Genome assembly of the hybrid grapevine *Vitis* 'Chambourcin'

Sagar Patel[1,2,3], Zachary N. Harris[1,2], Jason P. Londo[4], Allison Miller[1,2*] and Anne Fennell[5*]

[1]Saint Louis University, Department of Biology, 3507 Laclede Ave, St. Louis, MO 63103, USA

[2]Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, MO 63132, USA

[3]Eastern Virginia Medical School, School of Health Professions, Norfolk, VA 23501, USA

[4]School of Integrative Plant Science, Cornell University, 630 W. North Street, Geneva, NY 14456, USA

[5]South Dakota State University, Agronomy, Horticulture and Plant Science Department and BioSNTR, Brookings, SD 57006, USA

*** Correspondence:**
Sagar Patel
sgr308@gmail.com

Allison Miller
allison.j.miller@slu.edu

Anne Fennell
anne.fennell@sdstate.edu

Abstract

Background: 'Chambourcin' is a French-American interspecific hybrid grape variety grown in the eastern and midwestern United States and used for making wine. Currently, there are few genomic resources available for hybrid grapevines like 'Chambourcin'.

Results: We assembled the genome of 'Chambourcin' using PacBio HiFi long-read sequencing and Bionano optical map sequencing. We produced an assembly for 'Chambourcin' with 27 scaffolds with an N50 length of 23.3 Mb and an estimated BUSCO completeness of 98.2%. 33,265 gene models were predicted, of which 81% (26,886) were functionally annotated using Gene Ontology and KEGG pathway analysis. We identified 16,501 common orthologs between 'Chambourcin' gene models, *V. vinifera* 'PN40024' 12X.v2, VCOST.v3, *V. riparia* 'Manitoba 37' and *V. riparia* Gloire. A total of 1,589 plant transcription factors representing 58 different gene families were identified in 'Chambourcin'. Finally, we identified 310,963 simple sequence repeats (SSRs), repeating units of 1-6 base pairs in length in the 'Chambourcin' genome assembly.

Conclusions: We present the genome assembly, genome annotation, protein sequences and coding sequences reported for 'Chambourcin'. The 'Chambourcin' genome assembly provides a valuable resource for genome comparisons, functional genomic analysis, and genome-assisted breeding research.

Keywords: Vitis 'Chambourcin', grape genome, hybrid assembly, wine, PacBio HiFi, Bionano

INTRODUCTION

Grapevines (*Vitis* spp.) represent the most economically important berry-producing plants in the world, the fruits of which are used to make wine and other beverages and are consumed as fresh or dried fruit. The European grapevine *Vitis vinifera* L. ssp. *vinifera* is believed to have been domesticated approximately 8,000 years ago from wild populations of *V. vinifera* ssp. *sylvestris* growing in western Asia and eastern Europe (Myles et al. 2011). Grapevine growing (viticulture) spread rapidly through Europe and the Middle East, and eventually was introduced into North America as early as the mid 1700's and likely earlier (Pinney 1989). In addition to the introduced *V. vinifera*, North America is home to at least 20 different native *Vitis* species. Although European settlers in North America cultivated native North American *Vitis* spp.; today, few native North American grapevine species are used to make wine (e.g., *Vitis labrusca*)., Despite this, many native North American *Vitis* species have become a critical resources for viticulture through the development of disease resistant rootstocks (the below-ground portion of grafted vines) and hybrid scions (the above-ground portion of grafted vines) derived from interspecific hybridization between wild North American *Vitis* species and cultivated European *V. vinifera.* Hybrid derivatives of crosses between North American and European grapevine species make up a significant portion of the grapevines grown in eastern and midwestern North America, and hybrid rootstocks are used throughout most grape growing regions in the world.

Vitis sp. 'Chambourcin' ('Chambourcin' from here forward) is a hybrid wine grape variety derived from crosses between North American and European *Vitis* species. 'Chambourcin' was developed by the private breeder Joannes Seyve in France and was introduced into the USDA-ARS repository in Geneva, NY in 1985 (Foundation Plant Services). A complex hybrid, 'Chambourcin' is the product of Joannes Seyve 11369 and 'Plantet N' that includes several North American species in its background: *V. berlandieri* Planch., *V. labrusca* L., *V. lincecumii* Buckley, *V. riparia* Michx., *V. rupestris* Scheele, and *V. vinifera*. The full pedigree of 'Chambourcin' is available at https://www.vivc.de/. 'Chambourcin' produces black-skinned berries. Flavors of wine derived from 'Chambourcin' are described as black cherry, red fruit, herbaceous notes, black pepper, and chocolate (winetraveler.com). 'Chambourcin' is grown in parts of France and Australia, as well as in Colorado, Missouri, Nebraska, New Jersey, New York, Pennsylvania, and Virginia, among others.

'Chambourcin' is increasing in importance as a cultivated winegrape in the central and eastern United States and it has been used in experimental rootstock vineyards aimed at understanding rootstock effects on shoot system phenotypes (Migicovsky et al. 2019; Maimaitiyiming 2020; Awale et al. 2021; Harris et al. 2021, 2022). The goals of this study were 1) to develop a high-quality reference genome for 'Chambourcin'; 2) to identify and annotate gene models for more accurate functional genomic analysis for this disease resistant cultivar that is also the parent of new disease resistant cultivar 'Regent'. Work presented here advances understanding of hybrid grapevine genomics and will facilitate analyses of rootstock-scion interactions in 'Chambourcin' experimental vineyards.

Methods

PacBio HiFi sequencing and Bionano optical map sequencing

'Chambourcin' leaf material was obtained from a 12-year-old experimental vineyard located at the University of Missouri Southwest Research Station in Mount Vernon, Missouri, USA. For PacBio HIFI sequencing, high molecular weight (HMW) DNA was isolated using the Nucleobond Kit (Macherey-Nagel, Bethlehem, PA) as per manufacturer's protocol. Approximately 20 ug DNA was

sheared to a center of mass of 10-20 Kb in a Megaruptor 3 system, and a HiFi sequencing library was constructed following HiFi SMRTbell protocols for the Express Template Prep Kit 2.0 according to manufacturers' recommendations (Pacific Biosciences, California). The library was sequenced using Sequel binding and sequencing chemistry v2.0 in CCS mode in a Sequel II system with movie collection (file format of HiFi data) time of 30hrs. The HiFi reads were generated with the circular consensus sequencing (CCS) mode of pbtools using a minimum Predicted Accuracy of 0.990.

