Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Draft genome assembly and annotation of the Uzbek autochthonous grapevine cultivar ‘Parkent’ (*Vitis vinifera* L.)

Ilkhom Salakhutdinov
Center of Genomics and Bioinformatics
Dmitriy Mirzabaev
dmitriy08m@gmail.com

Center of Genomics and Bioinformatics
Venera Kamburova
Center of Genomics and Bioinformatics
Dilshod Usmonov
Center of Genomics and Bioinformatics
Farhod Radjapov
Center of Genomics and Bioinformatics
Shukhrat Shermatov
Center of Genomics and Bioinformatics
Zabardast Buriev
Center of Genomics and Bioinformatics
Ibrokhim Abdurakhmonov
Center of Genomics and Bioinformatics

Research Article

Keywords: Vitis vinifera L., genome assembly, genome annotation, cultivar ‘Parkent’, whole-genome sequencing, autochthonous cultivars

Posted Date: June 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9233976/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The genetic divergence between the autochthonous grapevine cultivars of Uzbekistan and European lineages represents a significant limiting factor for molecular genetic studies of Central Asian grapevines. Consequently, obtaining region-specific annotated genome assemblies is critical important for the accurate mapping of agronomically valuable traits and the development of marker-assisted selection programs. In this study, we present an accessible approach to genome assembly, along with the structural and functional annotation of the Uzbek autochthonous cultivar ‘Parkent’. The assembly was generated based on short-read sequencing data from the Illumina platform, using an initial *de novo* contig assembly to preserve cultivar-specific genomic features, followed by reference-guided scaffolding based on the telomere-to-telomere reference genome of *Vitis vinifera* L. cultivar ‘Pinot Noir 40024’. This methodology yielded a chromosome-scale draft genome assembly with a total size of 466.4 million base pairs and a high degree of completeness (98.2%), as assessed by Benchmarking Universal Single-Copy Orthologs. A total of 30,068 protein-coding genes were predicted, and key molecular-genetic features underlying the biological specificity of the cultivar ‘Parkent’ were identified, including enhanced sugar accumulation in berries and adaptation to arid climatic conditions and high solar radiation typical of the Central Asian region.

Key message

This study presents the first annotated draft genome assembly of the unique autochthonous Uzbek grapevine cultivar ‘Parkent’, establishing a critical foundation for the comparative genomics of Central Asian grapevines.

Introduction

Cultivated grapevine (*Vitis vinifera* L.) represents one of the most economically significant fruit crops in global agriculture (Grassi and Lorenzis 2021). The process of grape domestication, which began in the Middle East, resulted in the differentiation of the species into wild (subsp. *sylvestris*) and cultivated (subsp. *vinifera*) subspecies (Alleweldt 1997; Zohary et al. 2012). According to the ecological-geographical classification proposed by A. M. Negrul, cultivated grapevine varieties are classified into three main groups: *Proles occidentalis, Proles pontica*, and *Proles orientalis* (Valera et al. 2023; Tello et al. 2024).

Autochthonous grapevine cultivars of Uzbekistan (Temurov 2005) belonging to the *Proles orientalis* group are of particular scientific and practical importance. Among them, the ‘Parkent’ cultivar occupies a unique niche in regional breeding: it serves as a parent of the well-known early-mid ripening table-raisin cultivar ‘Rizamat’ (a cross between the ‘Katta-Kurgan’ and ‘Parkent’ cultivars) and is reported by local growers to act as an important fertile pollinator for the ‘Nimrang’ cultivar, a widely cultivated Central Asian table grape variety. Moreover, ‘Parkent’ is characterized by the absence of pubescence on the abaxial leaf surface, a trait determined by the *LH1* locus located on chromosome 5 (Kono et al. 2018).

This feature increases susceptibility to downy mildew caused by *Plasmopara viticola*, because it eliminates the primary constitutive barrier - the hydrophobic layer of prostrate trichomes - which normally prevents the formation of a water film required for the movement of *P. viticola* zoospores toward the stomata (Kono and Shimizu 2020). Such pronounced susceptibility to downy mildew makes this cultivar a potentially valuable model for studying the genetic mechanisms underlying physical defense responses in grapevine.

Realizing the full breeding potential of the 'Parkent' cultivar requires a comprehensive characterization of its genetic structure. However, such genomic analyses are often hindered by the substantial genetic divergence between Central Asian cultivars and the European reference genome of *V. vinifera* cultivar 'Pinot Noir' (PN40024) (Canaguier et al. 2017; Shi et al. 2023), which serves as a reference standard in grapevine genetics (Navarro-Payá et al. 2022; Velt et al. 2023). This divergence can reduce the accuracy of genomic studies, highlighting the need for region-specific annotated genome assemblies.

Currently, the prevailing approach to plant genome assembly integrates third-generation sequencing technologies (specifically high-fidelity PacBio HiFi long reads) with chromatin conformation capture (Hi-C) data (Yang et al. 2025). Although these methods ensure assembly contiguity at the chromosome-scale assembly contiguity, they entail stringent requirements for the extraction of high-molecular-weight DNA (Ova et al. 2025). Furthermore, the accurate resolution of repetitive sequences remains a complex challenge, resulting in substantial material, temporal, and computational costs (Yang et al. 2025).

Short-read sequencing platforms serve as an accessible and cost-effective alternative. Nevertheless, their application presents significant limitations: due to the short length of the generated fragments, they are unable to effectively resolve long repetitive genomic regions. This inevitably leads to high fragmentation during the *de novo* assembly of complex genomes when relying exclusively on such data. To overcome this challenge, as well as to minimize the loss of specific alleles due to reference bias, the optimal solution is a combined approach: an initial *de novo* contig assembly followed by reference-guided scaffolding. Within this strategy, the reference genome is utilized strictly as a macrostructural template for contig orientation and ordering at the chromosome level.

By implementing this strategy in practice, the present study introduces an accessible method for the genome assembly and annotation of highly heterozygous grapevine cultivars, combining an initial *de novo* contig assembly to preserve the unique genomic features of the cultivar with subsequent reference-guided scaffolding. Using the autochthonous cultivar 'Parkent' as an example, we generated a draft genome assembly and subsequently performed its structural and functional annotation. The application of the telomere-to-telomere (T2T) reference genome (PN_T2T) of cultivar 'Pinot Noir' (PN40024) (Shi et al. 2023) as a macrostructural template enabled the generation of a chromosome-level assembly of 466.4 million base pairs (Mbp) with high completeness (98.2%) as assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO), in which 30,068 protein-coding genes were predicted. The resulting assembly captures a substantial portion of the unique genetic diversity of this genotype, enabling the

identification of specific molecular-genetic features of the cultivar associated with its adaptation to arid conditions and high solar radiation levels typical of the Central Asian region.

