



OPEN

DATA DESCRIPTOR

# Haplotype-resolved T2T gap-free genomes of the winegrape cultivar Cabernet Sauvignon

Falak Sher Khan<sup>1,2,5</sup>, Tiepeng Sun<sup>2,5</sup>, Xiangfeng Wang<sup>2,3,4</sup>, Xuenan Zhang<sup>2</sup>,  
Naomi Abe-Kanoh<sup>3,4</sup>, Ying Hua Su<sup>1</sup>✉, Li Guo<sup>2</sup>✉ & Wenxiu Ye<sup>2</sup>✉

Cabernet Sauvignon (CS), a cultivar of winegrape (*Vitis vinifera*), is among the most renowned winegrape varieties globally. In this study, we released the haplotype-resolved telomere-to-telomere (T2T) CS genome assembled using a combination of PacBio HiFi, ONT ultra-long, and Hi-C sequencing data. The two T2T gap-free haplotype-resolved assemblies CS-T2T.Hap1 and CS-T2T.Hap2 sized 491.11 Mb (contig N50 = 25.09 Mb) and 491.90 Mb (contig N50 = 25.51 Mb), respectively. Genome annotation predicted a total of 36,456 genes in CS-T2T.Hap1 and 35,471 genes in CS-T2T.Hap2. By genome comparison, we discovered and validated megabase inversion events on Chromosome 03, 11, 18 and 19, which are not prevalent in other haplotype-resolved *V. vinifera* genomes. In summary, this haplotype-resolved T2T genome represents an essential genomic resource for Cabernet Sauvignon, and lays the foundation for its genetic studies, improvement and utilization.

## Background & Summary

The cultivated grapevine (*V. vinifera* L.) is one of the most valuable perennial fruit crops cultivated for wine-making and fresh consumption worldwide<sup>1</sup>. ‘Cabernet Sauvignon’ is a major red wine cultivar renowned for its thick, resilient skins and high concentration of phenolic compounds. Wine from this variety exhibits a robust tannic structure and complex flavor profile dominated by notes of blackcurrant, plum, and green bell pepper<sup>2</sup>. These traits are attributed to the grapevine’s unique genetic lineage and environmental adaptability, making it one of the most widely cultivated wine grape varieties globally. Originating from a natural cross between Sauvignon Blanc and Cabernet Franc in the Aquitaine region of France prior to the seventeenth century, its historical and viticultural significance is further underscored by its genetic stability and widespread adoption across diverse terroirs<sup>3</sup>.

Since the first *V. vinifera* genome was assembled in 2007 using the grapevine PN40024<sup>4</sup>, originally derived from 9 selfings of ‘Helfensteiner’ (cross of ‘Pinot noir’ and ‘Schiava grossa’)<sup>5</sup>, the field of grapevine genomics has progressed rapidly, with high-quality reference assemblies now available for diverse species and cultivars<sup>6–9</sup>. In 2023, of particular significance, telomere-to-telomere (T2T) genome has been completed for PN40024<sup>10</sup>. Since then, several economically vital varieties such as the white wine grape ‘Chardonnay’<sup>8</sup>, ‘Thomson seedless’<sup>11</sup>, ‘Baimunage’, ‘Black Monukka’, ‘Muscat Hamburg’, ‘Hongmunage’, ‘Manicule Finger’, ‘Thompson’<sup>9</sup> have been sequenced. The genome of ‘Cabernet Sauvignon’ was initially assembled in 2016<sup>12</sup>, followed by the release of several chromosomal-level versions<sup>8,13</sup>. Despite its global prominence as the most widely planted wine grape, ‘Cabernet Sauvignon’ lacks a T2T gap-free reference genome, which hinders comprehensive understanding of the molecular mechanism behind its complex traits, including disease resistance, berry development, and terroir-specific flavor profiles, thus limiting its improvement and utilization.

<sup>1</sup>State Key Laboratory of Wheat Improvement, College of Life Sciences, Shandong Agricultural University, Tai’an, 271018, Shandong, China. <sup>2</sup>Shandong Key Laboratory of Precision Molecular Crop Design and Breeding, Shandong Laboratory of Advanced Agricultural Sciences in Weifang, Peking University Institute of Advanced Agricultural Sciences, Weifang, 261325, China. <sup>3</sup>United Graduate School of Agricultural Science, Iwate University, 3-18-8 Ueda, Morioka, Iwate, 020-8550, Japan. <sup>4</sup>Department of Food, Life and Environmental Science, Faculty of Agriculture, Yamagata University, Wakaba-machi 1-23, Tsuruoka, Yamagata, 997-8555, Japan. <sup>5</sup>These authors contributed equally: Falak Sher Khan, Tiepeng Sun. ✉e-mail: [suyh@sdau.edu.cn](mailto:suyh@sdau.edu.cn); [li.guo@pku-iaas.edu.cn](mailto:li.guo@pku-iaas.edu.cn); [wenxiu.ye@pku-iaas.edu.cn](mailto:wenxiu.ye@pku-iaas.edu.cn)