For Bionano data, DNA was isolated from fresh young leaf tissue using the Prep™ Plant DNA Isolation and labeled using the Bionano Prep™ DNA Labeling Kit (Direct Label and Stain (DLS)), (Bionano Genomics, San Diego CA). In total, 500 ng uHMW DNA was used for the DLS reaction. DNA was incubated in the presence of DLE-1 Enzyme, DL-Green and DLE-1 Buffer for 3:20 h at 37 °C, followed by a proteinase K digestion at 50C for 30 minutes, double cleanup of unincorporated DL-Green label. The resulting DLS sample was combined with Flow Buffer, DTT and DNA stain, mixed at slow speed in a rotator mixer for an hour and incubated overnight at 4 °C. The labeled sample was then loaded onto a Bionano flow cell in a Saphyr System for separation, imaging, and creation of digital molecules according to the manufacturer's recommendations (https://bionanogenomics.com/support-page/saphyr-system). The raw molecule set was filtered to a molecule length of 250 kb and minimum labels of nine CTTAAG labels per molecule. Bionano maps were assembled without pre-assembly using the non-haplotype parameters with no CMPR cut and without extend-split. Bionano software (Solve, Tools and Access, v1.5.1), was used for data visualization, processing and assembly of Bionano maps. The PacBio HiFi and Bionano sequencing were done at Corteva Agriscience, Johnston, Iowa -USA.

Genome size estimation

The PacBio HiFi reads and 19 nt k-mers were used to estimate genome heterozygosity using jellyfish (v2.3.0) (Marçais et al., 2011). The resulting ".histo" file was visualized with GenomeScope (Vurture et al., 2017).

Genome Assembly

PacBio HiFi assembly was generated using the HiFiasm assembler (v0.13-r308) (Cheng et al., 2021) with default parameters. To reduce the number of small, low-coverage artifactual contigs often generated by Hifiasm, the assembly was filtered to exclude less than 70,

000 bp contigs. Resulting HiFi contigs were merged with the DLS Bionano maps with Bionano Solve (v3.5.1) using the hybridscaffold.pl script to get a hybrid assembly. Each scaffold of the hybrid assembly was then checked and small overlapping contigs were curated and removed to make a contiguous sequence. This curated diploid assembly was examined to identify alternative contigs using Purge Haplotigs (v1.1.1) (Roach et al., 2018) and the final assembly (primary assembly) and haplotig assemblies were prepared. In this study, we used only the primary assembly for all downstream analysis but the haplotigs are maintained to cover the total heterozygous genome. Scaffolds were aligned to the *V. vinifera* 'PN40024' 12X.v2 (Canaguier et al., 2017) reference genome using minimap2 (v2.17) (Heng Li et al., 2018) and renamed based on longest alignment with reference genome *V. vinifera* 'PN40024' 12X.v2 chromosomes. We mapped two thousand Chambourcin rhAmpSeq marker sequences to the 'Chambourcin' genome assembly using BWA aligner (Heng Li, et al, 2009). The rhAmpSeq markers were designed to target the core *Vitis* genome and were developed from gene rich collinear regions of 10 *Vitis* genomes (Zou, C., et al., 2020) These markers aid in mapping contig on chromosomes and checking orientation.

Genome assembly assessment and Dotplot

All assemblies generated by PacBio HiFi and Bionano data were assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO) (v5.4.2) (Simão et al., 2015) with genome mode and embryophyta_odb10 dataset. The dotplot was obtained using minimap2 (v2.17) (Heng Li et al., 2018) with default parameters where the 'Chambourcin' primary assembly was considered as query and *V. vinifera* 'PN40024' 12X.v2 considered as the reference genome.

De novo gene prediction, functional annotation and orthologous genes

De novo repeats were identified with RepeatModeler2 (v2.0.2a) (Jullien Li et al., 2020) and then repeats were masked by RepeatMasker (v4.1.1) (Smit et al., 2015). Paired-end RNA-seq data for 'Chambourcin' were mapped to the masked 'Chambourcin' primary genome assembly using HISAT2 (v2.1.0) (Kim et al., 2019) with default parameters. The resulting alignments and protein annotations from the *V. vinifera* 'PN40024' 12X.v2, VCost.v3 genome annotation (Canaguier et al., 2017) were used for gene prediction using BRAKER2 (v2.1.6) (Tomáš Brůna et al., 2021) with --prg=gth --gth2traingenes --gff3 parameters. The resulting gene predictions (proteins, coding sequences and annotations) were completed separately for the 'Chambourcin' primary assembly and for the 'Chambourcin' haplotig assembly. The quality of the predicted proteins were assessed using BUSCO (v5.4.2) (Simão et al., 2015) with protein mode and embryophyta_odb10 dataset. The predicted proteins of *Vitis* 'Chambourcin' primary assembly were then functionally annotated using EggNOG-mapper (http://eggnog-mapper.embl.de/) and related Gene Ontology (GO), KEGG pathway, and other functional information. The Gene Ontology plot was developed using WEGO tool (Ye J, 2018). For Orthologous gene models analysis, the sequences of 'Chambourcin' primary gene models, *V. vinifera* PN40024 12X.v2, VCost.v3 *V. riparia* 'Manitoba 37' (Patel et al., 2020), and *V. riparia* Gloire (Girollet et al., 2020) were analyzed using OrthoVenn2 (Ling Xu et al., 2019) (https://orthovenn2.bioinfotoolkits.net/home) with default settings using E-value: 1e-5 and inflation value: 1.5.