Materials and methods

Whole-genome paired-end library preparation and sequencing

Samples of young leaf tissue of the ‘Parkent’ cultivar were collected at the Research Institute of Plant Genetic Resources of the Republic of Uzbekistan. Genomic DNA was isolated using the magnetic bead-based Plant DNA Isolation Kit (diaGene, Russia) on the Auto-Pure 4800 automated nucleic acid purification system (Allsheng, China) according to the manufacturer’s instructions. The sequencing library was prepared using the Illumina DNA PCR-Free Library Preparation Kit following the manufacturer’s protocol. Paired-end sequencing (2×150 base pairs (bp)) was performed on the Illumina NextSeq 2000 platform.

After sequencing, the data were processed using various bioinformatics tools. All computational analyses were carried out on a high-performance computing cluster at the Center of Genomics and bioinformatics of the Academy of Sciences of the Republic of Uzbekistan. The commands used in all the tools are provided in Online Resource 1.

Read quality assessment, taxonomic classification, and decontamination

Initial quality analysis of raw reads was performed using FastQC v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Fastp v0.25.0 (Chen 2025) was used to remove adapters and low-quality bases. Kraken2 v2.1.3 (Lu et al. 2022) was used to remove contamination at the read level using the National Center for Biotechnology Information (NCBI) Core Nucleotide database (core_nt; index version: October 15, 2025). Reads classified as belonging to the target genus *Vitis* (TaxID: 3603) were extracted using the extract_kraken_reads.py script from the KrakenTools package.

K-mer analysis and heterozygosity estimation

Genome size and heterozygosity were estimated by analyzing *k*-mer distributions using Jellyfish v2.2.10 (Marçais and Kingsford 2011) (*k* = 21, 33, 63), followed by visualization in GenomeScope v2.0 (Ranallo-Benavidez et al. 2020).

De novo genome assembly, reference-guided scaffolding, and quality assessment

De novo genome assembly was performed using SPAdes v4.2.0 (Prjibelski et al. 2020) with the following set of *k*-mers: 21, 33 and 63. Contigs shorter than 200 bp were removed from the resulting assembly. Subsequently, the assembly underwent a single round of polishing using Pilon v1.24 (Walker et al. 2014).

Additional screening for vectors and adapters was performed using FCS-adaptor (https://github.com/ncbi/fcs/wiki/FCS-adaptor-quickstart). Contigs representing organellar DNA and ribosomal DNA (rDNA) (accession numbers: chloroplast, NC_007957.1; mitochondrion, NC_012119.1; rDNA, MW583073.1) were identified via BLASTn (Altschul et al. 1990) and subsequently filtered using custom Python scripts based on the filtering logic and coverage thresholds proposed by Vats et al. (2023). Redundant haplotypes were removed using Redundans v2.0.1 (Pryszcz and Gabaldón 2016).

The obtained contigs were aligned and oriented into scaffolds using RagTag v2.1.0 (Alonge et al. 2022) against the PN_T2T reference genome (NCBI RefSeq accession: GCA_030704535.1). The resulting scaffolds underwent three additional rounds of polishing with Pilon, after which contigs shorter than 200 bp were removed from the final assembly.

A comprehensive assessment of the quality and completeness of the final assembly was conducted using the bioinformatics tools BUSCO v6.0.0 (Tegenfeldt et al. 2025) with the eudicots_odb10-2024-01-08 dataset, as well as QUAST v5.2.0 (Gurevich et al. 2013). QUAST was utilized to calculate basic contiguity metrics (N50, NG50, NA50), the total genome size, and the duplication ratio, as well as to perform a detailed comparative analysis against the PN_T2T reference genome.

Identification of repetitive elements and genome annotation

Repetitive sequences were identified *de novo* using RepeatModeler v2.0.7 (Flynn et al. 2020). Subsequent genome masking was performed with RepeatMasker v4.1.2 (Tarailo-Graovac and Chen 2009) using the repeat family library generated in the previous step.

Structural and functional annotation was performed using the Funannotate v1.8.17 pipeline (Palmer et al. 2020), which included cleaning, prediction, and annotation steps. During the initial cleaning stage, the minimum sequence length was set to 10,000 bp to avoid generating a large number of fragmented gene models and false-positive predictions.

The prediction stage utilized *V. vinifera* expressed sequence tags (ESTs) as transcript evidence. Pipeline-specific settings for this stage included a maximum intron length of 10,000 bp, the use of cacao (*Theobroma cacao*) as the seed species, and the Embryophyta BUSCO dataset. Additionally, to enhance gene prediction accuracy, an AUGUSTUS model pre-trained directly on the assembled genome was integrated into the pipeline. This species-specific training was performed independently using the BUSCO v6.0.0 in --long mode against the eudicots_odb10 lineage dataset. The exact commands and

custom Bash scripts used for generating and configuring the training parameters are provided in Online Resource 1.

For the functional annotation of the predicted protein sequences, we utilized a comprehensive pipeline integrating multiple tools and databases. Functional features, domains, and orthologs were identified using InterProScan v5.69 (Jones et al. 2014) and eggNOG-mapper v2.1.13 (Cantalapiedra et al. 2021), leveraging the Clusters of Orthologous Genes (COG) and evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) databases. Further targeted functional assignments were performed by querying the protein sequences against UniProt (release 2025_04), the MEROPS database for peptidases v12.5, the Pfam protein family's database v38.0, and the dbCAN database v14.0 for carbohydrate-active enzymes (CAZymes).

Graphical representation of results

For the final visualization of the genomic metrics derived from the QUAST analysis, the RStudio v. 4.5.1 environment and the R package circlize were utilized. Statistical processing and graphical representation of the functional profile across Gene Ontology (GO) categories were performed using the dplyr and ggplot2 libraries.

