The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding

Xiaoya Shi[1, 2, #,] Shuo Cao[2, 3, #], Xu Wang[2, 6, #], Siyang Huang[2, 4], Yue Wang[2, 5], Zhongjie Liu[2], Wenwen Liu[2], Xiangpeng Leng[1], Yanling Peng[2], Nan Wang[2], Yiwen Wang[2], Zhiyao Ma[2], Xiaodong Xu[2], Fan Zhang[2], Hui Xue[2], Haixia Zhong[7], Yi Wang[8], Kekun Zhang[9], Amandine Velt[10], Komlan Avia[10], Daniela Holtgräwe[11], Jérôme Grimplet[12], José Tomás Matus[13], Doreen Ware[14,15]; Xinyu Wu[7], Haibo Wang[16], Chonghuai Liu[17], Yuling Fang[9], Camille Rustenholz[10, *], Zongming Cheng[18, *], Hua Xiao[2, 7, *], Yongfeng Zhou[2, 19, *] 1 College of Horticulture, Qingdao Agricultural University, 266109, Qingdao, China

2 State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

3 Key Laboratory of Horticultural Plant Biology Ministry of Education, Huazhong Agricultural University, Wuhan, People's Republic of China

4 Guangxi Key Lab for Sugarcane Biology, Guangxi University, Nanning, Guangxi, 530005 China

5 State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing, China

6 School of Agriculture and Food Science, University College Dublin, Belfield, Dublin 4, Ireland

7 Institute of Horticulture Crops, Xinjiang Academy of Agricultural Sciences, Urumqi, China

8 Beijing Key Laboratory of Grape Science and Enology, Institute of Botany, Chinese Academy of Sciences, Xiangshan, Beijing 100093, China

9 College of Enology, Northwest A&F University, Yangling 712100, China

10 SVQV, INRAE - University of Strasbourg, 68000 Colmar, France

11 Genetics and Genomics of Plants, CeBiTec & Faculty of Biology, Bielefeld University, 33615 Bielefeld, Germany

12 Unidad de Hortofruticultura, Centro de Investigación y Tecnología Agroalimentaria de Aragón (CITA), 50059 Zaragoza, Spain

13 Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, Paterna, 46908, Valencia, Spain

14 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA; ware@cshl.edu (D.W.)

15 USDA ARS NEA Robert W. Holley Center for Agriculture and Health, Agricultural Research Service, Ithaca, NY 14853, USA

16 Fruit Research Institute,Chinese Academy of Agricultural Sciences/Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (Germplasm Resources Utilization), Ministry of

Agriculture/Key Laboratory of Mineral Nutrition and Fertilizers Efficient Utilization of Deciduous Fruit Tree, Liaoning Province, Xingcheng, China

17 Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou, China

18 College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

19 State Key Laboratory of Tropical Crop Breeding, Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, China

#, these authors contributed equally to this study

* Corresponding authors: zhouyongfeng@caas.cn; xiaohua01@caas.cn; zmc@njau.edu.cn; camille.rustenholz@inrae.fr

Abstract

Grapevine is one of the most economically important crops worldwide. However, the previous versions of the grapevine reference genome consisted of thousands of fragments with missing centromeres and telomeres, which limited the accessibility of the repetitive sequences, the centromeric and telomeric regions, and the inheritance of important agronomic traits in these regions. Here, we assembled a telomere-to-telomere (T2T) gap-free reference genome for the pinot noir cultivar (PN40024) using the PacBio HiFi long reads. The T2T reference genome (PN_T2T) was 69 Mb longer with 9026 more genes identified than the 12X.v2 version (Canaguier et al., 2017). We annotated 67% repetitive sequences, 19 centromeres and 36 telomeres, and incorporated gene annotations of previous versions into the PN_T2T. We detected a total of 377 gene clusters, which showed associations with complex traits, such as aroma and disease resistance. Even though the PN40024 sample had been selfed for nine generations, we still found nine genomic hotspots of heterozygous sites associated with biological processes, such as the oxidation-reduction process and protein phosphorylation. The fully annotated complete reference genome, therefore, provides important resources for grapevine genetics and breeding.