Plant transcription factors prediction, phylogenetic tree and WRKY classification

The plant transcription factors for 'Chambourcin' primary assembly gene models and *V. vinifera* PN40024 12X.v2, VCost.v3 gene models were identified using PlantTFDB (5.0) (Jin JP et al., 2017) (http://planttfdb.gao-lab.org/). The identified transcription factors were divided into subfamilies according to their sequence relationship with *V. vinifera*. For circular phylogenetic tree and WRKY classification, WRKY sequences of 'Chambourcin' primary gene models and *V. vinifera* PN40024 12X.v2, VCost.v3 gene models retrieved from PlantTFDB (5.0) (Jin JP et al., 2017) (http://planttfdb.gao-lab.org/) and aligned using ClustalW method in MEGA7 (Kumar et al., 2016). Phylogenetic analysis was carried out using the Neighbor-Joining method with 1000 bootstrap replications and the evolutionary distances were computed using the Poisson correction method with Pairwise Deletion option. WRKY classification of 'Chambourcin' primary gene models carried out using the same method described in (Patel et al., 2020).

Synteny and Simple Sequence Repeats (SSRs)

The 'Chambourcin' masked primary genome assembly and gene models were aligned to *V. vinifera* 'PN40024' 12X.v2 (Canaguier et al., 2017) and *V. riparia* 'Gloire' (Girollet et al., 2020) genomes and gene models separately, using the promer option of the MUMmer program in SyMAP (v4.2) (C. Soderlund et al., 2010). We employed MIcroSAtellite (MISA) (Sebastian Beier et al., 2017)

(https://webblast.ipk-gatersleben.de/misa/) to find Simple Sequence Repeats (SSRs) in unmasked 'Chambourcin' primary genome assembly.

Results and Discussion

Genome Sequencing and Assembly of 'Chambourcin'

We generated a high quality and contiguous genome sequence of 'Chambourcin' using PacBio HiFi Sequencing and Bionano third-generation DNA sequencing. A total of 1,634,814 PacBio HiFi filtered reads were produced with an average length of 16,148 bp and genome coverage of 28X. The filtered Bionano data resulted in a subset of 1,243,428 molecules with a total length of 429,808,857 base pairs (bp). In total, 124 Bionano maps with a total length of 962,964 bp and an N50 of 13,725 bp were assembled, corresponding to the diploid complement. We estimated heterozygosity to be 2.28% in the 'Chambourcin' genome (Supplementary Figure 1), which is higher than estimates for heterozygosity in any of the other *Vitis* genomes sequenced to date (Canaguier et al., 2017; Girollet et al., 2020; Patel et al., 2020). Relatively higher levels of heterozygosity in the 'Chambourcin' compared to other *Vitis* species are expected given the complex interspecific pedigree of this cultivar. A GenomeScope plot of clean reads demonstrated two peaks of coverage; the first peak located at 25X coverage corresponds to the heterozygous portion of the genome, and the second peak at 52X coverage corresponds to the homozygous portion of the genome (Supplementary Figure 1).

A de novo 'Chambourcin' genome was assembled using HiFi and Bionano data. First, a contig assembly of the PacBio HiFi reads resolved the reads into 196 contigs with an N50 of 12,215,205 bp and total length of 949,347,381 bp (Supplementary Table 1). The PacBio HiFi contig assembly was then merged with Bionano maps to get an initial hybrid assembly comprising 67 scaffolds with a N50 length of 16,400,326 bp, a maximum scaffold length of 39,458,994 bp, and total scaffold lengths of 903,810,753 bp (Supplementary Table 1). After manual curation, the hybrid assembly included 64 scaffolds with an N50 length of 16,278,793 bp, maximum length of 39,458,994 bp and total length of 869,222,201 bp (Supplementary Table 1). The hybrid assembly was partitioned into a final primary assembly (501,530,462 bp) and a haplotig assembly (367,691,739) (Supplementary Table 1). The final primary assembly for 'Chambourcin' contained 27 scaffolds with an N50 length of 23,328,931 bp and a longest scaffold length of 39,458,997 bp (Table 1). The secondary haplotig assembly contained 37 haplotig scaffolds with a N50 length of 12,463,179 bp and a longest scaffold length of 28,443,568 bp (Table 1). We identified 98.2% Complete BUSCOs (C) for primary genome assembly and 73.1% Complete BUSCOs (C) for haplotig genome assembly (Table 1; Supplementary Table 2).

The 'Chambourcin' primary genome assembly was aligned to the reference genome *V. vinifera* 'PN40024' 12X.v2 (Canaguier et al., 2017) (Supplementary Table 3) and a dot-plot was generated to facilitate comparisons among genomes. Collinearity between 'Chambourcin' and *V. vinifera* 'PN40024' 12X.v2 was observed as a straight diagonal line without large gaps in the dot plot, confirming high synteny of the 'Chambourcin' genome and reference genomeq (Figure 1A). To further validate the Chambourcin' genome assembly, we mapped Chambourcin rhAmpSeq markers to the 'Chambourcin' genome assembly and found 99% of rhAmpSeq markers mapped to 'Chambourcin' scaffolds and mapped to the same chromosomes and positions that the markers were derived from in the collinear *Vitis* core genome (Supplementary Table 4).

Synteny analyses of the 'Chambourcin' primary genome assembly with *V. vinifera* 'PN40024' 12X.v2 and *V. riparia* 'Gloire' genomes were used to identify syntenic blocks (Svelitsky

et al. 2019) between species. The 'Chambourcin' primary assembly scaffolds aligned with larger syntenic blocks and covered the whole chromosomes of *V. vinifera* PN40024 12X.v2 and *V. riparia* 'Gloire' (Figure 1B). This alignment of the primary genome with *V. vinifera* PN40024 12X.v2 and *V. riparia* 'Gloire' indicated highly contiguous 'Chambourcin' scaffolds useful for a comparative genomic analysis.

Repeat Sequence Annotation

Repeated regions were binned into seven different classes: long interspersed nuclear elements (LINEs) (4.38%), long terminal repeats (LTRs) (16.63%), DNA transposons (2.10%), rolling-circles (0.50%), low complexity repeats (0.36%), simple repeats (1.24%) and unclassified repeats (31.63%) (Supplementary Table 5). The repetitive sequence content in the 'Chambourcin' primary genome assembly (56.83%) was higher than previously reported for *V. riparia* 'Manitoba 37' (46%) (Patel et al., 2020), *V. vinifera* 'PN40024' 12X.v2 (35.12%) (Canaguier et al., 2017) and *V. riparia* (33.94%) 'Gloire' (Girollet et al., 2020). Simple sequence repeats (SSRs) are tandem repeats of DNA that have been used to develop robust genetic markers. We identified 310,963 simple sequence repeats (SSRs), repeating units of 1-6 base pairs in length, in the 'Chambourcin' primary genome assembly (Figure 2; Supplementary Table 6).

