Two telomere-to-telomere, gap-free genome assemblies and comparisons revealed the conserved key genes associated with sugar accumulation in *Rubus* genus

Xiaopeng Li[1,†], Xianyan Han[1,3,†], Shoucheng Liu[1,†], Qunying Zhang[2,†], Jiyuan Guan[2,†], Yajun Tang[1,†], Meng Zhang[1], Hongli Lian[4], Pengbo Xu[4], Mengzhuo Zheng[1], Kui Li[1], Guilian Sun[1], Yan Sun[1], Yuanxin Dong[1], Xiaojie Lin[1], Ye Liang[1], Zhengdong Wang[1], Guochen Qin[1], Bosheng Li[1], Houcheng Zhou[3], Guohui Yang[5], Zhongchi Liu[6,7], Hang He[1,*], Junhui Zhou[1,*]

[1] Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences in Weifang, Shandong Provincial Key Laboratory of Precision Molecular Crop Design and Breeding, Shandong 261325, China

[2] Botanic Garden of Guizhou Province, Guiyang city, Guizhou Province 550081, China

[3] National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou 450009, China

[4] School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China

[5] College of Horticulture & Landscape Architecture, Northeast Agricultural University, Harbin 150030, China

[6] Faculty of Synthetic Biology, Shenzhen University of Advanced Technology, Shenzhen 518100, China

[7] Department of Cell Biology & Molecular Genetics, University of Maryland, College Park, MD 20742, United States

[*] Correspondence: junhui.zhou@pku-iaas.edu.cn, hehang@pku.edu.cn,

[†] These authors contributed equally to this work.

Abstract

Two telomere-to-telomere, gap-free genomes for Rubus species were sequenced and assembled by using an integrative approach that combined ultra-long reads from Oxford Nanopore Technology (ONT), PacBio high-fidelity (HiFi) long reads, Illumina paired-end short reads, and High-throughput Chromosome Conformation Capture (Hi-C) data, resulted in the first gap-free reference genomes of the Rubus genus. By comparative genomics and transcriptomics, we revealed the conserved key genes associated with sugar accumulation in Rubus genus, providing candidate target sites for improving fruit quality in Rubus. We also established a database for documenting all the genomic, transcriptomic, and phenotypic data generated in this study.

Dear Editor,

The genus *Rubus*, commonly known as brambles, comprise small perennial berry shrubs within the Rosaceae family. These plants are composed of over 740 species and boast a rich diversity of germplasm resources with a cosmopolitan distribution spanning diverse latitudes (Janick and Moore, 1996; Jennings, 1988), rendering them of exceptional value for evolutionary studies. *Rubus* species, recognized by the United Nations Food and Agriculture Organization as an international 'third-generation fruit', are particularly notable for their low sugar content and high abundance of bioactive compounds. These properties confer significant nutritional and medical value, including the ability to scavenge reactive oxygen species (ROS), reduce blood pressure, prevent carcinogenesis and cardiovascular diseases, and decelerate aging-related processes (Kaume et al., 2012; Moyer et al., 2002). Additionally, *Rubus* species exhibit a relatively short life cycle, asexual propagation, and most of them possess a small and simple genome (2n=2x=14), along with an efficient transformation system (Miller, 2019), endowing them with unparalleled potential to serve as an ideal model system for both fundamental and applied research in woody fruit crops.

The first genome of a *Rubus* species to be assembled and annotated was that of the black raspberry (*Rubus occidentalis* L.) (VanBuren et al., 2016). This was followed by the genomes of *Rubus chingii* Hu (Wang et al., 2021) and *Rubus idaeus* 'Anitra' (Davik et al., 2022). Previously, we also contributed to this field by assembling the most complete chromosome-level genome of the red raspberry cultivar 'Joan J' (*Rubus idaeus*) using third-generation sequencing technology (Zhou et al., 2023). However, despite these advancements, the red raspberry genomes still contain numerous unassembled and unannotated gaps, particularly in the centromeric and telomeric regions. These gaps present significant challenges for further genomic and functional

studies, highlighting the need for continued efforts to refine and improve the genomic resources available for *Rubus* species.