| Library type | Reads number | Total length (Gb) | N50 length (bp) | Coverage* |
|--------------|--------------|-------------------|-----------------|-----------|
| NGS          | 165,785,460  | 24.74             | 2 × 150         | 50×       |
| PacBio HiFi  | 4,752,646    | 87.86             | 18,062          | 178×      |
| Hi-C         | 413,992,094  | 61.78             | 2 × 150         | 126×      |
| ONT          | 531,173      | 49.71             | 100,000         | 101×      |

**Table 1.** Statistic of sequencing data for ‘Cabernet sauvignon’ genome assembly. \*The coverage was calculated based on the Hap1 genome assembly of 491.10 Mb.

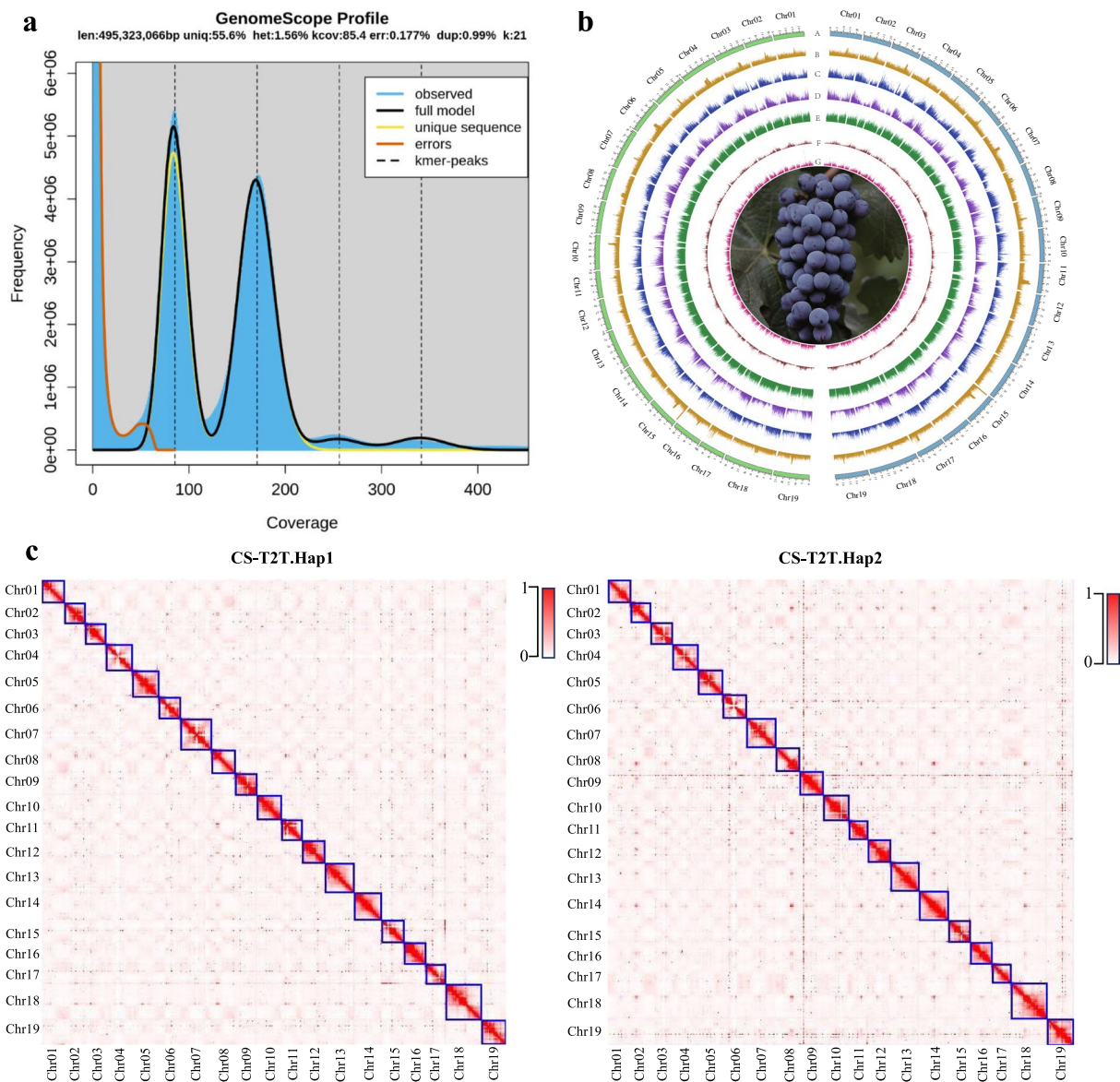
In this study, we assembled the haplotype-resolved T2T genome of ‘Cabernet Sauvignon’ using PacBio-HiFi, ONT ultra-long, and Hi-C sequencing data. The two haplotype-resolved T2T genome assemblies are gapless and sized 491.11 Mb and 491.90 Mb with N50 length of 25.09 Mb and 25.51 Mb, respectively. A total of 36,456 and 35,471 protein-coding genes were annotated in the two haplotype genomes combining ab initio prediction, homology proteins and transcriptome evidence. Vigorous quality assessment demonstrated the high accuracy and completeness of the assembly and annotation. These T2T assemblies and annotations improve the genomic resources available for *V. vinifera* cultivar ‘Cabernet Sauvignon’ and provide a crucial reference genome for its molecular studies and improvement.