Keywords: Viticulture, T2T, gap-free, gene cluster, centromere, telomere, inbreeding

Introduction

Since the first human genome was published in 2000, lots of reference genomes have been assembled successively in a variety of species (Lander et al., 2001; Venter et al., 2001; Rice and Green, 2019). The reference genome is essential for biological and genetic studies. Thus, acquiring a high-quality genome has persistently been pursued. However, there are many missing segments due to highly repetitive sequences clustered across the genome, especially three representative regions: telomere, centromere and ribosome DNA (rDNA) (Rice and Green, 2019; Giani et al., 2020; Nurk et al., 2022). Centromere, which hosts CENPA/CENH3-variant nucleosomes and where the kinetochore forms and attaches to spindle microtubules, play an essential role during cell division. It consists of alpha satellites, a highly repetitive DNA sequence. The alpha satellite is composed of monomeric DNA repeats known as Higher Order Repeats (HOR), which contains arranged monomers that ranged from 100 to 200 bp (Talbert and Henikoff, 2020; Naish et al., 2021; Sundararajan and Straight, 2022). Despite their conserved function across species, their structure and sequence can change rapidly within and between species, and diverse organizations were observed from one species to another. However, centromere shows concerted evolution within the genome (Liao et al., 2018, Rudd et al., 2006; Melters et al., 2013; Naish et al., 2021). Currently, the centromere remains mostly unknown for researchers. A similar situation also exists in telomeres, which are composed of tandem repeats of relatively conserved microsatellite sequences located at the end of a chromosome in eukaryotes (Fajkus et al., 2005; Podlevsky and Chen, 2016). Telomeres are important for protecting chromosome terminal sequences during cell division (Turner et al., 2019; Coulon and Vaurs, 2020; Yuan et al., 2020; Engin and Engin, 2021). The rDNA is one of the most abundant repetitive elements in the genome that are essential for ribosome formation and play an important role in driving cell growth and cell proliferation (Kobayashi, 2011; Xu et al., 2020; Sasaki and Kobayashi, 2021). Because of the missing information on previously assembled genomes, the investigation of these regions has been extremely limited in the past two decades.

Fortunately, benefiting from the improvement of sequencing technology and computational algorithms, the genome assembly ushered in a new era: the telomere-to-telomere genome (T2T genome, Kille et al., 2022). Compared with the fragmented genome, the T2T genome has fewer or no gap on it based on the third-generation sequencing using PacBio high-fidelity long reads (HiFi), ultra-long Oxford Nanopore Technologies (ONT) and Hi-C data. Moreover, the T2T genome includes nearly complete information of telomere, centromere and rDNA regions (Logsdon et al., 2020; Miga and Sullivan, 2021). Promisingly, the T2T genome allows us to access these regions, opening a window into understanding the structure of these regions and the function of genes in these regions. Since the first complete human X chromosome was published in 2020, T2T assembly quickly become a research hotspot (Miga et al., 2020; Logsdon et al., 2021). In plants, the first T2T genome was reported in *Arabidopsis thaliana* in 2021 (Naish et al., 2021; Wang et al., 2022a). Presently, T2T genome assembly has been obtained in several species, such as rice, banana and watermelon, which fascinated the research on genomic structure and function, and crop breeding (Belser et al., 2021; Deng et al., 2022; Zhang et al., 2022; Yue et al., 2022).