Here, for the first time, we present two highly continuous, completely gap-free reference genomes of the *Rubus* genus: *Rubus hirsutus* Thunb. 'Penglei' (XMM) and *Rubus eustephanos* Focke ex Diels 'Dahongpao' (DHP). These two species, widely distributed in southern China, exhibit close phylogenetic relationships (Supplemental Figure 1) and similar phenotypic traits (Supplemental Figure 2), yet they display distinct sugar accumulation levels in the ripe fruits (Supplemental Table 1), making them ideal candidates for investigating the mechanisms underlying sugar accumulation in the *Rubus* genus.

The telomere-to-telomere (T2T) genomes of XMM and DHP were assembled using an integrative approach that combined ultra-long reads from Oxford Nanopore Technology (ONT), PacBio high-fidelity (HiFi) long reads, Illumina paired-end short reads, and High-throughput Chromosome Conformation Capture (Hi-C) data (Supplemental Figures 3 and 4; Supplemental Table 2). This effort resulted in two of the most complete *Rubus* genomes to date (Supplemental Table 3), featuring fully resolved telomeres enriched with the plant telomeric repeat motif (TTTAGGG) (Supplemental Table 4). The genome sizes are 213.53 Mb for XMM and 218.26 Mb for DHP (Figure 1A; Supplemental Table 5). Both assemblies achieved exceptional BUSCO completeness scores of 99.18% for XMM and 99.23% for DHP. Furthermore, the mapping rates for HiFi and ONT reads were 99.90% and 99.99% for XMM (with 100% coverage each) and 99.17% and 99.98% for DHP (with coverage of 99.87% and 99.94%, respectively). The LTR Assembly Index (LAI) values exceeded gold standards, with scores of 26.49 for XMM and 20.53 for DHP, while consensus quality value (QV) scores ranged from 65 to 68 (Supplemental Table 6). Additionally, we identified two types of centromeric satellites (CEN18 and CEN17), which share 84% sequence

similarity and occupy 19.94% and 23.90% of the centromeric regions in XMM and DHP, respectively (Figure 1A; Supplemental Figure 5; Supplemental Table 7).

Through a combination of *ab initio* and homology-based approaches, 46.59% (99.48 Mb) and 47.18% (102.98 Mb) of repetitive sequences were identified in the XMM and DHP genomes, respectively (Supplemental Table 8). Following repeat masking, gene structure annotation was performed using a hybrid strategy that integrated *de novo* prediction, homology-based methods, and transcriptome evidence (Supplemental Figure 6). This comprehensive annotation revealed 28,204 and 28,569 protein-coding genes in the XMM and DHP genomes, respectively (Supplemental Table 9). These annotations achieved exceptional coverage of 98.50% and 98.45% of the complete BUSCO gene set, surpassing the quality of all previously released *Rubus* genomes (Supplemental Table 10).

To explore chromosomal variations and trait diversity between the XMM and DHP genomes, we conducted a comparative genomic analysis (Figures 1A and 1B). This analysis identified 380 syntenic blocks with over 79.00% 1-to-1 collinearity, encompassing 939,717 single nucleotide polymorphisms (SNPs), 180,328 insertions/deletions (InDels), and 867 structural variants, including 349 translocations, 408 duplications, and 110 inversions (Supplemental Table 11). Notably, a large inversion on chromosome 7, containing 33 annotated genes, was confirmed by aligning Hi-C reads from DHP to the XMM genome and visualizing the Hi-C interaction heatmap (Supplemental Figure 7A). Furthermore, a significant inversion was observed specifically in the centromeric region of chromosome 6 in both XMM and DHP genomes, which was validated through Sanger sequencing (Figures 1B and 1C; Supplemental Figure 7B; Supplemental Table 12).