## Methods

**Plant resources and genome sequencing.** The ‘Cabernet Sauvignon’ (CS) plants used in this study were obtained in 2022 from the National Repository for Grapevine and Peach (Zhengzhou) with accession number G0166. This study is based on a single clone derived from three individual plants propagated from this accession. The ‘Cabernet Sauvignon’ was cultivated in the vineyards at Peking University Institute of Advanced Agricultural Sciences, China. Genomic DNA was extracted from grapevine leaves using the CTAB (cetyltrimethylammonium bromide) method<sup>14</sup>. The Agilent 4200 Bioanalyzer (Agilent Technologies) was used to evaluate the DNA integrity. For HiFi sequencing library preparation, 15 µg of high-molecular-weight genomic DNA was utilized to generate a PacBio SMRTbell library with the SMRT Express Template Prep Kit 2.0 (Pacific Biosciences). Circular consensus sequencing (CCS) was conducted on a PacBio Sequel IIe system (Biomarker Technologies Corporation, Qingdao, China) to yield 87.86 Gb of HiFi data, corresponding to 178× genome coverage, with read N50 length of 18.06 kb (Table 1). Ultra-long-read sequencing was performed using the Oxford Nanopore Technologies (ONT) platform, where size-selected long DNA fragments were prepared according to the SQK-LSK109 Kit protocol. Sequencing was carried out on the PromethION sequencer at Single-Molecule Sequencing Platform of Annoroad. A total of 49.71 Gb ONT data was generated, corresponding to 101× genome coverage (Table 1). The Next-Generation Sequencing (NGS) and High-throughput chromosome conformation capture sequencing (Hi-C) datasets used in this study were obtained from published sources, corresponding to 50× and 126× genome coverage, respectively<sup>8</sup> (Table 1).

**Genome assembly.** To estimate fundamental characteristics of the ‘Cabernet Sauvignon’ genome, such as genome size, heterozygosity, we generated PacBio HiFi reads and performed a K-mer frequency analysis. The K-mer analysis was performed using Jellyfish v2.3.0<sup>15</sup> with a K-mer size of 21 and used for the genome survey by GenomeScope v1.0<sup>16</sup>. The genome size of ‘Cabernet Sauvignon’ was estimated at 495.32 Mb, with the heterozygosity rate of 1.56% (Fig. 1a). To assemble the haplotype-resolved T2T genomes for ‘Cabernet Sauvignon’, PacBio HiFi (178×), ONT ultra-long (101×), and Hi-C (126×) reads were integrated to produce an initial assembly of two haplotypes using hifiasm v0.24.0<sup>17</sup> with the parameters ‘-hom-cov-h1-h2-ul’. To eliminate mitochondrial and chloroplast sequences from the genome assembly, the genome contigs were aligned to the *V. vinifera* mitochondrial genome (GenBank: FM179380.1) and chloroplast genome (GenBank: DQ424856.1) using minimap2 v2.28<sup>18</sup> with the parameter ‘-x asm5’. Contigs exhibiting ≥ 50% alignment coverage to plastid sequences were excluded from the assembly. By comparing contigs with all NCBI RefSeq bacterial genomes using blast-2.13.0+ with the parameter ‘-task megablast’, other common contamination sequences (such as microbial DNAs) were further eliminated. The high-quality haplotype-resolved assemblies were scaffolded into chromosome-scale genomes using Hi-C data with Juicer<sup>19</sup> and 3d-DNA<sup>20</sup> pipeline. Manual corrections to the scaffolded pseudochromosomes were performed using Juicebox v1.11.08<sup>21</sup>, guided by Hi-C chromatin interaction patterns observed in contact heatmaps. To fill the gaps in the chromosome-scale assemblies, ONT ultra-long and HiFi reads were mapped to the haplotype-resolved genomes, and the aligned reads were used to fill gaps. The final genome assembly produced two gap-free haplotypes (Fig. 1b), haplotype 1 (CS-T2T.Hap1, 491.11 Mb) and haplotype 2 (CS-T2T.Hap2, 491.90 Mb) (Table 2) with each genome consists of 19 chromosomes (Fig. 1c). The scaffold N50 lengths for CS-T2T.Hap1 and CS-T2T.Hap2 were 25.09 Mb and 25.51 Mb, respectively (Table 2). The telomere sequences were identified using Tandem Repeat Finder (TRF) v4.09.1<sup>22</sup>, and the result file was converted into GFF3 format via TRF2GFF (<https://github.com/Adamtaranto/TRF2GFF>) to detect the seven-base repeats. In summary, ‘Cabernet Sauvignon’ assembly can be regarded as having achieved telomere-to-telomere and gap-free completeness.