Results

Sequencing data processing and k-mer based genome profiling

Whole-genome sequencing generated a total of 24.4 billion base pairs (Gbp) of raw reads, which was reduced to 19.6 Gbp of high-quality reads following adapter trimming with fastp and decontamination (Online Resource 2). These filtered reads were subsequently used to estimate the haploid genome length, heterozygosity level, and repeat content utilizing GenomeScope 2.0.

To determine the most accurate parameters, the evaluation was performed across multiple *k*-mer sizes (*k* = 21, 33, 63). The 21-mer analysis yielded a model fit of 65.35–98.44% with a read error rate of 0.175%. Increasing the *k*-mer size to 33 improved the model fit to 74.03–98.62% (error rate 0.173%). Optimal results were achieved using the 63-mer, which demonstrated the highest model fit (82.12–99.43%) and the lowest sequencing error rate (0.164%) (Fig. 1, Online Resource 3).

Based on the optimal 63-mer profile, the haploid genome length of the 'Parkent' cultivar was estimated to range from 452.7 to 474.4 Mbp. The maximum size of unique sequences was 339.4 Mbp (approximately 71.5%), while repetitive elements accounted for 135.0 Mbp (approximately 28.5%). The genome heterozygosity level was estimated at 1.06–1.32%. Considering the identified structural features and the obtained k-mer distribution profiles, the combination of *k* = 21, 33, and 63 was selected as optimal for the initial *de novo* genome assembly using the SPAdes algorithm.

De novo genome assembly, reference-guided scaffolding, and quality metrics

Following the initial *de novo* genome assembly using the SPAdes algorithm and the *k*-mers selected in the previous step, the total assembly length was 449.4 Mbp, distributed across 404,401 contigs. This initial assembly exhibited low contiguity: the N50 metric was 3,209 bp, and the maximum contig length reached 144,669 bp. Evaluation of the gene space completeness using BUSCO revealed 58.6% complete orthologous groups (with a low duplication rate of 1.0%), alongside a high proportion of fragmented (22.7%) and missing (18.7%) genes (Online Resource 4).

Based on the filtering results using the FCS-adaptor tool and BLASTn analysis, 128 contigs classified as external contaminants were identified and removed. Furthermore, 1,853 contigs corresponding to organellar sequences (chloroplasts and mitochondrial DNA) and rDNA were excluded from the assembly.

In the next stage, to overcome the initial fragmentation of the *de novo* assembly, scaffolding was performed in the RagTag using the PN_T2T reference genome, followed by three rounds of polishing in Pilon to correct local errors. This approach yielded a final assembly with a total size of 466.4 Mbp and led to a substantial increase in its contiguity: the N50 metric reached 17.76 Mbp, indicating a chromosome-scale assembly (Online Resource 5). The calculated duplication ratio was 1.13.

Moreover, an independent assessment of the gene space completeness BUSCO revealed 98.2% complete conserved orthologs. The vast majority of these (97.4%) are present in single copy, while only 0.8% were duplicated (Online Resource 5).

When evaluating the alignment to the PN_T2T reference genome, the genome fraction covered by the ‘Parkent’ cultivar assembly was 67.9%, while the total unaligned length reached 72.9 Mbp (Online Resource 5). Within homologous chromosome regions during reference alignment, the QUAST program identified 17,220 events classified as misassemblies.

However, because the reference genome used is related to but distinct from the sequenced cultivar, these differences may encompass both true assembly errors and actual structural variations (SVs), hereafter referred to as putative SVs. Notably, 15,202 of these events were already present at the contig level. Specifically, the algorithm identified 7,456 potential relocations, 7,689 translocations, and 57 inversions (Fig. 2, Online Resource 6). The presence of this array of discrepancies resulted in a decrease in the contiguity of aligned blocks, with the N50 metric dropping from 17.76 Mbp to 4,885 bp. Alongside macrostructural changes, genomic comparison revealed high local variability, characterized by 1,194.77 mismatches—representing putative single nucleotide polymorphisms (SNPs) and 191.14 small insertions/deletions (indels) per 100,000 bp (Fig. 2, Online Resource 6).

Identification and classification of repetitive elements

Repeat analysis using RepeatMasker identified 241.3 Mbp of repetitive sequences, accounting for 51.75% of the total assembly length (Online Resource 7). Among the identified repeats in the 'Parkent' cultivar assembly, interspersed repeats predominated, covering 232.5 Mbp (49.84% of the assembly). Long Interspersed Nuclear Elements (LINEs) accounted for 3.96%, while a significant proportion of the repetitive sequences remained unclassified (27.71%). Simple repeats constituted for a minor fraction at 1.44%.

Among the characterized transposable elements, retroelements were the most abundant (20.41%), predominantly represented by Long Terminal Repeat (LTR) retrotransposons (16.46%). Specifically, Ty1/Copia and Gypsy elements were the most abundant LTR superfamilies, comprising 9.30% and 6.34% of the total assembly length, respectively.

Structural and functional annotation of 'Parkent' *genome*

During structural and functional annotation using the Funannotate pipeline, the integration of *ab initio* predictions and ESTs yielded 35,847 consensus gene models. Following the filtering of artifacts (short sequences and models overlapping with assembly gaps and transposable elements), the final high-quality set comprised 30,068 protein-coding genes and 394 tRNA genes.

The functional characterization of the predicted proteins included alignment against the UniProt and MEROPS databases (846 annotated peptidases). Additionally, 32,236 Pfam domains, 996 carbohydrate-active enzymes (dbCAN), and 53,620 orthologous group assignments (COG/eggNOG) were identified within the proteome. The most comprehensive functional profile was provided by the InterProScan tool, which assigned 239,118 functional features.

Within this classification framework, GO terms were successfully assigned to 17,987 predicted protein sequences (Online Resource 8).

In the "Molecular Function" category, the largest number of proteins exhibited protein binding (GO:0005515; 2,589 sequences) and ATP binding (GO:0005524; 1,817 sequences) activities. A significant portion of the proteome was also associated with protein kinase activity (GO:0004672; 994 proteins) and DNA binding (GO:0003677; 942 proteins).

In the "Biological Process" category, the highest number of proteins was involved in phosphorylation (GO:0006468; 1,013) and regulation of transcription (GO:0006355; 1,009), as well as transmembrane transport (GO:0055085; 628) and carbohydrate metabolism (GO:0005975; 443).

Finally, in the "Cellular Component" category, proteins localized to the nucleus (GO:0005634; 1,929), membrane structures (GO:0016020; 1,819), and the cytoplasm (GO:0005737; 1,564) predominated (Fig. 3).