Additionally, we identified 2,013 and 2,146 presence/absence variation (PAV) sequences in the XMM and DHP genomes, respectively, totaling 1.33 Mb and 1.28 Mb. Over 87.50% of these PAV regions were shorter than 1 kb (Supplemental Figure 8A). Of particular interest, a sugar transporter gene (*MFS1*), present in the XMM genome but absent in DHP, exhibited higher expression levels across six tissues in XMM (Figure 1A; Supplemental Figures 8B-8D).

To explore the evolutionary history of the two *Rubus* species, we conducted a phylogenetic analysis using ten other published Rosaceae genomes (*Prunus armeniaca*, *Potentilla micrantha*, *Fragaria vesca*, *Rosa rugosa*, *Rosa wichuraiana*, *Rubus occidentalis*, *Rubus idaeus* 'Joan J', *Rubus argutus*, *Rubus chingii* Hu, and *Rubus corchorifolius*) from the Genome Database for Rosaceae (Jung et al., 2019). Based on sequence similarity, we identified 33,275 ortholog groups, encompassing 27,209 genes from XMM and 27,527 genes from DHP (Figure 1D). A phylogenetic tree was constructed for these 12 Rosaceae species using 3,969 conserved single-copy genes. The analysis revealed that XMM and DHP form a distinct clade, having diverged from the common ancestor of four *Rubus* species approximately 12.44 million years ago (Mya). The divergence between XMM and DHP occurred more recently, around 3.21 Mya. Additionally, *Rubus chingii* Hu and *Rubus corchorifolius* were found to be closely related to XMM and DHP.

Following the divergence of XMM and DHP, 198 gene families underwent significant expansion, while 617 families experienced contraction. Notably, gene families associated with glycoprotein biosynthetic and metabolic process expanded, whereas those involved in (1->3)-beta-D-glucan biosynthetic contracted (corrected $P < 0.05$) (Supplemental Figures 9 and 10). Further analysis of expanded and contracted gene families, along with functional gene annotations, revealed that two sugar-related gene