Gene Annotation and orthologous genes

A total of 33,265 gene models were predicted for the 'Chambourcin' primary genome assembly. We identified 95.7% complete BUSCOs (C); of these 85.8% were designated single-copy BUSCOs (S) and 9.9% were designated duplicated BUSCOs (D) (Supplementary Table 7). As evidenced by the high number of complete single-copy genes identified, the BUSCO results indicate that the 'Chambourcin' primary genome assembly offers comprehensive coverage of expected gene space. Functional annotation of the 'Chambourcin' primary gene models (33,265) was carried out using the EggNOG database (http://eggnog-mapper.embl.de/) (Supplementary Table 8). In total 26,886 'Chambourcin' primary gene models were annotated and 86% (23,026) of the 'Chambourcin' primary gene models annotated with *V. vinifera* V1 gene models (Supplementary Table 8). Out of the total 26,886 'Chambourcin' annotated primary proteins, 13,285 gene models were identified with Gene Ontology (GO) accessions (Supplementary Table 8) and further classified into three sub-ontologies: biological process (11,385), cellular component (11,455) and molecular function (9,903) (Supplementary Figure 2). A total of 8,319 'Chambourcin' primary gene models were annotated with KEGG pathway IDs (Supplementary Table 8). Using OrthoVenn2 (https://orthovenn2.bioinfotoolkits.net/home), we identified a total of 16,501 common orthologs between 'Chambourcin' primary gene models, *V. vinifera* PN40024 12X.v2 annotation , *V. riparia* 'Manitoba 37' and *V. riparia* 'Gloire' (Supplementary Figure 3A). In total, 17,105 orthologous gene models were found between the 'Chambourcin' and *V. riparia* 'Manitoba 37' and *V. riparia* 'Gloire' (Supplementary Figure 3B), while 19,387 gene models were orthologous with *V. vinifera* PN40024 12X.v2 VCost.v3 gene models (Supplementary Figure 3C).

Plant transcription factors and Chambourcin WRKY transcription factor classification

Using the Plant Transcription Factor Database (PlantTFDB 5.0 - http://planttfdb.gao-lab.org/), a total of 1,589 plant transcription factors representing 58 different gene families were identified from 'Chambourcin' (Supplementary Table 9). A similar number of transcription factors were identified for the AP2, NAC, RAV and WRKY gene families as found in *V. vinifera* 'PN40024' 12X.v2, VCost.v3 (Supplementary Table 9). There were 66 WRKY sequences identified in 'Chambourcin' and 62 in *V. vinifera* PN40024 12X.v2, VCost.v3 (Table 2; Supplementary Table 9). WRKY

transcription factors regulate many processes in plants and algae, such as the responses to biotic and abiotic stresses and seed dormancy. The Chambourcin WRKY subfamily classification was similar to *V. vinifera* 'PN40024' 12X.v2 and *V. riparia* 'Manitoba 37' (Figure 3; Table 2). These results show the high coverage of 'Chambourcin' gene models.

Conclusion

In this study, we present the first genome assembly of a complex interspecific hybrid grape cultivar, 'Chambourcin', using PacBio HiFi long read sequencing and Bionano third-generation sequencing data. The comparative genomic analyses of 'Chambourcin' with the reference genome of *V. vinifera* 'PN40024' 12X.v2 and *V. riparia* 'Gloire' indicated that the 'Chambourcin' genome aligns well to other grape genomes without any large structural variation. Ortholog analyses of 'Chambourcin' primary gene models, *V. vinifera* 'PN40024' 12X.v2, VCost.v3 , *V. riparia* 'Manitoba 37' and *V. riparia* 'Gloire' indicated that the 'Chambourcin' genome assembly and gene annotations are a high-quality grapevine resource for the research community.

Interspecific hybrids derived from two or more *Vitis* species are common in nature (Morales-Crus et al. 2021) and are the cornerstone rootstocks grown worldwide, cultivars cultivated in eastern and midwestern North America, and new disease resistant genotypes currently in development (Migicovksy et al. 2016). The sequence data, scaffold assemblies, and gene annotations the 'Chambourcin' genome assembly described here provide a valuable resource for genome comparisons, functional genomic analysis, and genome-assisted breeding research.

Data Availability

The PacBio HiFi reads deposited in the NCBI BioProject accession PRJNA754438. The Bionano maps, genome assembly, gene annotation, proteins and other data are available at figshare: 10.6084/m9.figshare.15505788

Private link for reviewers: https://figshare.com/s/7b04a3ad51fb1646fef2

Additional files

Supplementary Figure 1. GenomeScope plot estimating heterozygosity of the *Vitis* 'Chambourcin'.

Supplementary Figure 2. Gene Ontology plot of the 'Chambourcin' primary proteins.

Supplementary Figure 3. Venn diagram of 'Chambourcin' primary proteins with other grapevine species.

Supplementary Table 1. 'Chambourcin' genome assembly descriptive statistics.

Supplementary Table 2. BUSCO results of the 'Chambourcin' genome assembly.

Supplementary Table 3: Alignment of 'Chambourcin' primary genome assembly with the reference genome *V. vinifera* 'PN40024' 12X.v2.

Supplementary Table 4. Mapping of rhAmpSeq with 'Chambourcin' genome assembly.

Supplementary Table 5. Repetitive sequences in the 'Chambourcin' genome assembly.

Supplementary Table 6. The Simple Sequence Repeats (SSRs) of the 'Chambourcin' primary genome assembly.

Supplementary Table 7. BUSCO results for protein sequences of the 'Chambourcin' and other grapevine species.

Supplementary Table 8. Functional annotation of the 'Chambourcin' primary proteins.