The grapevine (*Vitis vinifera ssp. vinifera*), a Near East originated fruit tree, is one of the most widely cultivated and economically valuable crops worldwide (Grassi and De Lorenzis, 2021). Domesticated grapes often have highly heterozygous genomes (Zhou et al., 2019), which greatly impeded the acquirement of a high-quality genome. For instance, ~15% of genes were hemizygous in the Chardonnay genome (Zhou et al., 2019). Fortunately, PN40024, a highly homozygous Pinot Noir genotype that originated from successive selfings, was sequenced and the reference genome (8X) of grapevine was first obtained in 2007, which was also the first one for fruit crops (Jaillon et al., 2007). Subsequently, several updated versions have been released: the

12X.v2 version and its upgraded annotation VCost.v3 in 2017, and the PN40024.v4.1 version in 2021 (Canaguier et al., 2017; Navarro-Payá et al., 2021). In addition, fragmented genome assemblies of various grape cultivars have been produced in recent years such as Black Corinth (Massonnet et al., 2020), Cabernet Franc (Vondras et al., 2021; Minio et al., 2022), Cabernet Sauvignon (Chin et al., 2016; Minio et al., 2019b; Minio et al., 2022), Carménère (Minio et al., 2019a), Chardonnay (Roach et al., 2018; Zhou et al., 2019), Merlot (Massonnet et al., 2020), and Nebbiolo (Maestri et al., 2022). As a representative dicotyledonous plant in fruit trees, the high-quality genome will greatly facilitate the research on gene function, genetic structure and evolution of *Vitis* and eudicots species.

However, the previous incomplete genome assembly of grapes makes it difficult to access the highly repetitive regions on the genome. Here, we generated a T2T-level gap-free grape reference genome using the PN40024 material and aimed to address the following four questions. First, with the application of third-generation sequencing and assembly technologies, high-fidelity long reads have contributed to gap-free genome assemblies (Cheng et al., 2021; Mascher et al., 2021). Can we complete the reference genome of grapes using these new sequencing and assembly approaches? Second, the studies on centromere, telomere and rDNA have long been neglected. We analyzed the feature, structure, and distribution of these regions based on the assembled gapless grape genome. Third, the annotation of TE and genes in highly repetitive regions were improved based on the T2T genome, which could further improve our understanding of their biological functions, especially the gene clusters. Finally, the PN40024 genome was almost fully homozygous (Jaillon et al., 2007), but some sites remained heterozygous after nine generations of selfing. It is worthwhile to investigate the genomic distribution and the genetic effects of such heterozygous sites.

Results

A telomere to telomere gap-free reference genome for PN_T2T

PN40024, a highly homozygous Pinot Noir inbred line (Jaillon et al., 2007), was used for T2T genome assembly. In total, 21 Gb (21,024,461,524 bp, ~42X coverage) HiFi reads were generated by the PacBio platform. For the preliminary assembly, HiFiasm was used to assemble the HiFi reads. Using the published grapevine genomes as the reference, we then used Mummer to order the 38 contigs into 19 chromosomes (Figure 1). Only one gap left after initial assembly into contigs (Figure S1). After filling the gap with HiFi reads, a gap-free PN_T2T genome was finally generated (494.87 Mb), which is 69 Mb longer than 12X.v2 (426.18 Mb, Table 1) using the same statistical method. The K-mer was used to evaluate genomic homozygosity, estimated at 99.8% (Figure S2A-D). The BUSCO was used to evaluate genomic completeness, 98.5% of the core conserved plant genes were found complete in the genome assembly (Figure S2E), which is 4.8% more than the 12X.v2 (93.7%, Table 1).