To verify true-to-type identity for our ‘Cabernet Sauvignon’ (CS) plant material, a k-mer-based genomic analysis was performed via MASH v2.3<sup>23</sup>. To verify genomic consistency, we retrieved publicly available CS sequencing data (SRR3346861, SRR3346862) from the NCBI Sequence Read Archive (SRA), followed by pairwise comparisons with the assembled CS haplotype genomes and the published PN40024<sup>10</sup> reference genome (Supplementary Table S1). The results demonstrate no detectable genetic divergence between the CS genome



**Fig. 1** Haplotype-resolved genome assembly of ‘Cabernet Sauvignon’. **(a)** Genome survey based on K-mer analysis using GenomeScope1.0. **(b)** Genomic features of CS genome. Tracks (A) to (G): chromosomes, GC content, gene density, exon density, repeat density, LTR/Gypsy density, LTR/Copia density. Windows: 50 Kb. Left (green): Haplotype 2. Right (blue): Haplotype 1. **(c)** Hi-C contact map of CS assembly.

| Genomic feature        | CS-T2T.Hap1 | CS-T2T.Hap2 |
|------------------------|-------------|-------------|
| Number of contigs      | 19          | 19          |
| Total length (bp)      | 491,107,025 | 491,898,969 |
| Contig N50 (bp)        | 25,085,764  | 25,511,701  |
| Number of gaps         | 0           | 0           |
| Number of telomeres    | 36          | 36          |
| GC content (%)         | 34.94       | 35.00       |
| Repeat content (%)     | 53.03       | 52.88       |
| BUSCO (%)              | 98.3        | 98.2        |
| LAI                    | 16.46       | 18.25       |
| QV                     | 49.13       | 49.83       |
| HiFi mapping ratio (%) | 99.99       | 99.99       |
| ONT mapping ratio (%)  | 100.00      | 100.00      |

**Table 2.** Assembly statistics for the two haplotypes of ‘Cabernet sauvignon’.

| Repeat type     | CS-T2T.Hap1 |         | CS-T2T.Hap2 |         |
|-----------------|-------------|---------|-------------|---------|
|                 | Length (bp) | Percent | Length (bp) | Percent |
| LTR             | 194,499,814 | 39.60%  | 199,568,310 | 40.60%  |
| LINE            | 20,117,018  | 4.10%   | 20,023,284  | 4.07%   |
| SINE            | 1,195,499   | 0.24%   | 1,214,893   | 0.25%   |
| DNA Transposons | 38,595,821  | 7.86%   | 38,400,562  | 7.81%   |
| Total           | 254,408,152 | 51.80%  | 259,207,049 | 52.70%  |

**Table 3.** Statistics of annotated repeat elements in the genome of ‘Cabernet Sauvignon’.

| Feature                 | CS-T2T.Hap1 | CS-T2T.Hap2 |
|-------------------------|-------------|-------------|
| Annotated genes         | 36,456      | 35,471      |
| NR                      | 34,128      | 32,634      |
| eggNOG                  | 28,549      | 27,110      |
| Pfam                    | 25,668      | 24,367      |
| Swiss-port              | 23,419      | 23,174      |
| KEGG                    | 13,548      | 13,444      |
| GO                      | 12,704      | 12,902      |
| Non-redundant annotated | 34,142      | 32,648      |