Discussion

In the present study, we performed whole-genome sequencing of the Uzbek autochthonous grapevine cultivar ‘Parkent’ using short-read technology. To generate an annotated draft genome assembly of this Central Asian cultivar, a combined bioinformatics approach was employed: an initial *de novo* contig assembly followed by reference-guided scaffolding based on the PN_T2T reference genome. The application of this methodology yielded a genome assembly with a high degree of completeness, which served as the foundation for subsequent structural and functional annotation.

In the first stage, the fundamental architectural properties of the investigated genome were characterized. We analyzed the *k*-mer frequency distribution at various lengths (*k* = 21, 33, and 63). Comparative analysis demonstrated that the optimal result, characterized by the highest fit to the theoretical model and the lowest error rate, was achieved using the 63-mer. Based on this profile, the estimated haploid genome size of the ‘Parkent’ cultivar was 452.7–474.4 Mbp, which is close to the size of the PN_T2T reference genome (494.9 Mbp) and falls well within the known genome size variation range (450–500 Mbp) described for *V. vinifera* (Lodhi and Reisch 1995; González et al. 2022).

Furthermore, the estimated heterozygosity level, determined from the 63-mer profile (1.06–1.32%), is highly characteristic of cultivated *V. vinifera* varieties. Cultivated grapevines generally exhibit higher heterozygosity levels (1.29%) compared to wild forms (0.70%) (Luo et al. 2025). This is explained by the fact that most grapevine cultivars originated from crosses between genetically divergent parental lineages. Such origins, along with the continuous accumulation of somatic mutations during vegetative propagation, naturally lead to the high heterozygosity observed in their genomes (Zhou et al. 2019). Additionally, the *k*-mer analysis revealed that the proportion of unique sequences was approximately 71.5%, while repetitive elements accounted for approximately 28.5%.

The combination of such pronounced heterozygosity and the presence of repetitive elements poses a fundamental computational challenge for whole-genome assembly from short reads. The length of these fragments is insufficient to span long repeats and accurately resolve divergent alleles, inevitably leading to assembly graph collapse (Michael and VanBuren 2020). Consequently, the initial *de novo* assembly was predictably highly fragmented. To overcome this limitation, reference-guided scaffolding was performed. This approach yielded a final assembly with a total size of 466.4 Mbp and resulted in a substantial increase in contiguity: the N50 reached 17.76 Mbp, and the gene space completeness, as assessed by BUSCO, was 98.2%, indicating a pseudochromosome-level genome organization inferred from reference-guided scaffolding. The high quality of the result is further corroborated by the near equality of the N50 and NG50 metrics. This supports the high contiguity and comprehensive coverage of the assembly, minimizing the likelihood of artificial metric inflation due to unassembled regions (Gurevich et al. 2013).

Beyond genome-level completeness, a critical quality criterion for highly heterozygous genomes is the absence of false allelic duplications. During the assembly of highly heterozygous plants genomes, algorithms frequently misassemble divergent alleles of a single locus as independent contigs (Chin et al. 2016). Our obtained duplication ratio of 1.13 indicates the successful minimization of artificial repeat

multiplication and the correct merging of heterozygous regions into a single haploid consensus. This conclusion is also supported by the low duplication rate (0.8%) detected during the BUSCO assessment, and the absence of redundant haplotigs as verified using the Redundans algorithm, even at the initial contig assembly stage.

The generation of an assembly lacking redundant haplotigs confirms the high internal integrity and establishes a robust foundational baseline for subsequent comparative genomic analysis relative to the reference genome. Nevertheless, when transitioning from the evaluation of internal assembly metrics to comparative analysis against a reference genome, a methodological challenge arises regarding the classification of the identified structural differences.

According to literature sources and the alignment results of the final ‘Parkent’ genome assembly against the PN_T2T reference genome, the observed pattern (characterized by a genome fraction of 67.9% and 72.9 Mbp of unaligned sequences) demonstrates profound genetic divergence. In comparative plant genomics, the presence of such extensive unaligned sequences is considered a biological norm, reflecting the presence of cultivar-specific elements, highly variable intergenic regions, and unique patterns of transposon insertions (Bayer et al. 2020). Furthermore, within homologous chromosomal regions, the QUAST algorithm identified numerous putative SVs, classified as relocations, translocations, and inversions. Alongside these, a high degree of divergence is further evidenced by the high density of SNPs and short insertions/deletions.

Crucially, it is important to clarify that the vast majority of these structural differences had already been captured during the initial *de novo* contig assembly phase, prior to the application of reference-guided scaffolding. Although this excludes the possibility that they are strictly alignment artifacts resulting from mapping to a reference genome, their true nature requires cautious interpretation. As noted in methodological studies (Gurevich et al. 2013), in such cases, these differences may represent either artifacts of the assembly process itself or true biological SVs. In the context of comparing evolutionarily distant cultivars, a significant portion of the putative SVs may reflect the actual genetic divergence of the ‘Parkent’ cultivar from the reference genome. This assumption is corroborated by the high content of retrotransposons we detected during the genome masking stage, specifically LTR elements of the Ty1/Copia and Gypsy superfamilies. As previously demonstrated by researchers (Zhou et al. 2019; Luo et al. 2025), mobile element insertions and LTR activity (especially Copia) contribute substantially to the formation of cultivar-specific chromosomal rearrangements, acting as a key factor of grapevine genome evolution. Overall, such a substantial number of events is also consistent with data regarding the high genomic plasticity of *V. vinifera*, wherein SVs can affect up to 15% of the genome (Zhou et al. 2019).

Nevertheless, such an interpretation in the case of our assembly requires further evidence. Specifically, to reliably differentiate true biological variability from potential assembly artifacts, a comprehensive *de novo* assembly of the autochthonous cultivar utilizing long-read sequencing technologies is necessary. Furthermore, identifying an equally divergent genotype is required to establish a specialized reference genome. Fulfilling these objectives will provide a robust foundational baseline for conducting an

accurate comparative analysis of the entire diversity of unique autochthonous grapevine cultivars in Uzbekistan.