Compared with the 12X.v2 genome, a substantial improvement was observed in our PN_T2T assembly. The contig N50 length of PN_T2T was ~ 250 times higher than that of 12X.v2 (26.89 Mb VS 102 kb), and all the 9429 gaps in 12X.v2 were filled in PN_T2T genome (Table 1, Table S1, Figure 1A). As shown in Figure 1C, 28 gaps in 12X.v2 were filled in PN_T2T with the largest gap being 16,951 bp in the 1Mb syntenic region on chromosome 18 (Figure 1C). Many orientation errors in 12X.v2 were also corrected such as inversions and translocations compared to PN_T2T (Figure 1A, Figure S3). For example, two large inversions, which were located surrounding the centromere of chromosome 3 and at the ends of chromosome 5, with the length of 0.9 M and 4.9 M were observed between two versions of assemblies, respectively (Figure 1A, Figure 1B). Moreover, 19 centromeres and 36 out of the 38 telomeres were detected on the PN_T2T genome assembly, except one telomere on chromosome 15 and one telomere on chromosome 17, which were missing in all previous grape genome assemblies. A total of 37,534 genes and 41,064 transcripts were annotated, among which 24526 (86.01%), 27696 (78.83%), 27717 (78.75%) were shared with older versions PN40024.V2.1 (https://phytozomenext.jgi.doe.gov/info/Vvinifera_v2_1), PN40024.v4, and PN40024.v4.1 (https://integrape.eu/resources/genes-genomes/genome-accessions/), respectively (Table S2). A total of 5472 (14.58%) genes were not found to correspond in any of the three versions. A total of 97.9% single-copy genes completely assembled was assessed by the BUSCO analysis, and structural domains were detected in 35508 sequences out of 40307 unique sequences (88.1%) while PN40024.V4.1 has 38364 unique sequences, and 29688 sequences were detected with structural domains (77.4%, Table S2).

Based on the species-specific Pan-TE database constructed by RepeatModeler2, the repeats were detected with a pipeline shown in Figure 2A. Finally, 66.47% of our gap-free grape genome was marked as repetitive sequences (Figure 1D). As a comparison, 62.47% of the repetitive sequences were identified in 12X.v2 genome using the same pipeline (Table S3). Among the repeats predicted in PN_T2T genome, the largest portion is transposable elements (TEs, 63.90%) with a total length of 316 Mb (59.96% and 292 Mb in 12X.v2). The TEs mainly consisted of the long terminal repeat (LTR) type (47.54%), predominantly Gypsy (20.22%) and Copia (19.67%) elements. In total, we detected 276 rDNA sequences, representing 0.019% of the genome.

The telomeres and centromeres

To access the telomeric and centromeric regions in PN_T2T, we identified the telomeres and centromeres using the pipeline in Figure 2A. For telomeres, we checked the 150 kb sequences at both ends of each chromosome, and the length of the telomere repeat unit was set to range from 5 to 12 bp. Finally, the telomere repeat unit (TTTAGGG/CCCTAAA) was detected, which was the most abundant in the genome and carried by all chromosomes. The same telomere repeat unit was reported in grapes by Melters et al. 2013 and Castro et al. 2021. We further predicted the telomeres in 36 out of 38 telomeres in the PN T2T genome, except the short arms of chromosome 15 and chromosome 17 (Figure 1A, Figure 2B and Table S4). Among them, the longest telomere (31 kb) was in the short arm of chromosome 8 with 4,479 repeats while the shortest telomere (1,260 bp) was in the long arm of chromosome 7 with only 180 repeats.

To detect the centromeric region, we scanned candidate repeats from 30 to 500 bp along the genome. The Tandem Repeats Finder (TRF) found 470 different repeat units in the PN_T2T genome, of which the 107 bp repeats were the most abundant unit in the whole genome, which had 182,620.5 (copies ≥ 2) repetitions accounted for about 3.95% of the genome, followed by 321 bp (2.45%), 214 bp (1.94%), and 135 bp (1.05%) (Figure 3A). Interestingly, we found the sequences of 214 bp and 321 bp repeat units consisted of two and three copies of the 107 bp repeat unit, respectively. The TE analyses also support the centromeric feature of the 107bp repetitive region (Figure 2). Thus, the centromeres were recognized mainly based on 107 bp repeat units, and localized on all 19 chromosomes (Figure 1A, Figure 2B, and Table S5). As shown in Figure 3B, the total length of 107 bp repeats varied from 1.4 kb to 3.8 Mb, but the sequences of the 107 bp repeats were highly conserved among chromosomes (Figure 3C). The 107 bp repeats were the most abundant in all chromosomes, except chromosomes 3, 14 and 18 (Figure 3D-H, Table S6). We found that the 187 bp was the main repeat unit in chromosome 14, it was scattered throughout the whole chromosome, and that 51 bp, 56 bp, 105 bp and 107 bp repeat units were highly overlapped and enriched in the centromere, which showed a core region of chromosome. The centromeric repeat unit in chromosome 3 was the 135 bp repeats and its integer multiples (270 bp and 405 bp). As for chromosome 18, 66 bp and its integer multiple 132 bp were the main repeat units (Figure S4).