**Table 4.** Summary of protein-coding genes functional annotation for the two haplotypes of ‘Cabernet sauvignon’.

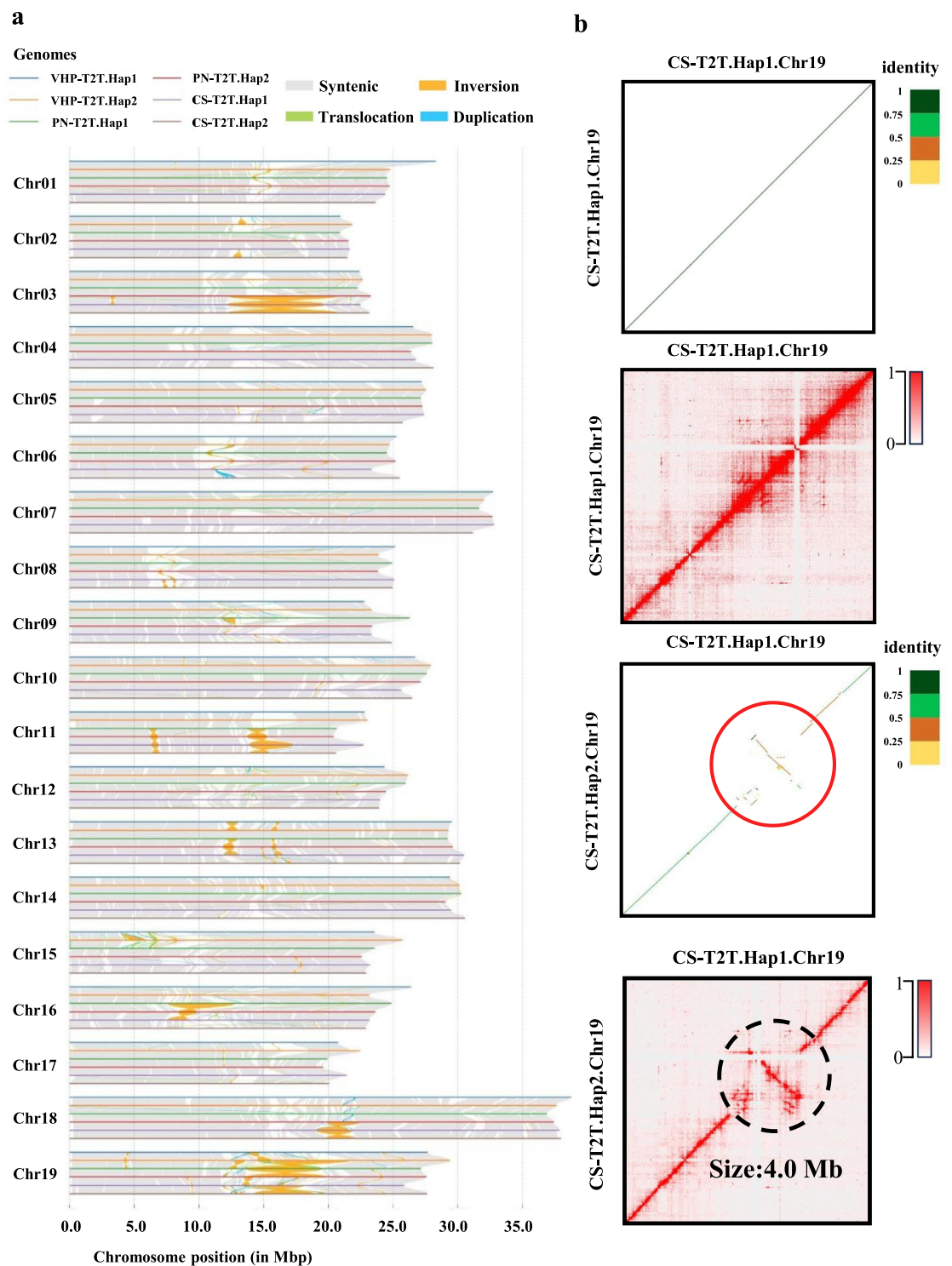
assemblies generated in this study and the published CS genome assemblies. These findings further confirm that the plant material used in this study represents authentic ‘Cabernet Sauvignon’.

**TE annotation.** Before genome annotation, a nonredundant transposable element (TE) library was constructed using the Extensive *de novo* TE Annotator (EDTA) v2.2.1<sup>24</sup> with parameters ‘–sensitive 0–anno 1–step all’. For haplotype-resolved assembly, approximately 51.80–52.70% of genome was classified as repetitive elements, which included 39.60–40.60% long terminal repeats (LTRs), 4.07–4.10% long interspersed nuclear elements (LINEs), 0.24–0.25% short interspersed nuclear elements (SINEs), and 7.86–7.81% DNA transposon elements, respectively (Table 3).

**Genome annotation.** Gene model prediction was performed using a comprehensive pipeline that integrated multiple data sources after masking repetitive elements, including: ab initio predictions, homologous protein comparisons, and RNA-seq evidence. The ab initio gene prediction was implemented by training the GeneMark-ET model with BRAKER2 v2.1.6<sup>25</sup> and further refined it by training the semi-HMM model SNAP<sup>26</sup> via MAKER v3.01.03<sup>27</sup>. The homologous protein sequences were derived from published data on *Vitis*<sup>8,10</sup> species and the UniProt Swiss-Prot database (<https://www.uniprot.org/downloads>). The homologous protein dataset was filtered using CD-HIT v4.8.1<sup>28</sup> with default parameters to remove redundant sequences. The transcriptome data were obtained from published sources and derived from the same cultivar<sup>8</sup>. Transcriptome data were mapped to the haplotype-resolved assembly using HISAT2 v2.2.1<sup>29</sup>, and subsequently processed with StringTie v2.2.1<sup>30</sup> for genome-guided transcript assembly. The trained gene prediction models, homologous protein sequences and transcriptome evidence were integrated into MAKER to predict gene structures. The final gene model was predicted by the MAKER pipeline with protein-coding genes having an AED < 0.5. A total of 36,456 and 35,471 protein-coding genes were depicted for CS-Hap1 and CS-Hap2, respectively (Table 4).