Transitioning from the analysis of the genome's structural organization to its functional regions, we identified 30,068 protein-coding genes during the subsequent annotation stage. This figure is consistent with previously published *V. vinifera* genomes and falls within the expected range, showing a minor deviation from the 30,799 protein-coding genes annotated in the PN_T2T reference genome. Furthermore, this value is comparable to the 30,434 genes described for a highly homozygous grapevine genotype (Jaillon et al. 2007), and aligns with metrics from recent high-quality assemblies, in which 29,235 and 29,139 protein-coding genes were identified for two alternative haplotypes (Ricciardi et al. 2025). Finally, by remaining below the 32,425 genes identified in the ‘Carménère’ cultivar haplotigs, our result successfully avoids gene count inflation frequently observed in highly heterozygous unphased primary assemblies, e.g., 40,684 genes (Minio et al. 2019).

In the final stage, functional annotation was performed for the 30,068 identified protein-coding genes, based on the assignment of GO terms to 17,987 predicted protein sequences. The functional distribution of these sequences underscores the high metabolic activity and well-developed cellular signaling systems of the autochthonous ‘Parkent’ cultivar. The presence of a large number of protein-coding genes associated with carbohydrate metabolism (GO:0005975) and active transmembrane transport (GO:0055085) reflects the biological specificity of the cultivar—namely, the intensive import and accumulation of sugars in the berries during ripening (Shangguan et al. 2017). Concurrently, the high representation of protein kinases (GO:0004672), proteins involved in phosphorylation processes (GO:0006468), and transcription regulators (GO:0006355) indicates a robust system of cellular signaling and abiotic stress response. As demonstrated in studies of grapevine adaptation (Cakir and Kilickaya 2015; Haider et al. 2017), processes associated with protein kinase activity and transcription regulation play a crucial role in osmotic regulation and plant survival under extreme conditions. Thus, the identified functional profile of the annotated proteome is directly linked to the physiological adaptation of this cultivar to the arid conditions and high solar insolation levels of the Central Asian region.

Conclusion

This study demonstrates the successful generation of an annotated draft genome assembly for the autochthonous ‘Parkent’ cultivar, featuring a total length of 466.4 Mbp and a BUSCO completeness of 98.2%, utilizing short-read sequencing, initial *de novo* assembly, and subsequent reference-guided scaffolding. The high completeness of the resulting assembly enabled a detailed structural and functional annotation of the genome, resulting in the prediction of 30,068 protein-coding genes and the identification of molecular mechanisms underlying the biological specificity of cultivar. In particular, the observed abundance of genes associated with carbohydrate metabolism and transmembrane transport reflects the genetic basis for traits such as intensive sugar import and accumulation in the berries during ripening. Concurrently, the high content of protein kinases, as well as proteins involved in phosphorylation processes and transcriptional regulation, indicates the presence of a robust cellular

signaling and abiotic stress response system. Collectively, these specific functional features determine the physiological adaptation of the ‘Parkent’ cultivar to the arid conditions and high levels of solar insolation typical of Central Asia environments. The successful genomic characterization of this unique genotype confirms the efficacy of the chosen approach and lays the foundation for the analysis of other autochthonous grapevine cultivars of Uzbekistan using the presented methodology.

Declarations

Ethical Approval

Not Applicable.

Competing interests

The authors declare no competing interests.

Author Contribution

All authors contributed to the study conception and design. Material preparation, DNA extraction, and sequencing were performed by D.U. and F.R. Bioinformatic analysis and the writing of the first draft of the manuscript were performed by D.M. Study supervision, manuscript review, and editing were conducted by I.S. and V.K. Funding acquisition, resource provision, and general research administration were managed by S.Sh., Z.B., and I.A.

Acknowledgement

This work was supported by the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan (Grant number: AL-11125078106).

Data Availability

The raw data has been submitted to NCBI Sequence Read Archive under accession ID PRJNA905061 and are available at the following URL: https://www.ncbi.nlm.nih.gov/sra/PRJNA905061.

References

1. Alleweldt G (1997) Genetics of Grapevine Breeding. In: Behnke HD, Lüttge U, Esser K et al (eds) Progress in Botany, vol 58. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-60458-4_19

2. Alonge M, Lebeigle L, Kirsche M et al (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol 23:258. https://doi.org/10.1186/s13059-022-02823-7
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
4. Babraham Bioinformatics (n.d (2026) FastQC: A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 11 Mar
5. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020) Plant pan-genomes are the new reference. Nat Plants 6:914–920. https://doi.org/10.1038/s41477-020-0733-0
6. Cakir B, Kilickaya O (2015) Mitogen-activated protein kinase cascades in Vitis vinifera. Front Plant Sci 6:556. https://doi.org/10.3389/fpls.2015.00556
7. Canaguier A, Grimplet J, Di Gaspero G et al (2017) A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genom Data 14:56–62. https://doi.org/10.1016/j.gdata.2017.09.002
8. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825–5829. https://doi.org/10.1093/molbev/msab293
9. Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinf 25. https://doi.org/10.1002/0471250953.bi0410s25. 4.10.1–4.10.14
10. Chen S (2025) fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. iMeta 4:e70078. https://doi.org/10.1002/imt2.70078
11. Chin CS, Peluso P, Sedlazeck FJ et al (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054. https://doi.org/10.1038/nmeth.4035
12. Flynn JM, Hubley R, Goubert C et al (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117:9451–9457. https://doi.org/10.1073/pnas.1921046117
13. González R, Vallès J, Garnatje T (2022) Genome Size Variation Assessment in Vitis vinifera L. Landraces in Ibiza and Formentera (Balearic Islands). Plants 11:1892. https://doi.org/10.3390/plants11141892
14. Grassi F, De Lorenzis G (2021) Back to the Origins: Background and Perspectives of Grapevine Domestication. Int J Mol Sci 22:4518. https://doi.org/10.3390/ijms22094518
15. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086
16. Haider MS, Kurjogi MM, Khalil-ur-Rehman M et al (2017) Insights into grapevine defense response against drought as revealed by biochemical, physiological and RNA-Seq analysis. Sci Rep 7:13134. https://doi.org/10.1038/s41598-017-13464-3