Supplementary Table 9. The plant transcription factors of the 'Chambourcin' primary and *V. vinifera* 'PN40024' 12X.v2, VCost.v3 proteins.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This project was funded by NSF Plant Genome Research Program 1546869 to A.J.M., A.F. and J.P.L.

Author Contributions

S.P., Z.H., A.F. and A.M. conceived of and designed this study, A.F. provided computational resources and guidance for this study and J.P.L. provided paired-end RNA-Seq data for gene prediction and 'Chambourcin' rhAmpseq marker haplotypes sequences. S.P. processed DNA sequences for 'Chambourcin', assembled the genome and conducted synteny analysis. S.P. processed RNASeq data for gene prediction, conducted gene prediction and gene annotation. S.P. conducted comparative genomics analysis, and uploaded sequences to NCBI and figshare; S.P. wrote the first draft of the manuscript. S.P., A.M., Z.H., A.F., and J.P.L. reviewed and finalized the manuscript.

Acknowledgments

We acknowledge Laszlo Kovacs for collecting 'Chambourcin' samples for sequencing and Roberto Villegas-Diaz for assistance with installing and debugging tools in the South Dakota University High Performance Cluster.

Reference

1. Awale, Mani, Connie Liu, and Misha Kwasniewski. "A Metabolomics-Based Approach to Differentiate Volatiles between a European and Hybrid Grapes and Wines." HORTSCIENCE. Vol. 56. No. 9. 113 S WEST ST, STE 200, ALEXANDRIA, VA 22314-2851 USA: AMER SOC HORTICULTURAL SCIENCE, 2021.
2. C. Soderlund, M. Bomhoff, and W. Nelson (2010). SyMAP: A turnkey synteny system with application to plant genomes. Nucleic Acids Res 39(10):e68
3. Canaguier, A. et al. A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genomics Data 14, 56–62 (2017).

4. Cheng, H., Concepcion, G.T., Feng, X. et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18, 170–175 (2021).
5. Girollet, N. et al. De novo phased assembly of Vitis riparia grape genome. Sci. Data 6, 127 (2019).
6. Harris, Z. N., Pratt, J. E., Bhakta, N., Frawley, E., Klein, L. L., Kwasniewski, M. T., Migicovsky, Z., & Miller, A. J. (2022). Temporal and environmental factors interact with rootstock genotype to shape leaf elemental composition in grafted grapevines. Plant Direct, 6(8), e440. https://doi.org/10.1002/pld3.440
7. Heng Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics, Volume 34, Issue 18, 15 September 2018, Pages 3094–3100, https://doi.org/10.1093/bioinformatics/bty191
8. Heng Li, Richard Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics, Volume 25, Issue 14, 15 July 2009, Pages 1754–1760.
9. Jin JP, Tian F, Yang DC, Meng YQ, Kong L, Luo JC and Gao G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Research, 45(D1):D1040-D1045
10. Jullien M. Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, Arian F. Smit., RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences Apr 2020, 117 (17) 9451-9457; DOI: 10.1073/pnas.1921046117
11. Kim, D., Paggi, J.M., Park, C. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915 (2019).
12. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009 Sep;19(9):1639-45. doi: 10.1101/gr.092759.109. Epub 2009 Jun 18. PMID: 19541911; PMCID: PMC2752132.
13. Kumar S., Stecher G., and Tamura K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular Biology and Evolution 33:1870-1874.
14. Ling Xu, Zhaobin Dong, Lu Fang, Yongjiang Luo, Zhaoyuan Wei, Hailong Guo et al., OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species, Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W52–W58.
15. Maimaitiyiming, M.; Sagan, V.; Sidike, P.; Maimaitijiang, M.; Miller, A.J.; Kwasniewski, M. Leveraging Very-High Spatial Resolution Hyperspectral and Thermal UAV Imageries for Characterizing Diurnal Indicators of Grapevine Physiology. Remote Sens. 2020, 12, 3216. https://doi.org/10.3390/rs12193216
16. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770 (2011).
17. Migicovsky, Z., Sawler, J., Money, D. et al. Genomic ancestry estimation quantifies use of wild species in grape breeding. BMC Genomics 17, 478 (2016). https://doi.org/10.1186/s12864-016-2834-8
18. Morales-Cruz, A., Aguirre-Liguori, J.A., Zhou, Y. et al. Introgression among North American wild grapes (Vitis) fuels biotic and abiotic adaptation. Genome Biol 22, 254 (2021).
19. Patel, S., Robben, M., Fennell, A. et al. Draft genome of the Native American cold hardy grapevine Vitis riparia Michx. ‘Manitoba 37’. Hortic Res 7, 92 (2020).
20. Pinney T. 1989. A History of Wine in America. Volume 1. From the Beginnings to Prohibition. University of California Press, Berkeley, CA. Proceedings of the National Academy of Sciences Mar 2011, 108 (9) 3530-3535; DOI: 10.1073/pnas.1009363108
21. Roach, M.J., Schmidt, S.A. & Borneman, A.R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19, 460 (2018). https://doi.org/10.1186/s12859-018-2485-7
22. Sean Myles, Adam R. Boyko et al. Genetic structure and domestication history of the grape

23. Sebastian Beier, Thomas Thiel, Thomas Münch, Uwe Scholz, Martin Mascher, MISA-web: a web server for microsatellite prediction, Bioinformatics, Volume 33, Issue 16, 15 August 2017, Pages 2583–2585.
24. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210-2. doi: 10.1093/bioinformatics/btv351. Epub 2015 Jun 9. PMID: 26059717.
25. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org
26. Tomáš Brůna, Katharina J Hoff, Alexandre Lomsadze, Mario Stanke, Mark Borodovsky, BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database, NAR Genomics and Bioinformatics, Volume 3, Issue 1, March 2021, lqaa108.
27. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33, 2202–2204 (2017).
28. Ye J, Zhang Y, et al. Nucleic Acids Res. 2018 Jul 2;46(W1):W71-W75.
29. Zachary N Harris, Mani Awale, Niyati Bhakta, Daniel H Chitwood, Anne Fennell, Emma Frawley, Laura L Klein, Laszlo G Kovacs, Misha Kwasniewski, Jason P Londo, Qin Ma, Zoë Migicovsky, Joel F Swift, Allison J Miller, Multi-dimensional leaf phenotypes reflect root system genotype in grafted grapevine over the growing season, GigaScience, Volume 10, Issue 12, December 2021, giab087, https://doi.org/10.1093/gigascience/giab087
30. Zoë Migicovsky, Zachary N Harris, Laura L Klein, Mao Li, Adam McDermaid, Daniel H Chitwood, Anne Fennell, Laszlo G Kovacs, Misha Kwasniewski, Jason P Londo, Qin Ma, Allison J Miller, Rootstock effects on scion phenotypes in a 'Chambourcin' experimental vineyard, Horticulture Research, Volume 6, 2019, 64, https://doi.org/10.1038/s41438-019-0146-2
31. Zou, C., Karn, A., Reisch, B. et al. Haplotyping the Vitis collinear core genome with rhAmpSeq improves marker transferability in a diverse genus. Nat Commun 11, 413 (2020).

