverage (158.01 million reads) provided the best quality over the others (Table S2). After filling gaps using all the raw sequencing data by GapCloser (Luo et al. 2012), the draft genome assembly was 280.33 Mb, with the contig N50 and scaffold N50 sizes of 63.65 kb and 2.48 Mb, respectively (Table 2). Compared to the former assembly of sweet cherry genome by Shirasawa *et al. (Shirasawa et al. 2017)*, our assembly provided a slightly higher coverage and much better contiguity (Table 2). The scaffold assembly increased in size from 272.36 to 280.33 Mb, whereas the N50 from 0.22 to 2.48 Mb.

To evaluate the quality of our sweet cherry assembly, 150 million reads were sampled and aligned to the assembled genome sequence using BWA (Li & Durbin 2009). 99.02% of the reads were reliably aligned to our genome assembly (Table S3). CEGMA (Parra et al. 2007) and BUSCO (Simao et al. 2015) were utilized to evaluate the completeness of the Supernova assembly (Table S4). Out of 248 core eukaryotic genes, 231 and 13 were detected as complete and partial genes in the CEGMA assessment. The BUSCO analysis showed that our assembly captured 1,403 (97.43%) of the 1,440 single-copy orthologous of embryo plants, of which 1,381 (95.9%) were complete (1,345 single-copy and 36 duplicated-copy), implying a high completeness of our assembly.

Chromosome-scale pseudomolecule construction

Using previously reported sweet cherry genetic maps (Calle et al. 2018; Guajardo et al. 2015; Klagges et al. 2013; Peace et al. 2012; Shirasawa et al. 2017; Wang et al. 2015), we constructed a consensus map to guide the chromosome-scale pseudomolecule construction. GMAP (Wu & Watanabe 2005) and ALLMAPS (Tang et al. 2015) were used to organize scaffolds onto eight chromosome-scale pseudomolecules as described by Hulse-Kemp (Hulse-Kemp et al. 2018). Eventually, 494 scaffolds representing over 214 Mb sequences were anchored to eight chromosome-scale pseudomolecules by using 7,838 markers (36.6 markers per Mb). Among the 214 Mb anchored sequences, 202.6 Mb were oriented (Table S5 and Figure 1). These results illustrated a higher contiguity and quality than the previous reference genomes of sweet cherry that had 905 scaffolds spanning 191.7 Mb (Shirasawa et al. 2017).

Annotation of repeat sequences

By searching against the Repbase library and repetitive motif identification, we found that 32.71% (over 91 Mb) of the genome assembly was repetitive (Table 3). Among all the repetitive elements, long-terminal-repeat retrotransposons (6.39%) were the predominant component. We found a nearly 28.4Mb of annotated repeat sequence length shorter in our assembly than the former assembly (Shirasawa et al. 2017), which might be a reasonable explanation that the k-mer method estimated a smaller genome size for our assembly than the former assembly (299.17 over 352.9 Mb).

cDNA assembly and noncoding RNA (ncRNA) annotation

More than 78 million paired-end cDNA reads were generated with the length of 150 nt (Table S6). The high quality cDNA reads were assembled by Trinity. A total of 33,401 transcripts with a total length of 42.6 Mb were generated. The length of assembled transcripts ranged from 201 to 15,591 nt, with a mean length of 1,276 nt. These assembled contigs were considered as unigenes, and the length distribution is shown in Table S7.

Noncoding RNA includes miRNA, rRNA, snoRNA, tRNA, and tRNA pseudogene, with different structures. A total of 109,277 ncRNAs were generated, with a total length of 7.35 Mb, representing 2.63% of the whole sweet cherry genome (Table 4). Compared to the annotation in Shirasawa et al. , our annotation predicted fewer tRNA and rRNA.

Protein-coding gene prediction and functional annotation

In total, 30,439 genes coding for 30,975 proteins were predicted in our assembly (Table 5), which is fewer than the previous assembly version (Shirasawa et al. 2017) with 43,349 genes. Our newly *de novo* annotated gene models were fewer than the previous assembly, 30,439 genes vs 43,349 genes. This reduction may be due to the overestimation of tandem duplicated genes in the fragmentation of the previous genome assembly, or due to the different prediction method. Similar decreases were also observed in apple (Daccord et al. 2017) and Brassica rapa (Zhang et al. 2018).