17. Jaillon O, Aury JM, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467. https://doi.org/10.1038/nature06148
18. Jones P, Binns D, Chang HY et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031
19. Kono A, Ban Y, Mitani N et al (2018) Development of SSR markers linked to QTL reducing leaf hair density and grapevine downy mildew resistance in Vitis vinifera. Mol Breed 38:138. https://doi.org/10.1007/s11032-018-0889-8
20. Kono A, Shimizu T (2020) Leaf Trichomes as an Effective Structure for Disease Resistance: The Case of Grapevine Downy Mildew. JARQ 54:293–298
21. Lodhi MA, Reisch BI (1995) Nuclear DNA content of Vitis species, cultivars, and other genera of the Vitaceae. Theor Appl Genet 90:11–16. https://doi.org/10.1007/BF00220990
22. Lu J, Rincon N, Wood DE et al (2022) Metagenome analysis using the Kraken software suite. Nat Protoc 17:2815–2839. https://doi.org/10.1038/s41596-022-00738-y
23. Luo Y, Liu Z, Jin Z et al (2025) Phased T2T genome assemblies facilitate the mining of disease-resistance genes in Vitis davidii. Hortic Res 12:uhae306. https://doi.org/10.1093/hr/uhae306
24. Marçais G, Kingsford C (2011) 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
25. Michael TP, VanBuren R (2020) Building near-complete plant genomes. Curr Opin Plant Biol 54:26–33. https://doi.org/10.1016/j.pbi.2019.12.009
26. Minio A, Massonnet M, Figueroa-Balderas R et al (2019) Diploid Genome Assembly of the Wine Grape Carménère. G3. Genes Genomes Genet 9:1331–1337. https://doi.org/10.1534/g3.119.400030
27. Navarro-Payá D, Santiago A, Orduña L et al (2022) The Grape Gene Reference Catalogue as a Standard Resource for Gene Selection and Genetic Improvement. Front Plant Sci 12:803977. https://doi.org/10.3389/fpls.2021.803977
28. NCBI (n d.) FCS adaptor quickstart. ncbi/fcs Wiki. GitHub. https://github.com/ncbi/fcs/wiki/FCS-adaptor-quickstart. Accessed 11 Mar 2026
29. Ova AÖ, Joffre E, Zandi Shafagh R et al (2025) Improved Isolation of Ultra-High-Molecular-Weight Genomic DNA Suitable for Third-Generation Sequencing. Microorganisms 13:534. https://doi.org/10.3390/microorganisms13030534
30. Palmer JM, Stajich J, Palmer JM et al (2020) Funannotate v1.8.1: Eukaryotic genome annotation. Zenodo. https://doi.org/10.5281/ZENODO.1134477
31. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A (2020) Using SPAdes De Novo Assembler. Curr Protoc Bioinf 70:e102. https://doi.org/10.1002/cpbi.102
32. Pryszcz LP, Gabaldón T (2016) Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res 44:e113. https://doi.org/10.1093/nar/gkw294

33. Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11:1432. https://doi.org/10.1038/s41467-020-14998-3
34. Ricciardi V, Minio A, Massonnet M et al (2025) The genome of Vitis vinifera cv. Mgaloblishvili reveals resistance and susceptibility factors to downy mildew in the Rpv29 and Rpv31 loci. Hortic Res 12:uhaf055. https://doi.org/10.1093/hr/uhaf055
35. Shangguan L, Mu Q, Fang X et al (2017) RNA-Sequencing Reveals Biological Networks during Table Grapevine ('Fujiminori') Fruit Development. PLoS ONE 12:e0170571. https://doi.org/10.1371/journal.pone.0170571
36. Shi X, Cao S, Wang X et al (2023) The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding. Hortic Res 10:uhad061. https://doi.org/10.1093/hr/uhad061
37. Tegenfeldt F, Kuznetsov D, Manni M et al (2025) OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. Nucleic Acids Res 53:D516–D522. https://doi.org/10.1093/nar/gkae987
38. Tello J, Todić S, Ferradás Y et al (2024) The genetic characterization of grapevines prospected in old Serbian vineyards reveals multiple relationships between traditional varieties of the Balkans. Front Plant Sci 15:1391679. https://doi.org/10.3389/fpls.2024.1391679
39. Temurov S (2005) Uzumchilik [Viticulture]. O'zbekiston milliy ensiklopediyasi, Tashkent (in Uzbek)
40. Valera J, Matilla-Seiquer G, Obón C, Alcaraz F, Rivera D (2023) Grapevine in the Ancient Upper Euphrates: Horticultural Implications of a Bayesian Morphometric Study of Archaeological Seeds. Horticulturae 9:803. https://doi.org/10.3390/horticulturae9070803
41. Vats S, Kumar V, Mandlik R et al (2023) Reference Guided De Novo Genome Assembly of Transformation Pliable Solanum lycopersicum cv. Pusa Ruby Genes 14:570. https://doi.org/10.3390/genes14030570
42. Velt A, Frommer B, Blanc S et al (2023) An improved reference of the grapevine genome reasserts the origin of the PN40024 highly homozygous genotype. G3. Genes Genomes Genet 13:jkad067. https://doi.org/10.1093/g3journal/jkad067
43. Walker BJ, Abeel T, Shea T et al (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9:e112963. https://doi.org/10.1371/journal.pone.0112963
44. Yang Y, Du W, Li Y, Lei J, Pan W (2025) Recent advances and challenges in de novo genome assembly. Genom Commun 2:e014. https://doi.org/10.48130/gcomm-0025-0015
45. Zhou Y, Minio A, Massonnet M et al (2019) The population genetics of structural variants in grapevine domestication. Nat Plants 5:965–979. https://doi.org/10.1038/s41477-019-0507-8
46. Zohary D, Hopf M, Weiss E (2012) Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin, 4th edn. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:osobl/9780199549061.001.0001