Table 1. 'Chambourcin' genome assembly and gene prediction results.

	Details	'Chambourcin' primary assembly	'Chambourcin' haplotig assembly
Genome assembly results			
	Number of scaffolds	27	37
	Total size of scaffolds	501,530,462	367,691,739
	Longest scaffold	39,458,997	28,443,568
	Shortest scaffold	5,521,010	1,571,397
	Number of scaffolds > 1M nt	27	37
	Number of scaffolds > 10M nt	22	16
	N50 scaffold length	23,328,931	12,463,179
	Scaffold %N	0.04	1.74
Gene prediction results			
	Total CDS and protein	33,283	23,764
	Total CDS bp	36,489,387	25,598,412
	Mean CDS bp	1,096.3	1,077.2
	Longest CDS bp	32,841	21,870
	Total protein bp	12,129,857	8,509,049
	Mean protein bp	364.4	358.1

	Longest protein bp	10,946	7,289
BUSCO results			
	Complete BUSCOs (C) = (S) + (D)	1,586 (98.2%)	1,180 (73.1%)
	Complete and single-copy BUSCOs (S)	1,553 (96.2%)	1,146 (71%)
	Complete and duplicated BUSCOs (D)	33 (2%)	34 (2.1%)
	Fragmented BUSCOs (F)	17 (1.1%)	23 (1.4%)
	Missing BUSCOs (M)	11 (0.7%)	411 (25.5%)
	Total BUSCOs	1,614	1,614

Table 2. WRKY transcription factor classification comparison of the 'Chambourcin' with other grape species.

		Group II						
Species	**Group I**	**II a**	**IIb**	**IIc**	**IId**	**IIe**	**Group III**	**Total**
Chambourcin	14	4	8	17	7	9	7	66
V. vinifera V3	12	3	8	17	8	8	6	62
*V. riparia**	13	3	8	19	8	9	7	67

* (Patel et al., 2020)