**Functional annotation of protein-coding genes.** The predicted protein-coding genes were functionally annotated utilizing a robust multi-database approach to ensure comprehensive annotation, including Non-Redundant (NR), Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (eggNOG), Protein Families Database (Pfam), Swiss-Prot protein database, Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases. The eggNOG database was engaged via eggNOG-mapper (<http://eggno-mapper.embl.de/>). The non-redundant (NR) and Swiss-Prot protein databases were employed for functional annotation using Diamond BLASTP v2.1.6<sup>31</sup>. A total of 34,142 and 32,648 non-redundant protein-coding genes were annotated using the various databases for CS-T2T.Hap1 and CS-T2T.Hap2, respectively (Table 4).

**Genome synteny analysis.** To compare published *Vitis* genomes<sup>8,10</sup> with the assembled genome, whole-genome alignments were generated using minimap2 v2.28<sup>18</sup>, followed by sorting and indexing the alignment BAM files with SAMtools v1.20<sup>32</sup>. Genomic synteny and structural variants were identified by SyRI<sup>33</sup>, and the results visualized using Plotsr (<https://github.com/schneebergerlab/plotsr>). Notably, within the two ‘Cabernet Sauvignon’ haplotype genomes, we discovered megabase-scale inversions, which are not prevalent in other haplotype-resolved *V. vinifera* genomes (Fig. 2a). To validate the chromosomal inversions, we compared the two haplotype genomes of ‘Cabernet Sauvignon’ using minimap2 v2.28<sup>18</sup> and visualizing by D-genies<sup>34</sup>. For example, a chromosomal inversion of approximately 4 Mb was detected through comparative analysis of chromosome



**Fig. 2** Genome synteny and structural variants analysis. (a) The collinearity and structural variants among six genomes. VHP: ‘Chardonnay’. PN: ‘PN40024’. CS: ‘Cabernet Sauvignon’. (b) Example of an inversion event on chromosome 19 in two haplotypes validated by Hi-C data. Up, dot plot showing genome alignment. Down, Hi-C contact heatmap.

19 between CS-T2T.Hap1 and CS-T2T.Hap2 (Fig. 2b). Furthermore, chromosomal inversions were validated using Hi-C interaction maps, which confirm a local inversion in this region (Fig. 2b). Minimap2 v2.28<sup>18</sup> was employed to perform sequence alignment between CS haplotype 1 and 2 (Supplementary Table S2), and major chromosomal inversion breakpoints on chromosome 03 and 19 were identified (Supplementary Tables S3–S4). Subsequently, HiFi and ONT reads were mapped to the two haplotypes to validate the breakpoints, and the results were visualized using the Integrative Genomics Viewer (IGV) (Figures S2–S5, see Supplementary Information).

The resulting figures indicate that approximately half of the reads span the breakpoints intact, providing strong support for the inversion events. To further verify the existence of genomic inversion events in ‘Cabernet Sauvignon’, we additionally performed a genome-wide synteny comparison between the two published ‘Cabernet Sauvignon’ haplotype genomes ([https://www.grapegenomics.com/pages/VvCabSauv/VvCabSauv08\\_v2.0/](https://www.grapegenomics.com/pages/VvCabSauv/VvCabSauv08_v2.0/)), which revealed multiple megabase-scale inversion events (Figure S1, see Supplementary Information). These results indicate that inversion events are prevalent in ‘Cabernet Sauvignon’ grapes and are not restricted to the plant material analyzed in this study.

## Data Records

The raw genomic sequencing data and genome assemblies have been deposited in the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) (<https://www.ebi.ac.uk/>) under the study number of PRJEB94081<sup>35</sup>. The accession numbers of PacBio HiFi and ONT sequencing data are publicly accessible as ERR15313518<sup>36</sup> and ERR15314151<sup>37</sup>, respectively. The accession numbers of genome assemblies are publicly accessible as GCA\_965644325<sup>38</sup> and GCA\_965644335<sup>39</sup>, respectively. The annotation files have been deposited in Zenodo (<https://doi.org/10.5281/zenodo.16300203>).