Figures

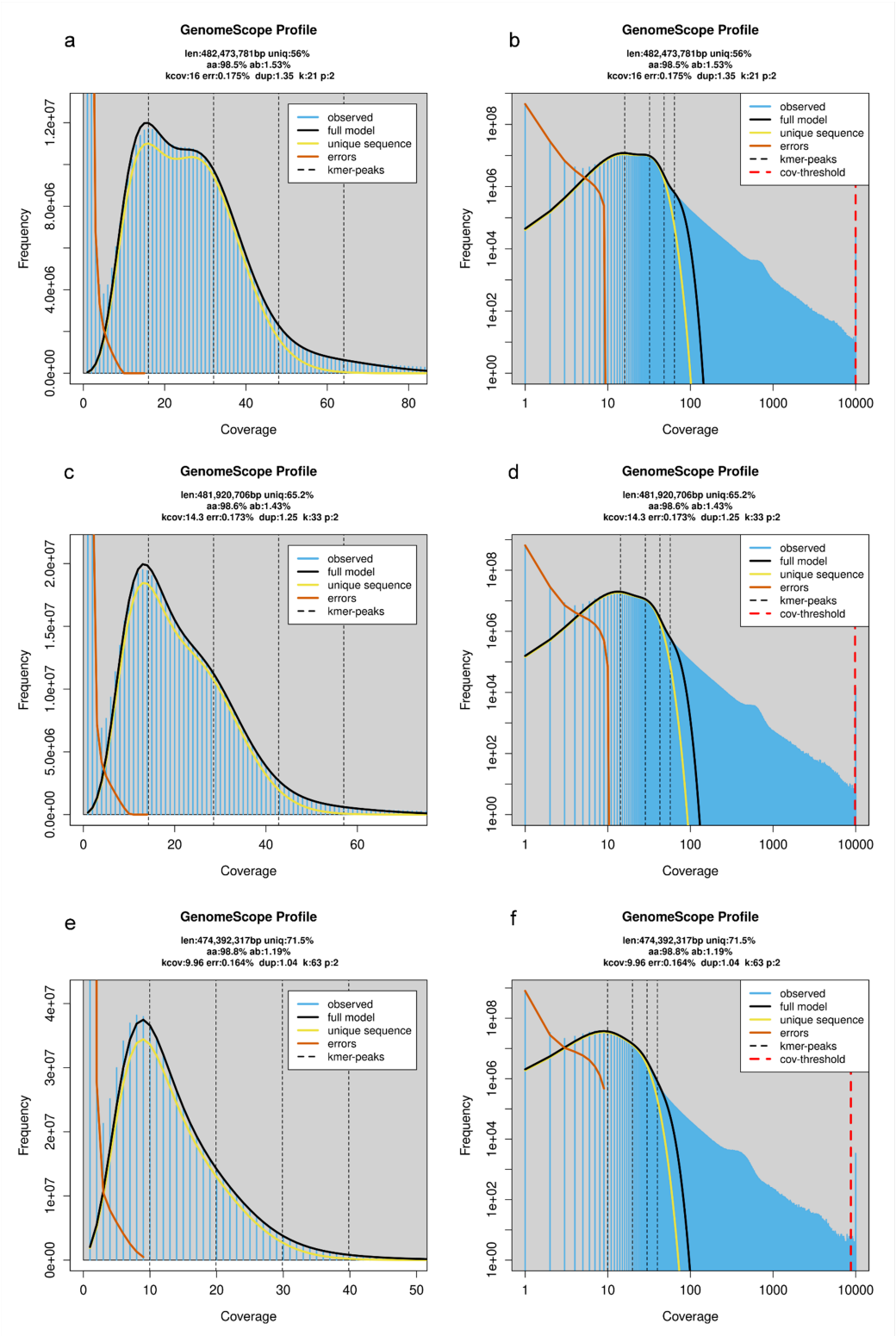


Figure 1

GenomeScope 2.0 *k*-mer frequency profiles for *Vitis vinifera* L. cultivar 'Parkent'. Linear (a, c, e) and log-scaled (b, d, f) plots correspond to *k* = 21, 33, and 63, respectively. The graphs illustrate observed *k*-mer frequencies (shaded light blue area), the full calculated model (upper solid black line), unique sequences

(middle solid yellow line), sequencing errors (lower steep solid orange line), *k*-mer peaks (vertical dashed black lines), and the maximum coverage threshold (secondary vertical dashed red line). As indicated in the plot headers, the 63-mer profiles (e, f) display the highest proportion of unique sequences (71.5%) and the lowest error rate (0.164%)