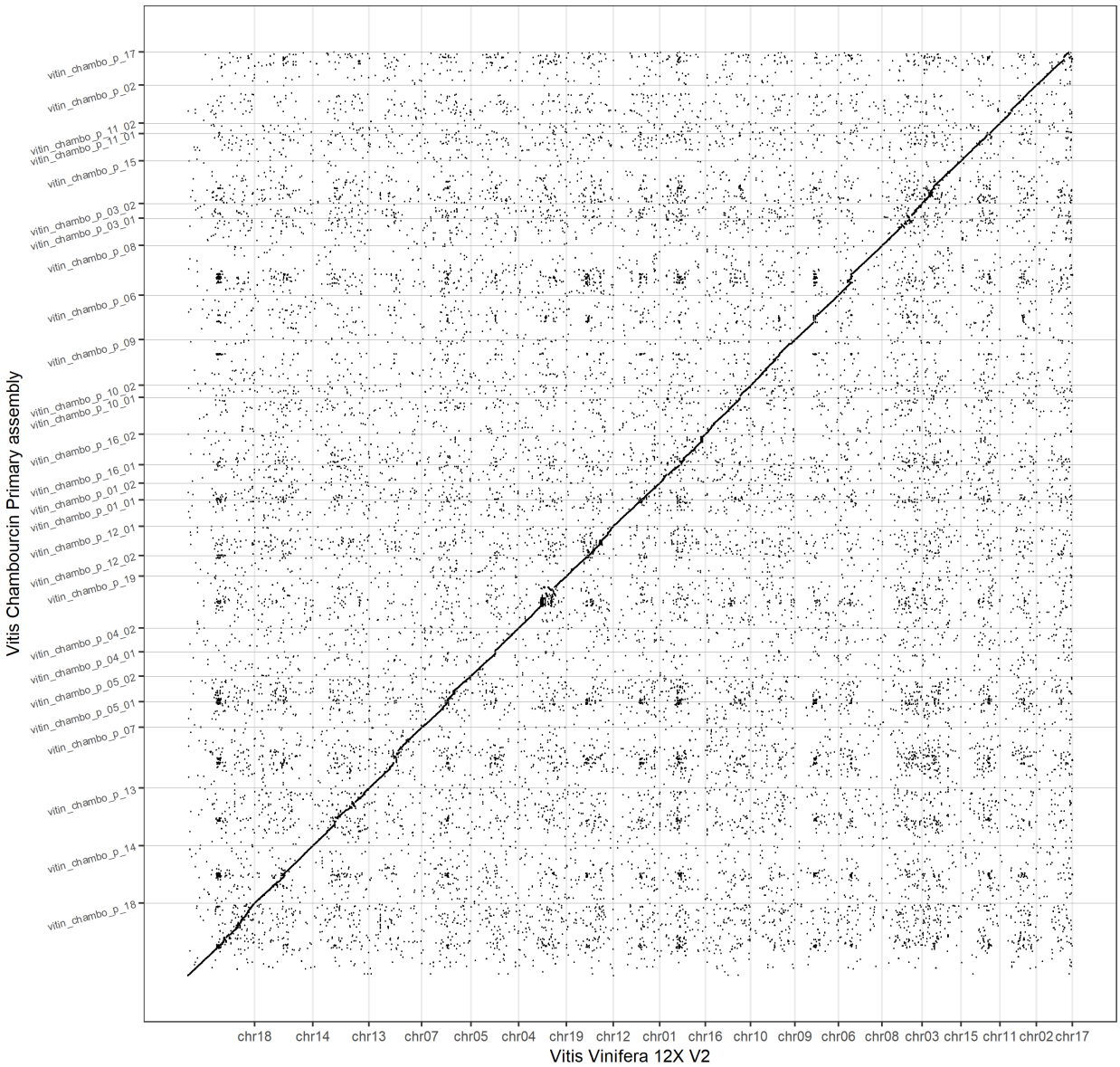
Figures

Figure 1. Comparative study of the 'Chambourcin' genome assembly. (A) Dotplot of the 'Chambourcin' primary genome assembly and *V. vinifera* 'PN40024' 12X.v2. (B) Venn diagram of Orthologous genes in the 'Chambourcin' primary proteins, *V. vinifera* PN40024 12X.v2, VCost.v3, *V. riparia* 'Manitoba 37' and *V. riparia* Gloire. (C) Orthologous genes in 'Chambourcin' primary proteins, *V. riparia* 'Manitoba 37' and *V. riparia* Gloire species, (D) Orthologous genes in 'Chambourcin' primary proteins and *V. vinifera* PN40024 12X.V3 proteins and (E) Synteny between 'Chambourcin' primary genome assembly, *V. vinifer*a PN40024 12X.v2 genome and *V. riparia* Gloire genome.

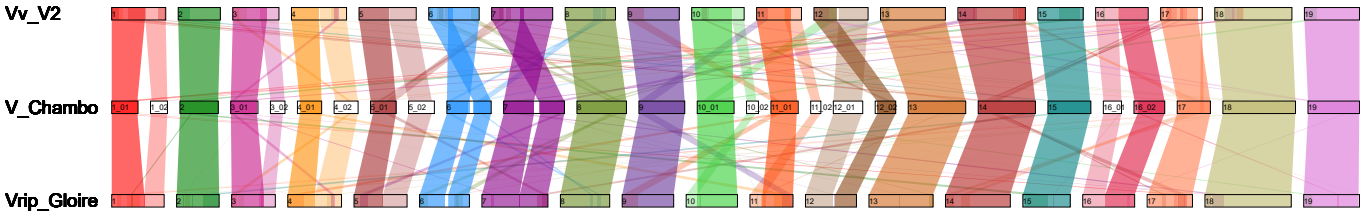
Figure 2. Circos plot. Outer ring represents all scaffolds of 'Chambourcin' primary genome assembly in different colors. The second ring of purple color represents the Simple Sequence Repeats (SSRs). The third ring of green color represents the Repetitive sequences, and the fourth ring of blue color represents gene annotations.

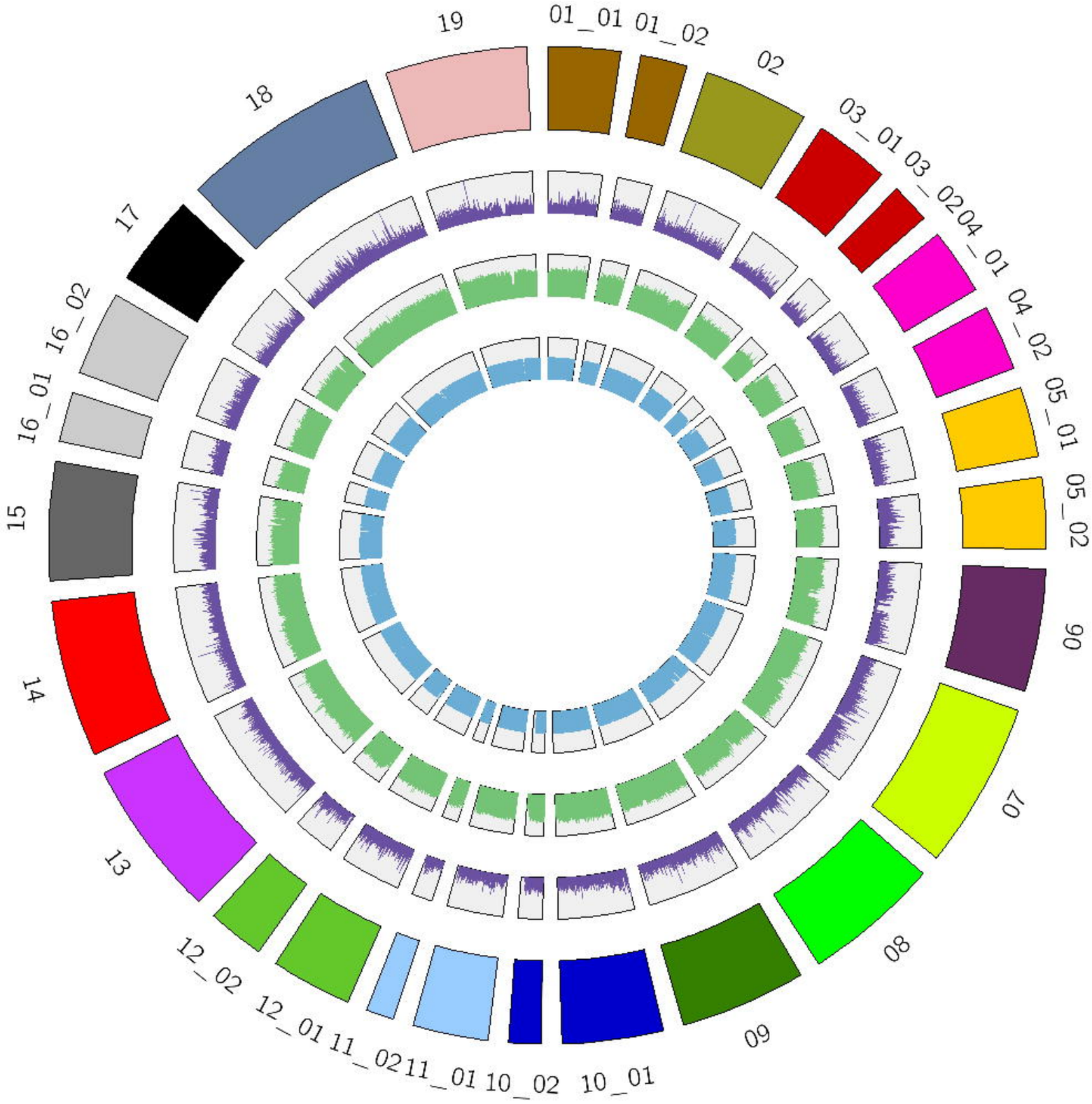
Figure 3. 'Chambourcin' and *V. vinifera* PN40024 12X.v2, VCost.v3 WRKY transcription factors. Blue dots represent 'Chambourcin' and red dots represent *V. vinifera* PN40024 12X.v2, VCost.v3. Bootstrap values displayed are at nodes.