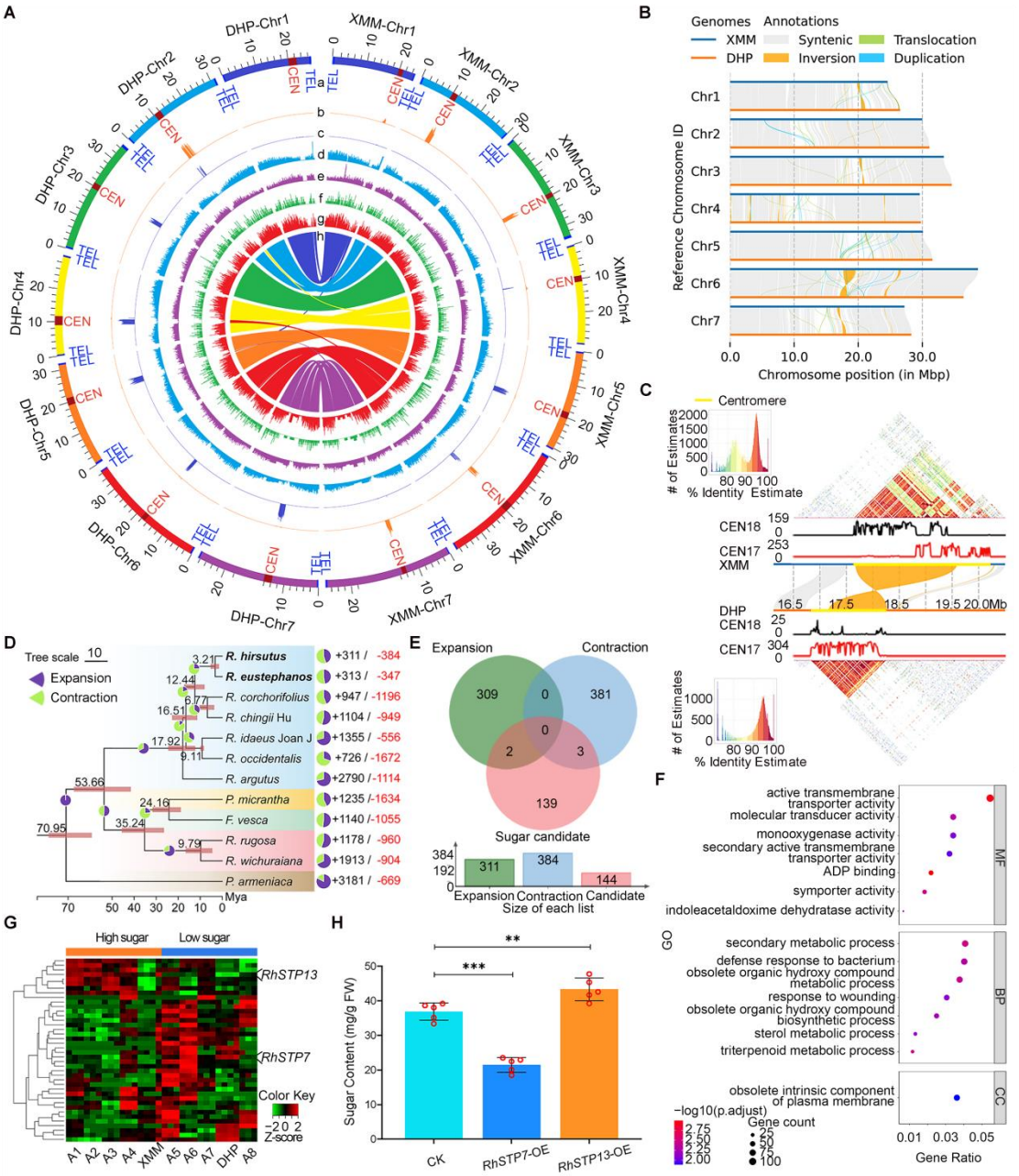
families in XMM have undergone expansion, while three sugar-related gene families have undergone contraction (Figure 1E; Supplemental Table 13).

Based on our UPLC-MS/MS results (Supplemental Figure 11A), XMM exhibited a remarkably high sugar content, approximately 3.64 times greater than that of DHP. To broaden our analysis, we carried out transcriptome sequencing and annotated eight other *Rubus* materials. These 10 materials were categorized into high-sugar and low-sugar groups according to their total fruit sugar content, revealing 3,531 differentially expressed genes (DEGs) (Supplemental Table 14) which were significantly enriched in the molecular function 'active transmembrane transporter activity' (GO:0022804) (Figure 1F; Supplemental Table 15). We further identified 40 candidate genes primarily involved in sugar transport processes that likely contribute to the differences in sugar accumulation between XMM and DHP by integrating differential gene expression analysis, homologous gene comparison, and functional annotation (Figure 1G; Supplemental Table 16).

Then we selected two most striking DEGs, *RhSTP13* and *RhSTP7*, which were enriched in the 'active transmembrane transporter activity' and 'symporter activity' GO terms and exhibited highest contrasting expression patterns: *RhSTP13* was upregulated in high-sugar samples, while *RhSTP7* was downregulated. To investigate their functional roles during *Rubus* fruit development, we predicted the protein structures of both genes and found no significant structural differences (Supplemental Figures 11B and 11C). We then constructed overexpression vectors for *RhSTP13* and *RhSTP7* and performed transient transformation assays in *Rubus argutus*, a cultivar with medium sugar content (Supplemental Figure 11D). Overexpression of *RhSTP13* led to a 17.48% increase in sugar concentration, whereas overexpression of *RhSTP7* resulted in a 41.73% decrease in total sugar content (Figure 1H; Supplemental Table 17). The relative expression levels of *RhSTP13* and *RhSTP7* were positively correlated with the expected changes

in sugar content (Supplemental Figure 11E). In addition, *RhSTP13* belongs to an expanded gene family in XMM. These high-quality, gap-free genomes provide a robust foundation for advancing genomic and functional studies in the *Rubus* genus, particularly in understanding the genetic basis of sugar accumulation and other economically important traits.