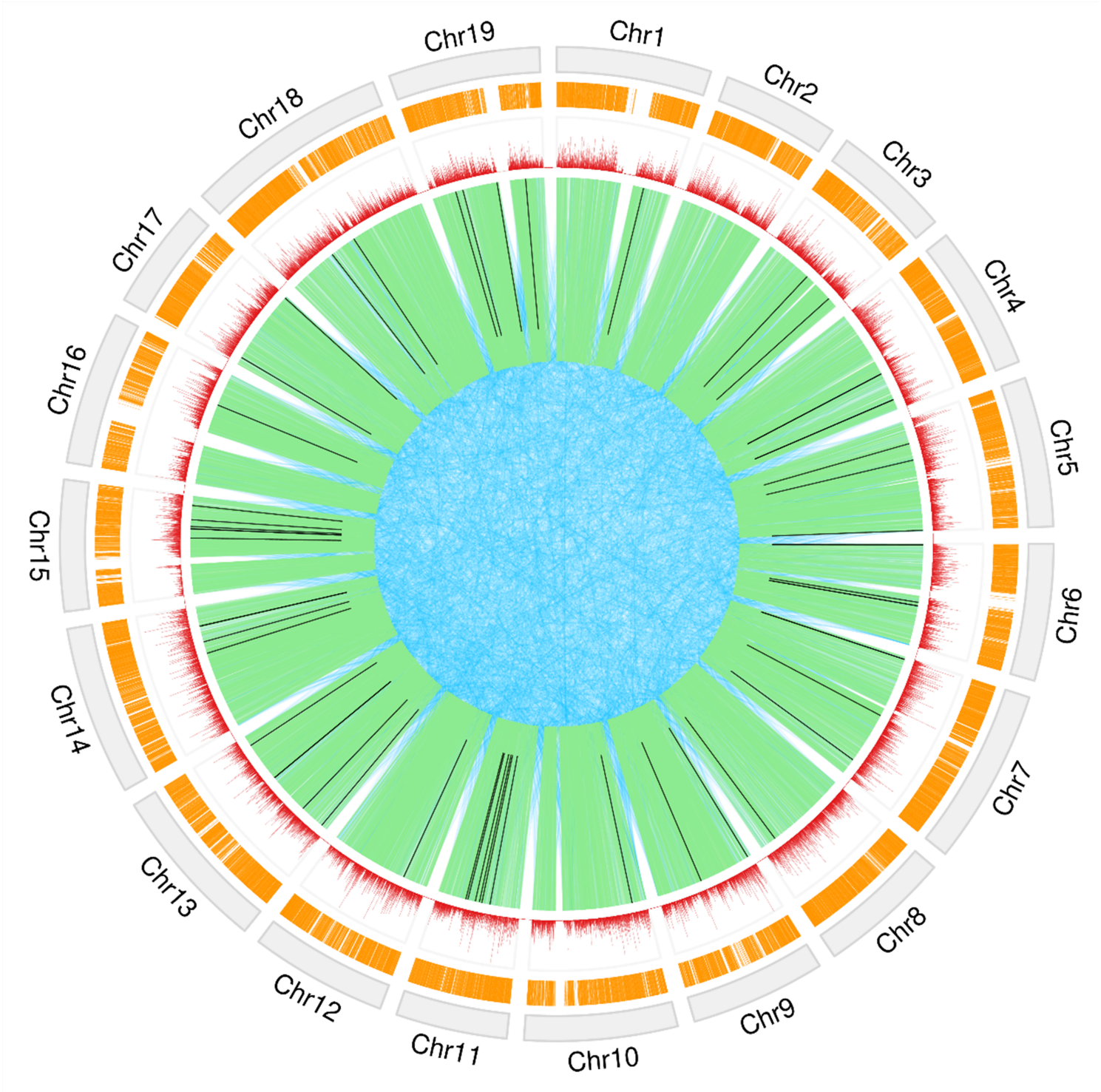


Figure 2

Genomic architecture and structural variation of the *Vitis vinifera* L. cultivar 'Parkent' assembly. The circular plot illustrates the distribution of genomic features based on a comparison with the telomere-to-

telomere reference genome (PN_T2T). The tracks represent, from outer to inner: the outer ring (grey) denotes the 19 reference chromosomes; the second track (orange blocks) shows the aligned regions of the ‘Parkent’ assembly; and the inner track (red) displays the micro-mutation density (single nucleotide polymorphisms and short insertions and deletions (indels)) calculated in 100,000 bp windows. Central links visualize large-scale genomic discordances, representing either potential misassemblies or putative structural variations: inter-chromosomal translocations are indicated by light blue lines, intra-chromosomal relocations by light green arcs, and inversions are highlighted by black arcs

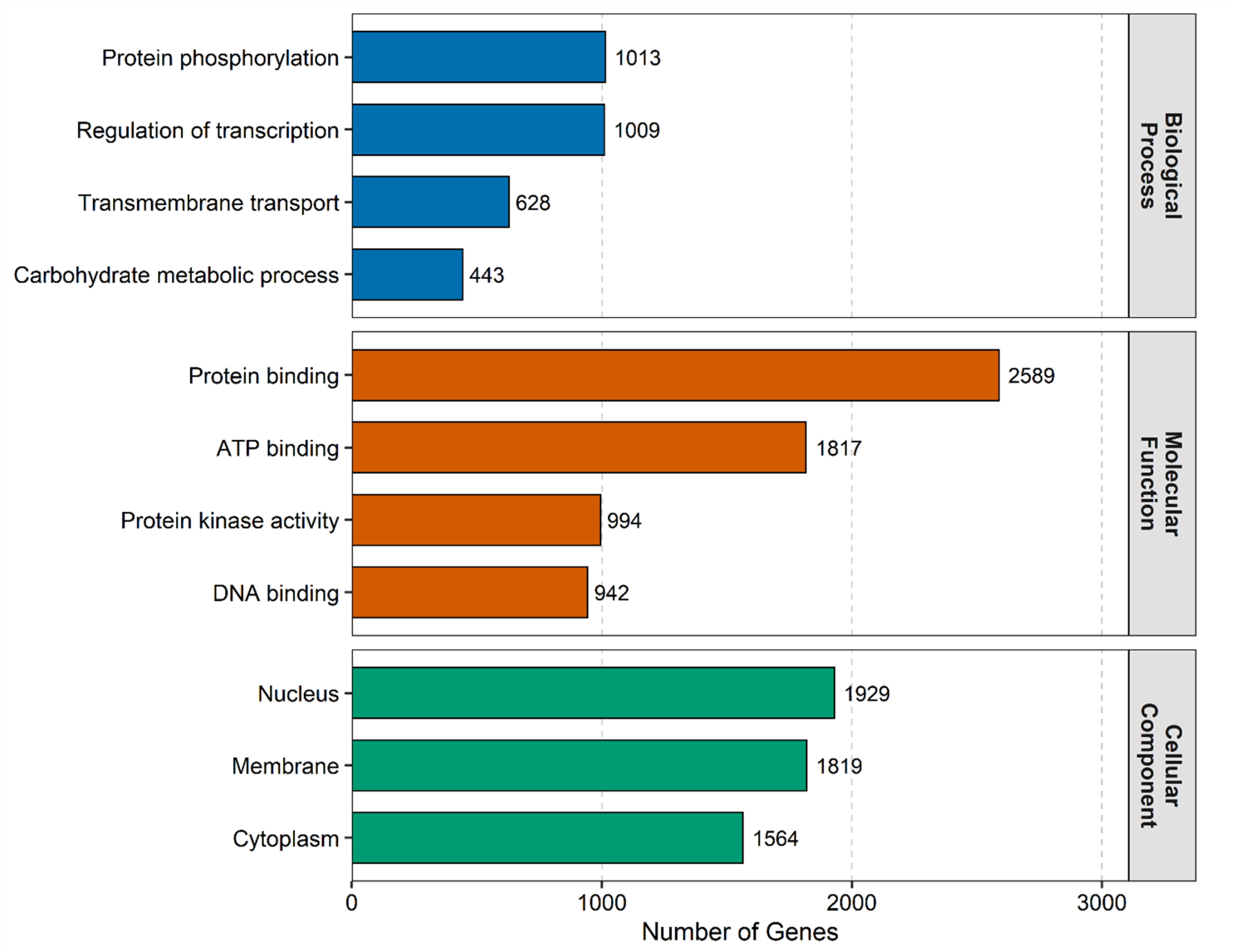


Figure 3

Functional classification of the predicted *Vitis vinifera* L. cultivar ‘Parkent’ proteome based on Gene Ontology (GO) assignments. The bar chart illustrates the distribution of the most highly represented GO terms across the three main ontological domains: Biological Process (top panel), Molecular Function (middle panel), and Cellular Component (bottom panel). The x-axis indicates the absolute number of protein-coding genes assigned to each specific GO category.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2.xlsx
- ESM3.xlsx
- ESM4.xlsx
- ESM5.xlsx
- ESM6.xlsx
- ESM7.xlsx
- ESM8.xlsx