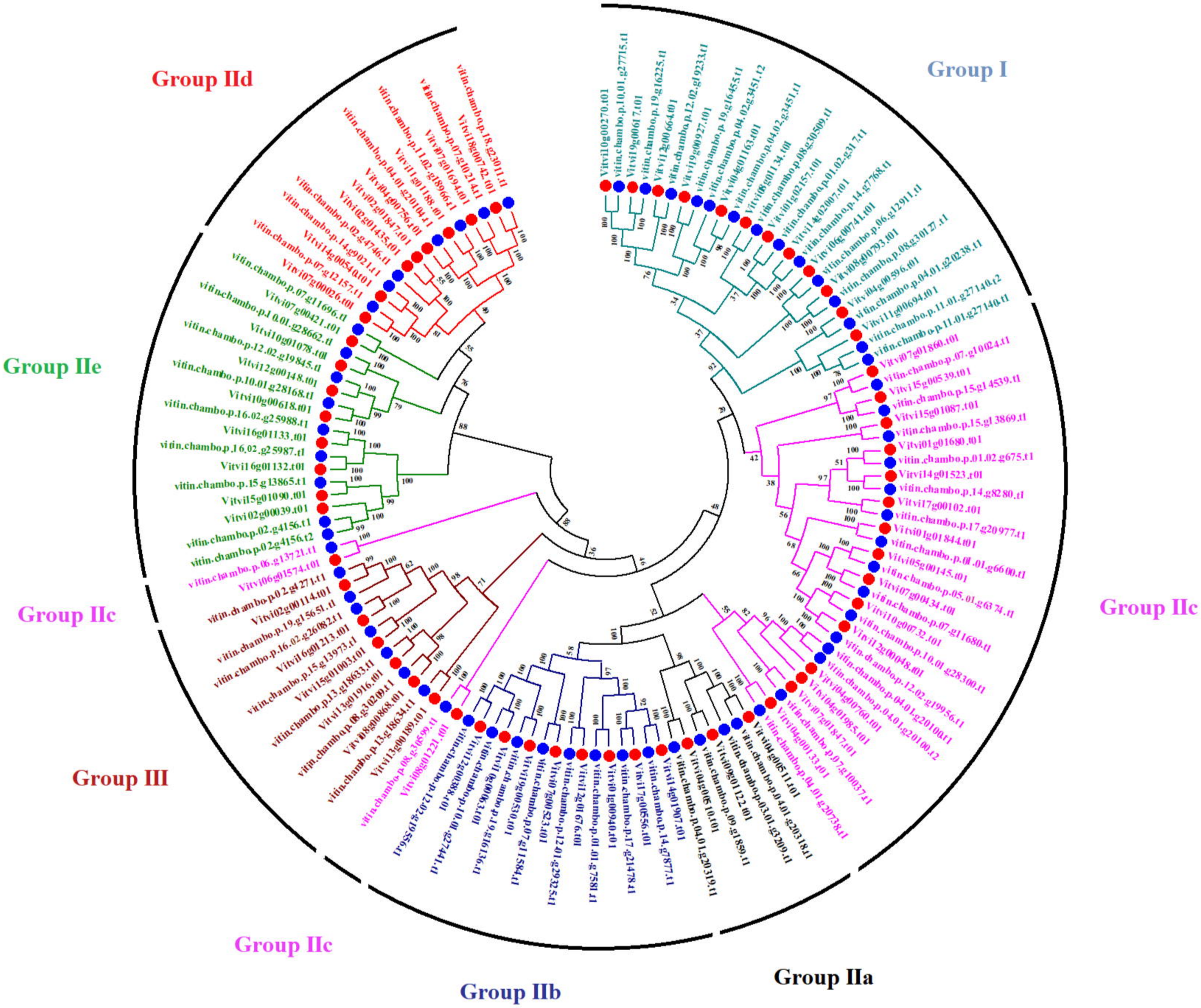
A


Vitis Chambourcin Primary assembly
vitin_chambo_p_17
vitin_chambo_p_02
vitin_chambo_p_11_02
vitin_chambo_p_11_01
vitin_chambo_p_15
vitin_chambo_p_03_02
vitin_chambo_p_03_01
vitin_chambo_p_08
vitin_chambo_p_06
vitin_chambo_p_09
vitin_chambo_p_10_02
vitin_chambo_p_10_01
vitin_chambo_p_16_02
vitin_chambo_p_16_01
vitin_chambo_p_01_02
vitin_chambo_p_01_01
vitin_chambo_p_12_01
vitin_chambo_p_12_02
vitin_chambo_p_19
vitin_chambo_p_04_02
vitin_chambo_p_04_01
vitin_chambo_p_05_02
vitin_chambo_p_05_01
vitin_chambo_p_07
vitin_chambo_p_13
vitin_chambo_p_14
vitin_chambo_p_18
chr18 chr14 chr13 chr07 chr05 chr04 chr19 chr12 chr01 chr16 chr10 chr09 chr06 chr08 chr03 chr15 chr11 chr02 chr17
Vitis Vinifera 12X V2

B


Vv_V2
V_Chambo
Vrip_Gloire


01_01
01_02
02
03_01
03_02
04_01
04_02
05_01
05_02
06
07
08
09
10_01
10_02
11_01
11_02
12_01
12_02
13
14
15
16_01
16_02
17
18
19


Group IId
Group I
Group IIe
Group IIc
Group IIc
Group III
Group IIc
Group IIb
Group IIa