To locate the centromeric repeats, we further examined the relationship between TE and centromeres. LTR retrotransposons or centromeric retrotransposons (CRs), were usually mixed with tandem repeats and enriched in plant centromeric regions (Guo et al., 2016; Fernandes et al., 2019). We found (Figure 4A) that the genes and TE repeats, such as LTR (Gypsy and Copia), DNA TE(MULE-MuDR) and RC (Helitron), have a low density in the special region where the enormous centromeric tandem repeats enriched in the chromosome were viewed in IGV (Figure 2 and Figure S4). We then inferred the region with centromeric repeats and low TE density as the centromeres after zooming one by one (Figure S4 and Table S5). The pattern of 107 bp was the target, which was highly linked with the centromeric region in grapes. However, there were likely different repeat units and patterns that appeared on chromosomes 3, 14, and 18 (Figure 3F-H). The scattering of transposons and the distribution of the centromere showed that specific sequence-defined repeat superfamilies are correlated or anticorrelated to various levels with centromeric proximity (Figure 2B, Figure 4A), forming density gradients that are the main chromosome-scale repeat-associated features, presumably reflecting overall chromatin structure (Figure S4).

To detect the captured genes, we then screened all genes in these regions in the highly linked centromeric region. Interestingly, we found 343 genes (Table S7) captured in the centromeres, which included 179 genes with Uniprot ID through blastp. Through GO (Gene Ontology) functional annotation, 12 genes were enriched in protein binding (molecular function, MF), such as *VvAMP1* (*Vitis01g01298* and *Vitis13g01021*; Uniprot ID: Q9M1S8) involved in ethylene (ETH), gibberellin (GA), and abscisic acid (ABA) signaling pathways (Saibo et al., 2007; Shi et al., 2013). In addition, we found 10 genes enriched in the cellular component (CC) of the cytosol, mitochondrion and cytoplasm respectively, including auxin transport protein *VvBIG* (*Vitis02g01141*; Uniprot ID: Q9SRU2), which influences general growth and development in plants (Gil et al., 2001); fumarate hydratase 1 *VvFUM1* (*Vitis02g01128*; Uniprot ID: P93033) catalyzes the active of mitochondrial Krebs cycle-associated enzyme (Zubimendi et al., 2018); 6-phosphogluconate dehydrogenase, decarboxylating 2 *VvPGD2* (*Vitis02g01123*, Uniprot ID: Q9FWA3) plays a key role in the development of the male gametophytes and interaction between the pollen tube and the ovule (Hölscher et al., 2016). Moreover, RNA modification, protein autophosphorylation, DNA integration, DNA recombination and photomorphogenesis were enriched in the biological process (BP) (Figure 4C).

Gene clusters in the grapevine reference genome

To infer the gene clusters in the grapevine genome, protein-to-protein alignments among the Pinot protein coding gene exposed a rich panoply of duplication structures in terms of genomic positions and functions. Prominent and complex tandem-like blocks of high-similarity genes can be seen via visualizations of all–vs.–all alignments (Figure S5). We found a total of 377 gene clusters in the grapevine reference genome (Table S8). These duplications often involve local rearrangements and can extend into megabases with dozens to hundreds of genes involved (Figure 5). On chromosome 16 (23-27 Mb), there were 599 genes enriched domains mainly including WAKs (Wall associated receptor kinase galacturonan binding), PPR repeat, Leucine-rich, ABC transporter, Intergrase domain, Peptidase family, Protein kinase and Reverse transcriptase (Figure 5A). And on chromosome 18 (25~36 Mb), there were 1237 genes enriched domains mainly including Intergrase domain, C JID domain, NB ARC domain, Leucine rich repeat, Multicopper oxidase, Reverse transcriptase, Terpene synthase and TIR. The results showed that many of the strongly enriched structural domains were part of the structural domains of plant disease resistance genes (R genes), including NB-ARC, TIR and structures identified by the Colis database. We analyzed the domain architecture of our 41,064 PN_T2T PCGs and identified 3,381 possible R genes. Collectively, these R genes and gene clusters in grapes indicated a tremendous opportunity for exploring plant defense mechanisms.