## Technical Validations

**Genome assembly quality assessment.** To evaluate the completeness of the ‘Cabernet Sauvignon’ haplotype-resolved T2T gap-free genome, we employed the Benchmarking Universal Single-Copy Orthologous (BUSCO) (v5.4.2)<sup>40</sup> with the eudicots\_odb10 database. The evaluation of genome completeness demonstrated BUSCO scores of 98.3% and 98.2% for CS-T2T.Hap1 and CS-T2T.Hap2, respectively (Table 2). The ONT and HiFi reads were aligned to genomes with minimap2 v2.28 (-ax map-ont, -ax map-pb)<sup>18</sup>. The mapping rate, mean sequencing depth, and genome coverage were calculated via SAMtools v1.20<sup>31</sup> from BAM files generated by minimap2. The haplotype-resolved genome assemblies showed high mapping ratio: HiFi reads reached 99.99%, and ONT reads achieved 100% mapping to both haplotypes (Table 2), reflecting superior assembly continuity. Long terminal repeats (LTRs) were predicted using LTR\_FINDER v1.1<sup>41</sup>, followed by classification and filtering using LTR\_retriever v2.9.0<sup>42</sup> to calculate the LTR Assembly Index (LAI), reflecting genome assembly continuity. LAI scores of 16.46 (CS-T2T.Hap1) and 18.25 (CS-T2T.Hap2) indicated high-quality haplotype-resolved assemblies (Table 2). The k-mer-based quality of the haplotype-resolved assemblies was assessed using Meryl v1.3 and Merqury v1.3<sup>43</sup>. The results revealed quality values (QV) of 49.13 for CS-T2T.Hap1 and 49.83 for CS-T2T.Hap2 (Table 2), indicating high accuracy in both haplotypes. Collectively, these results demonstrate the high accuracy and completeness of the T2T haplotype-resolved genome assembly for the grapevine cultivar ‘Cabernet Sauvignon’.

**Gene annotation assessment.** Functional validation of annotated genes was conducted through alignment with different public protein databases. For the two haplotype genomes, revealed that the following proportions of proteins were annotated across databases: NR (92.82%), EggNOG (77.38%), Pfam (69.56%), Swiss-Prot (64.78%), KEGG (37.53%), and GO (35.60%). Overall, 93.65% (Hap1) and 92.04% (Hap2) non-redundant genes were mapped to public protein databases, confirming the comprehensive of the gene annotations.

## Data availability

The accession numbers of PacBio HiFi and ONT sequencing data are publicly accessible as ERR15313518 and ERR15314151 respectively. The accession numbers of genome assemblies are publicly accessible as GCA\_965644325 and GCA\_965644335 respectively. The annotation files have been deposited in Zenodo (<https://doi.org/10.5281/zenodo.16300203>).

## Code availability

All software used in this work are publicly available, and the software versions and parameters used are described in the Methods section. The parameters not mentioned in the analysis were used as default parameters suggested by the developer. No specific script was developed in this study.

Received: 28 July 2025; Accepted: 17 February 2026;

Published online: 26 February 2026

## References

1. Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D. & Gaut, B. S. Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 11715–11720, <https://doi.org/10.1073/pnas.1709257114> (2017).
2. Ren, Y. *et al.* Assessment of ‘Cabernet Sauvignon’ Grape Quality Half-Véraison to Maturity for Grapevines Grown in Different Regions. *International Journal of Molecular Sciences* **24**, 4670, <https://doi.org/10.3390/ijms24054670> (2023).
3. Bowers, J. E. & Meredith, C. P. The parentage of a classic wine grape, Cabernet Sauvignon. *Nature Genetics* **16**, 84–7, <https://doi.org/10.1038/ng0597-84> (1997).
4. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–7, <https://doi.org/10.1038/nature06148> (2007).
5. Velt, A. *et al.* An improved reference of the grapevine genome reasserts the origin of the PN40024 highly homozygous genotype. *G3 Genes|Genomes|Genetics* **13**, jkad067, <https://doi.org/10.1093/g3journal/jkad067> (2023).
6. Cochetel, N. *et al.* A super-pangenome of the North American wild grape species. *Genome Biology* **24**, 290, <https://doi.org/10.1186/s13059-023-03133-2> (2023).
7. Wang, Y., Ding, K., Li, H., Kuang, Y. & Liang, Z. Biography of *Vitis* genomics: recent advances and prospective. *Horticulture Research* **11**, uhae128, <https://doi.org/10.1093/hr/uhae128> (2024).