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

Gene family analysis compared with other plant species

We applied OrthoFinder to identify the potential orthologous genes between thirteen plant genomes (Emms & Kelly 2015). Gene family clustering identified 23,129 orthogroups in common that was consisted of 375,493 genes (81.1% of the total genes) in these genomes (Table S8). 8,465 orthogroups were present in all species, and 246 of 8,465 orthogroups were single-copy genes. In sweet cherry genome, a total of 46 orthogroups (124 genes) were unique and 2,062 orphan genes were identified that could not be clustered with any genes in these thirteen genomes. A species tree were also constructed by using STRIDE (Emms & Kelly 2017), as part of OrthoFinder (Figure 2). To study the expansion or contraction of these gene families, a comparison was conducted using CAFÉ (version 4.2)(De Bie et al. 2006). Compared with other plant genomes, 1,017 gene families had expanded and 3,643 gene families had contracted in sweet cherry genome (Figure 2).

CONCLUSION

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

Conflicts of interests: The authors declare that there is no conflict of interest.

Funding statement: This study was supported by Shandong Provincial Key Laboratory for Fruit Biotechnology Breeding and also funded by the Special Fund for Innovation Teams of Fruit Trees in Agricultural Technology System of Shandong Province (SDAIT-06-04).

Availability of supporting data: Raw sequencing reads have been deposited in GenBank under Bioproject ID PRJNA503752 (Reviewer link created for BioProject PRJNA503752 https://dataview.ncbi.nlm.nih.gov/object/PRJNA503752?reviewer=uoh9d3ruu1fv7vd34he6hqrcag, and reviewer link for genome assembly, annotation and chromosome-scale pseudomolecule construction https://figshare.com/s/1eb14a4d516656d789e3).

Author Contributions: JW, WL and QL conceived the project. JW collected the samples and extracted the genomic DNA. JW, WL, DZ, PH, YT and HW performed the genome assembly and data analysis. JW, WL, HZ, XZ, LX, LZ, XC and QL wrote the paper. All authors read and approved the final version of the manuscript.

References

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

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

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

Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, van de Geest H, Bianco L, Micheletti D, Velasco R, Di Pierro EA, Gouzy J, Rees DJG, Guerif P, Muranty H, Durel CE, Laurens F, Lespinasse Y, Gaillard S, Aubourg S, Quesneville H, Weigel D, van de Weg E, Troggio M, and Bucher E. 2017. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nature Genetics 49:1099-1106. 10.1038/ng.3886

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

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

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

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

The UniProt C. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45:D158-D169. 10.1093/nar/gkw1099

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

Wu TD, and Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859-1875. 10.1093/bioinformatics/bti310

Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, Liang J, Cai C, Liu Z, Liu B, Wang F, Li S, Liu F, Li X, Cheng L, Yang W, Li MH, Grossniklaus U, Zheng H, and Wang X. 2018. Improved Brassica rapa reference genome by single-molecule sequencing and chromosome conformation capture technologies. Hortic Res 5:50. 10.1038/s41438-018-0071-9

Table 1(on next page)

Summary statistics of sequence data.

Table 1. Summary statistics of sequence data.

Sample	Sweet cherry
Raw Reads	810,734,866
Raw Base (G)	121.61
Clean Reads	750,890,534
Clean Based (G)	112.63
Error Rate (%)	0.02
Q20 (%)	97.52
Q30 (%)	94.24
GC Content (%)	40.8
Clean Ratio (%)	92.62
Low Ratio (%)	5.51
N Ratio (%)	0.01
Adapter Ratio (%)	1.86

Table 2(on next page)

Comparison of sweet cherry (*Prunus avium*) genome assembly between cv. 'Tieton' in the current study and cv. 'Satonishiki' in previous study.

Table 2. Comparison of sweet cherry (*Prunus avium*) genome assembly between cv. 'Tieton' in the current study and cv. 'Satonishiki' in previous study (Shirasawa et al. 2017).