The genetic heterozygosity after the ninth generation of selfing

We are interested in the genomic changes associated with the inbreeding process. Based on the reference genome of PN_T2T, the resequencing data of four PN40024 clones were downloaded from NCBI and analyzed (Jaillon et al., 2007, Magris et al., 2019). A total of 244,215 SNPs were detected, among which 208,330 SNPs (85.3%) were shared in all four samples while the other 35,886 SNPs were only presented in 1-3 samples (Figure 6A). Interestingly, we found nine hotspots of heterozygous SNPs on chromosomes 1, 2, 3, 4, 7, 10, 11 and 16 (Figure 5A, Figure S6). To further investigate the highly heterozygous region, we examined the top 5% heterozygosity windows and identified a total of nine large continuous fragments (chromosome 1: 1.1-1.3 M, chromosome 2: 4.2-7.2 M, chromosome 3: 9.4-9.9 M, chromosome 4: 21.8-22.9 M, chromosome 7: 15.3-26.2 M, chromosome 10: 0.7-6.5 M, 17.6-18.3 M, chromosome 11: 7.1-7.8 M, chromosome 16: 13.0-13.5 M). The GO enrichment analysis on the genes in these regions showed that the most significantly enriched terms were response to water deprivation, protein phosphorylation, cell division, response to oxidative stress and response to salt stress, which were closely associated with key physiological activities in plants (Table S9, Figure 6C, Figure S7).

Discussion

The complete reference genome is essential for crop genetics and breeding. The previous versions of the grapevine reference genome have thousands of gaps with errors in repetitive regions and missing centromeres and telomeres, which limited the access of variants within these regions. Sometimes, such unreachable regions are underlying QTL of important agronomic traits, such as the berry color and sex determination on chromosome 2 (Fournier-Level et al., 2009; Zhou et al. 2017; Zhou et al. 2019; Zou et al. 2021) and disease resistance on chromosome 14 (Riaz et al., 2008; Morales-Cruz et al. 2022). The complete reference genome has great potential to reveal the missing heritability of important polygenic agronomic traits, therefore, it could increase the genetic gain in grapevine breeding. More and more investigations suggested the important functions of gene clusters, a total of 377 gene clusters were detected in the PN_T2T. The grapevine genome is also widely used in studies of plant evolution and comparative genomics because of its important phylogenetic position on the evolution of eudicots (Jaillon et al., 2007). The T2T version could be widely used in plant evolutionary genomics, especially at the repetitive sequences, centromeres and telomeres. The T2T gap-free reference genome had incorporated gene annotations of previous versions with more accurate TE annotation (up to ~ 67% of the genome), which will be an important resource for grapevine functional genomics and breeding.

The architecture and context of plant centromeres

The centromeric region ranges from Kbs to Gbs in length, including > 90% tandem repeats (McKinley and Cheeseman, 2016). The centromere is among the last pieces of great unknowns in genomics since it was inaccessible by previous sequencing technologies. The assemblies often collapse due to the highly repetitive nature of the centromeric region. We assembled and annotated centromeres for all 19 chromosomes of the grapevine genome (Figure 1). Most of the chromosomes have a single centromere while others could have multiple centromeric regions so called holocentromere (Steiner and Henikoff, 2014; Hofstatter et al., 2022). On chromosomes 16 and 18, we found tandem repeats in many regions while on other chromosomes only a single peak was detected (Figure 2B), suggesting that the structure of the centromeric region might be more complicated and requires further investigations.