We have also established a comprehensive database for the *Rubus* genus to catalog all genomic, transcriptomic, and phenotypic data generated in this study. Currently, the database includes the telomere-to-telomere gap-free genome assemblies of XMM and DHP, which were utilized in this research. Additionally, this database offers basic analysis tools, such as batch query and BLAST, to support gene discovery and genomic investigations across all species within the *Rubus* genus.



206

207 **Figure 1. Genome assemblies, variation, evolution, and sugar-related genes of**
208 ***Rubus hirsutus* Thunb. 'Penglei' (XMM) and *Rubus eustephanos* Focke ex Diels**
209 **'Dahongpao' (DHP).**

210 **(A)** Circular diagram of XMM and DHP reference genomes. (a) Chromosomes are
211 represented by centromeres (dark red) and telomeres (blue); (b) CEN18 density; (c)
212 CEN17 density; (d) Gene density; (e) TE density; (f) PAV density; (g) SNP density; (h)
213 Collinear lines at the center of diagram highlight homoeologous chromosomes
214 relationships and non-homoeologous regions.

(**B**) Genomic alignments between XMM and DHP. Inversions, duplications and translocations are marked with orange, blue and green ribbons, respectively.

(**C**) Synteny, structural rearrangements, monomer, and identity distribution in the 16-20 Mbp region of chromosome 6 between XMM and DHP.

(**D**) Phylogenetic relationships and divergence times between raspberries and other Rosaceae species. The black numbers close to the divergence nodes indicate the divergence times and the red bars represent the 95% confidence intervals. The red and black numbers indicate expanded and contracted ortholog groups at the corresponding node. Scale bar corresponds to 10 Mya.

(**E)** Venn diagram depicts the family of sugar-related, contraction and expansion gene family in XMM.

(**F**) GO enrichment analyses of DEGs in ripe fruits between high-sugar and low-sugar groups.

(**G**) Sugar-relate allelic expression profiles between five high-sugar and five low-sugar samples. A1-A8 represented *Rubus idaeus* Tulamen, *Rubus idaeus* Erika, *Rubus idaeus* Orange Miracle, *Rubus chingii* Hu, *Rubus setchuenensis* Bureau et Franch, *Rubus argutus* Hillquist, *Rubus idaeus* Heritage and *Rubus occidentalis*, respectively. The color gradient ranging from red to green is indicative of the relative expression values of the genes.

(**H**) Total sugar content in mature fruits of *RhSTP7*-OE and *RhSTP13*-OE based on UPLC-MS/MS results. Cas9-OE infiltrated-fruits were used as control. Two biological replicates consisting of five fruits were prepared for determination of sugar content.

Data Availability

All the raw sequencing data generated for this project, including the Nanopore ONT data, PacBio HiFi data, and Illumina NovaSeq data, as well as the assemblies, are archived at the National Genomics Data Center.

SUPPLEMENTAL INFORMATION

Supplementary information for this online version is accessible on the online platform.

FUNDING

This study was supported by the Taishan Scholars Program (J.Z.), the Dean's Foundation from Peking University Institute of Advanced Agricultural Sciences, the funding from Shandong Laboratory of Advanced Agricultural Sciences at Weifang, and Guizhou Provincial Scientific and Technological Program (QKHFQ[2024]004-1).