Genome	Tieton	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	0.286
Number of contig	19,420	2,046,201
Longest of contig (kb)	670.29	19.97
Total contig length (Mb)	237.92	407.82
G+C content (%)	37.86	37.7
Ns (%)	15.12	9.34

Table 3(on next page)

Distribution of repeats and unique sequences

Table 3. Distribution of repeats and unique sequences

Type		Number	Total length (bp)	Percent (%)
Unique sequence		-	188,625,714	67.29
Repeat type	**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

Table 4(on next page)

Summary of the none-coding RNA analysis

Table 4. Summary of the none-coding RNA analysis

Gene type	Gene number	Total length (bp)	Percent (%)
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
All nc-RNA	109,277	7,358,805	2.63

Table 5(on next page)

Summary statistics for protein-coding gene prediction

Table 5. Summary statistics for protein-coding gene prediction

Prediction method or software*	Number of genes	mRNA number	Average RNA length	Exon number	Average exon length	Intron number	Average intron length
De novo	47866	47866	2118.8	179067	302.9	131201	359.5
RNA-seq	16512	16512	4032.3	91646	228.5	75134	344.6
EVM	30455	30455	2433.3	139225	275.8	108770	328.3
PASA	30439	30975	2720.6	140185	277	109210	329.2

*EVM = EVidenceModeler; PASA = Program to Assemble Spliced Alignments.

Table 6(on next page)

Summary statistics for functional annotation

Table 6. Summary statistics for functional annotation

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

NR = non-redundant protein sequences; GO = gene ontology; KEGG = Kyoto Encyclopaedia of Genes and Genomes

Figure 1

Pseudomolecule construction of sweet cherry by assigning scaffolds to seven genetic maps

KS, CP, CK, AC, VG, JWF, and JWF1 are the 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 .