In the PN40024 grapevine reference genome, there are three major repetitive patterns in the 19 chromosomes, suggesting different chromosomal evolutionary histories (Figure 3D-H). On chromosomes 3, 14 and 18, we found 135 bp, 56bp and 66 bp tandem repeats, respectively (Figure S4), while on other chromosomes, the major unit of tandem repeats is 107 bp (Figure 3D-H, Figure 4B, Figure 4D). The evolutionary histories of the grapevine centromere of each chromosome are still an open question to be addressed with all *Vitis* genomes. Previous comparative genomic analyses suggested that centromere is conservative among closely related species with a constant number of chromosomes (Liao et al., 2018). The transformation of centromeric structures happens during the chromosomal division and fusion when the number of chromosomes changes in evolution. The muscadine grape has 20 chromosomes with chromosomes 7 and 13 collinear with *Vitis* chromosome 7, which is associated with a chromosome fusion event (Cochetel et al., 2021). Only one centromeric region is left on chromosome 7 in our grapevine reference genome (Figure 2B, Figure S4), suggesting one centromere was lost during the evolvement of the genus *Vitis*.

The centromeric architecture shaped the content within the genome, population genetic diversity within species and genetic differentiation among species. Population genetic analyses revealed that the genetic variants in the centromeric region are highly linked with much lower genetic diversity compared to chromosome arms (Kawabe et al., 2008). The centromeres capture tens to thousands of genes that are highly linked with the centromeric tandem repeats. These genes along with the centromeric region are functional as supergenes. In total, we found 343 captured genes (Table S7) in the centromeric region in the grapevine reference genome. Interestingly, the genes are mainly involved in the ethylene (ETH), gibberellin (GA), and abscisic acid (ABA) signaling pathways (Saibo et al., 2007; Shi et al., 2013).

The hotspots of heterozygous variants in selfing plants

The current grapevine reference genome was generated by a Pinot Noir sample (PN40024) selfed for nine generations with high homozygosity at 99.8% of the genome (Figure S2A-D). However, we still interested in the remained heterozygous sites. Thus, we collected Illumina resequencing reads for four clones of PN40024 maintained in different international labs. Interestingly, the heterozygous SNPs and SVs were enriched in specific regions when mapped to the PN_T2T. In total, we found 208,330 heterozygous SNPs shared by the four samples, and 35,886 SNPs specific to 1-3 samples. The former is more likely the original variants of PN40024 after nine generations of selfing while the latter could be somatic mutations generated during the distribution and tissue culture in different labs. Interestingly, we found the hotspots of common variants were enriched in central biological processes including the oxidation-reduction process and protein phosphorylation. The hotspots on chromosome 2 also cover the sex-determination QTL region (Figure 6), which complicated the mining of the sex-determination genes (Zhou et al. 2019; Zou et al., 2021), because the candidate genes were not presented in the old version of the reference genome. It has been reported that during the clonal reproduction of fruit trees, such heterozygous deleterious variants accumulate in the genome (Zhou et al. 2019; Xiao et al., 2023). The clonal processes hide recessive deleterious variants including small SNPs and indels and large structural variants in a heterozygous state (Zhou et al., 2017; Zhou et al., 2019). A strong inbreeding depression has been commonly observed in clonal crops, including potato, cassava, citrus and grapevine (Zhou et al., 2017; Ramu et al., 2017; Zhang et al., 2021; Wang et al., 2022b;) since the strongly deleterious variants in these genomic regions has been exposed to lethal or strong recessive selection during selfing cycles. In grapevine breeding, the inbreeding and outcrossing depression were commonly detected because the hidden heterozygous recessive deleterious variants increased during clonal propagation has been exposed during sex reproduction.