8. Guo, L. *et al.* Super pangenome of *Vitis* empowers identification of downy mildew resistance genes for grapevine improvement. *Nature Genetics* **57**, 741–753, <https://doi.org/10.1038/s41588-025-02111-7> (2025).
9. Liu, Z. *et al.* Grapevine pangenome facilitates trait genetics and genomic breeding. *Nature Genetics* **56**, 2804–2814, <https://doi.org/10.1038/s41588-025-02111-7> (2024).
10. Shi, X. *et al.* The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Horticulture Research* **10**, <https://doi.org/10.1093/hr/uhad061> (2023).
11. Wang, X. *et al.* Telomere-to-telomere and gap-free genome assembly of a susceptible grapevine species (Thompson Seedless) to facilitate grape functional genomics. *Horticulture Research* **11**, uhad260, <https://doi.org/10.1093/hr/uhad260> (2023).
12. Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**, 1050–1054, <https://doi.org/10.1038/nmeth.4035> (2016).
13. Massonnet, M. *et al.* The genetic basis of sex determination in grapes. *Nature Communications* **11**, 2902, <https://doi.org/10.1038/s41467-020-16700-z> (2020).
14. Schenk, J. J. *et al.* What is the “modified” CTAB protocol? Characterizing modifications to the CTAB DNA extraction protocol. *Applications in Plant Sciences* **11**, 3, <https://doi.org/10.1002/aps3.11517> (2023).
15. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, <https://doi.org/10.1093/bioinformatics/btr011> (2011).
16. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, <https://doi.org/10.1093/bioinformatics/btx153> (2017).
17. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175, <https://doi.org/10.1038/s41592-020-01056-5> (2021).
18. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, <https://doi.org/10.1093/bioinformatics/bty191> (2018).
19. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell systems* **3**, 95–8, <https://doi.org/10.1016/j.cels.2016.07.002> (2016).
20. Dudchenko, O. *et al.* *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, <https://doi.org/10.1126/science.aal3327> (2017).
21. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell systems* **3**, 99–101, <https://doi.org/10.1016/j.cels.2015.07.012> (2016).
22. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–80, <https://doi.org/10.1093/nar/27.2.573> (1999).
23. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology* **17**, 132, <https://doi.org/10.1186/s13059-016-0997-x> (2016).
24. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* **20**, 275, <https://doi.org/10.1186/s13059-019-1905-y> (2019).
25. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* **3**, lqaa108, <https://doi.org/10.1093/nargab/lqaa108> (2021).
26. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–9, <https://doi.org/10.1093/bioinformatics/btn564> (2008).
27. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **18**, 188–96, <https://doi.org/10.1101/gr.6743907> (2008).
28. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–2, <https://doi.org/10.1093/bioinformatics/bts565> (2012).
29. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–60, <https://doi.org/10.1038/nmeth.3317> (2015).
30. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290–5, <https://doi.org/10.1038/nbt.3122> (2015).
31. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366–368, <https://doi.org/10.1038/s41592-021-01101-x> (2021).
32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9, <https://doi.org/10.1093/bioinformatics/btp352> (2009).
33. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* **20**, 277, <https://doi.org/10.1186/s13059-019-1911-0> (2019).
34. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *Peer Journal* **6**, e4958, <https://doi.org/10.7717/peerj.4958> (2018).
35. EMBL-EBI ENA <https://www.ebi.ac.uk/ena/browser/view/PRJEB94081> (2025).
36. EMBL-EBI ENA <https://www.ebi.ac.uk/ena/browser/view/ERR15313518> (2025).
37. EMBL-EBI ENA <https://www.ebi.ac.uk/ena/browser/view/ERR15314151> (2025).
38. EMBL-EBI ENA [https://www.ebi.ac.uk/ena/browser/view/GCA\\_965644325](https://www.ebi.ac.uk/ena/browser/view/GCA_965644325) (2025).
39. EMBL-EBI ENA [https://www.ebi.ac.uk/ena/browser/view/GCA\\_965644335](https://www.ebi.ac.uk/ena/browser/view/GCA_965644335) (2025).
40. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols in Bioinformatics* **1**, e323, <https://doi.org/10.1002/cpz1.323> (2021).
41. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–8, <https://doi.org/10.1093/nar/gkm286> (2007).
42. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research* **46**, e126, <https://doi.org/10.1093/nar/gky730> (2018).
43. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245, <https://doi.org/10.1186/s13059-020-02134-9> (2020).

## Acknowledgements

This research was supported by the Key Research and Development Program of Shandong Province (2025LZGC027, 2024CXPT031), and Shandong Provincial Natural Science Foundation (SYS202206), Taishan Scholars Program of Shandong Province and Weifang Key Laboratory of Grapevine Improvement and Utilization, China.

## Author contributions

W.Y., L.G. and Y.H.S. conceived and supervised the project, X.Z. prepared plant materials. F.S.K. and T.S. performed bioinformatics analysis, prepared the tables, figures, uploaded the raw data and wrote the manuscript. X.W., N.K., L.G. and W.Y. revised the manuscript. All authors have read and approved the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-026-06910-3>.

**Correspondence** and requests for materials should be addressed to Y.H.S., L.G. or W.Y.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026