AUTHOR CONTRIBUTIONS

J.Z., H.H., and Q.Z. conceived and supervised the research. X.L., and S.L. performed the genome assembly and other bioinformatics analyses. X.H., J.G., Y.T., and G.Y. performed phenotyping and collected the samples for sequencing. X.H., and G.S. performed experiments for functional validation of candidate genes. M.Z., H.L., P.X., Y.S., Y.D., X.J.L., Y.L., Z.W., G.Q., B.L., H.Z., and Z.L. provided assistance and offered suggestions. X.L., and S.L. designed and constructed the framework of the database. X.L., X.H., S.L., M.Z.Z., K.L., H.H., and J.Z. wrote and revised the manuscript. J.Z., and Q.Z. provided the funding support. All authors read and approved the manuscript.

ACKNOWLEDGEMENT

We extended our sincere appreciation to the Bioinformatics Platform at Peking University Institute of Advanced Agricultural Sciences (PKU-IAAS) for providing high-performance computing resources. We would like to thank Dr. Xiaoli Lin and Zhiying Lou (PKU-IAAS) for helping with biochemical experiments and protein structure prediction. We would like to thank Dr. Dongdong Lu (PKU-IAAS) for helping with the flow cytometry experiment. We would also like to thank Bailong Song and Hongxiao Sun from PKU-IAAS Mass Spectrometry Platform for their preliminary

experiment in quantification of sugar content. Accurate quantification of total sugar content in fruits of *Rubus* species was performed by UPLC-MS/MS platform from Guocangjian Biotech (targetcrop.com).

References

Davik, J., Roen, D., Lysoe, E., Buti, M., Rossman, S., Alsheikh, M., Aiden, E.L., Dudchenko, O., and Sargent, D.J. (2022). A chromosome-level genome sequence assembly of the red raspberry (Rubus idaeus L.). PLoS One **17**:e0265096. https://doi.org/10.1371/journal.pone.0265096.

Janick, J., and Moore, J.N. (1996). Fruit breeding, tree and tropical fruits (John Wiley & Sons).

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

Jung, S., Lee, T., Cheng, C.H., Buble, K., Zheng, P., Yu, J., Humann, J., Ficklin, S.P., Gasic, K., Scott, K., et al. (2019). 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. Nucleic Acids Res. **47**:D1137-d1145. https://doi.org/10.1093/nar/gky1000.

Kaume, L., Howard, L.R., and Devareddy, L. (2012). The blackberry fruit: a review on its composition and chemistry, metabolism and bioavailability, and health benefits. J. Agric. Food Chem. **60**:5716-5727. https://doi.org/10.1021/jf203318p.

Miller, S. (2019). Gene editing of red raspberry (Rubus idaeus L.) with CRISPR/Cas9 knocking out F3'H. Norwegian University of Life Sciences, Ås.

Moyer, R.A., Hummer, K.E., Finn, C.E., Frei, B., and Wrolstad, R.E. (2002). Anthocyanins, phenolics, and antioxidant capacity in diverse small fruits: vaccinium, rubus, and ribes. J. Agric. Food Chem. **50**:519-525. https://doi.org/10.1021/jf011062r.

VanBuren, R., Bryant, D., Bushakra, J.M., Vining, K.J., Edger, P.P., Rowley, E.R., Priest, H.D., Michael, T.P., Lyons, E., and Filichkin, S.A. (2016). The genome of black raspberry (Rubus occidentalis). Plant J. **87**:535-547.

Wang, L., Lei, T., Han, G., Yue, J., Zhang, X., Yang, Q., Ruan, H., Gu, C., Zhang, Q., Qian, T., et al. (2021). The chromosome-scale reference genome of Rubus chingii Hu provides insight into the biosynthetic pathway of hydrolyzable tannins. Plant J. **107**:1466-1477. https://doi.org/10.1111/tpj.15394.

Zhou, J., Li, M., Li, Y., Xiao, Y., Luo, X., Gao, S., Ma, Z., Sadowski, N., Timp, W., Dardick, C., et al. (2023). Comparison of red raspberry and wild strawberry fruits reveals mechanisms of fruit type specification. Plant Physiol. **193**:1016-1035. https://doi.org/10.1093/plphys/kiad409.