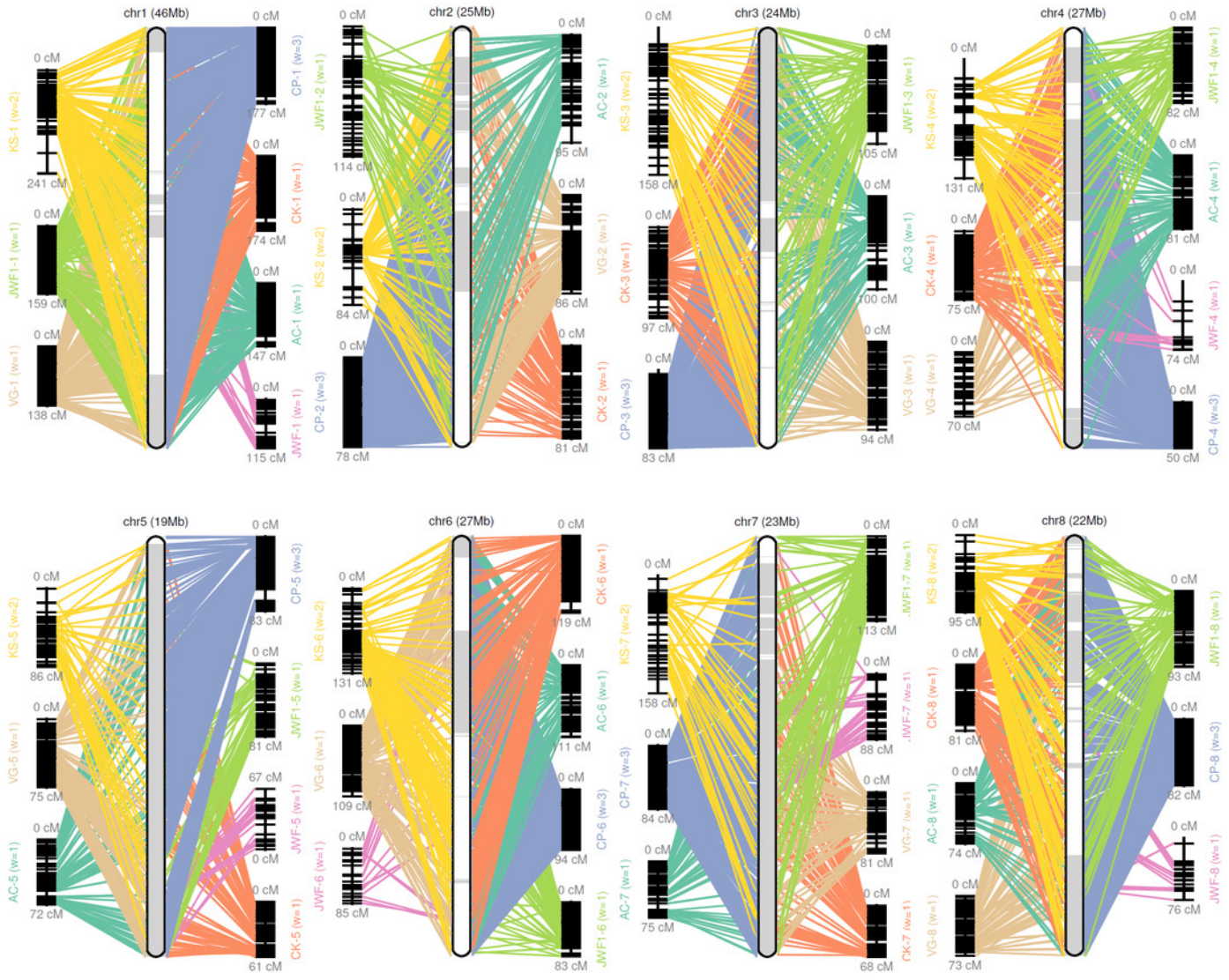
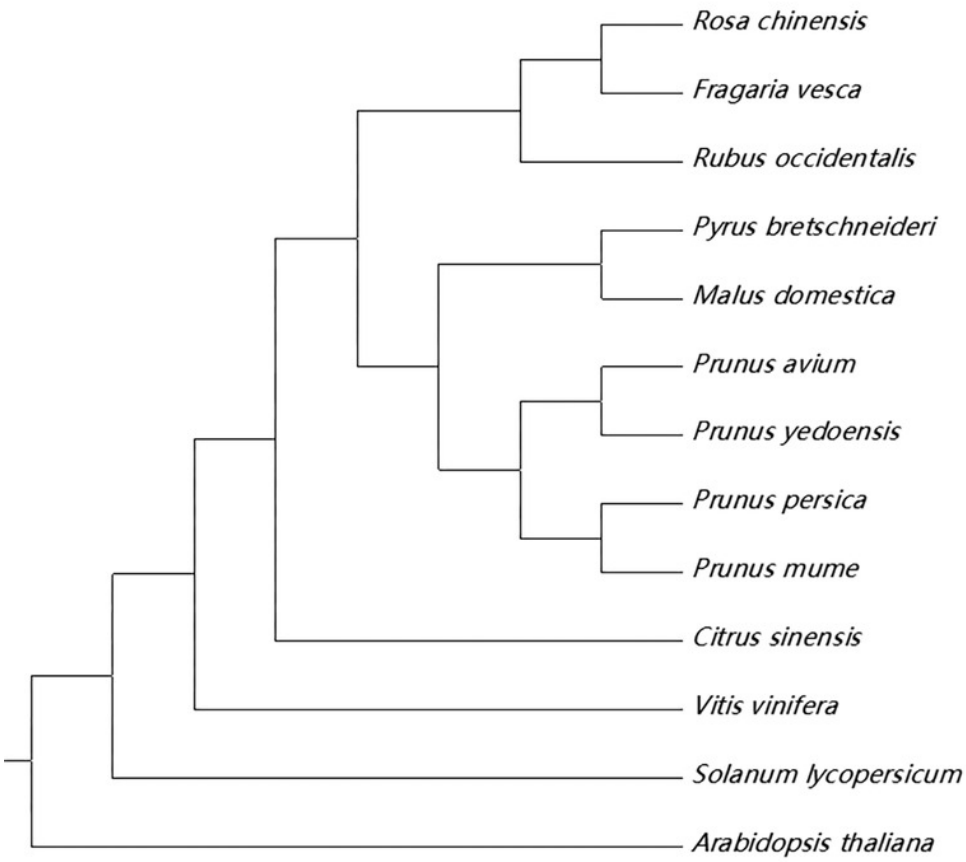


Figure 2

Species tree and gene family expansion analysis.

A species tree were also constructed by using STRIDE , as part of OrthoFinder. A comparison was conducted using CAFÉ (version 4.2) . Compared with other plant genomes, 1,017 gene families had expanded and 3,643 gene families had contracted in sweet cherry genome



Gene family	
Expansions	Contractions
2606	1112
1246	3052
1337	4254
7963	3052
6028	1052
1012	3642
6026	2505
687	2051
3201	1744
1344	10218
1289	10857
2835	9505
2437	10805