Methods

Sample collection and genome sequencing

PN40024 is a line that belongs to one of the near homozygous lines originally derived from Pinot Noir by successive selfing steps, estimated the close to 97% homozygosity tested by SSR markers (Jaillon et al., 2007). We got this inbred material from INRAE under Material Transfer Agreement (MTA) and transplanted it in the greenhouse belonged to AGIS (Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China) for subsequent experiments. Young leaves and ovules from PN40024 were flash-frozen in liquid nitrogen. Genomic DNA and RNA were isolated using the DNeasy Plant Mini kit (Qiagen) following the manufacturer's instructions. For PacBio HIFI sequencing, two single-molecule real-time cells were sequenced on a PacBio Sequel II platform, and a total of 21 Gb of HiFi read was generated using CCS (https://github.com/PacificBiosciences/ccs) with the default parameter for the sequenced accessions. From each RNA-seq sample, isolate poly (A) mRNA 10 μg of total RNA was used to prepare Illumina RNA-seq libraries. These libraries were then sequenced using the Illumina HiSeqTM 2000 system in accordance with the manufacturer's instructions.

T2T genome assembly

Initially, PN40024 was assembled genome by incorporating PacBio single-molecule real-time long-read sequences. Reads generated by the PacBio Sequel II platform were self-corrected, trimmed and assembled by hifiasm, using default parameters (https://github.com/chhylp123/hifiasm, Cheng et al., 2021). The initial output of hifiasm (v.0.13) yielded the p_ctg draft assembly. Genome heterozygosity was estimated using a k-mer-based approach by GenomeScope2.0 (Ranallo-Benavidez et al., 2020), estimated close to 99.8% homozygosity (Figure S2A-D). Then, homology-based scaffolds were generated with MUMmer (v.4.0.0) (Marcais et al., 2018) "scaffold", using the 12X.v2 reference genome (Figure S3). By applying MUMmer tools, we order and orient the contig-level assemblies into 19 chromosomes, and join the adjacent contigs to generate a scaffold with 100 N. Finally, we adjusted the assembly manually through aligning the genome sequencing data from previous version of PN40024, which was mapped to the genome assembly by Minimap2 (v.2.21) and visualized in IGV (v.2.12.3) software to observe whether the gap regions were supported by reads (Figure S1). The filling and close of the gaps with the selected and assigned contigs were performed by mapping the 50 bp-sequences around the gap to continuous long reads (CLR) of PN40024.v4 and obtaining the gapless telomere-to-telomere PN40024 assembly for all 19 grape chromosomes. Assembly was inspected based on BUSCO (Simão et al., 2015) completeness and the duplication score.

The annotation of genes and TEs

We have used an self-developed method for genome annotation. The putative genes were first searched for by using transcripts and uniprot as evidence. A preliminary gene model was then built for the putative genes and the further search was performed using AUGUSTUS (V3.4.0) (Stanke et al., 2006). All the found putative genes fragments were then filtered, including genes involving duplicated regions, genes with CDS lengths shorter than 90 and genes not supported by any evidence. The missing genes were attempted to be complemented and the complete genes were subjected to the alternative splicing analyses. Finally, all the results were examined by hidden

Markov models downloaded from the Pfam database to obtain the final gene models. Interproscan (v.5.56-89.0, Jones et al., 2014) was used to function annotation for our assembly, Pfam (v.34.0, Mistry et al., 2021) and Coils (v.2.2.1, Fitzkee et al., 2005) was used for the identification of structural domains (https://github.com/unavailable-2374/Genome-Wide-Annotation-Pipeline).

The primary repeat analysis was outlined in Figure 2A and began with the construction of a Pan-Vitis database of repeat families by RepeatModeler (open-2.0.3, Flynn et al., 2020) and a series of scripts, which was then applied with RepeatMasker (open-4.1.2). For building this Pan-Vitis repeat database we download 17 Vitis genomes from NCBI, then use RepeatModeler2 to identify TE family. After that, we got 17 consensus fasta files of TE family, by removing the single copy and failed annotations we aggregated these files. We used NCBI-BLAST+2.9.0 (Altschul et al., 1990) to remove some redundancy sequences (-i 80%, -l 80%). After all, we got the final file of repeat identity, then we used deepTE (Yan et al., 2020) with the Plant model to classify those unclassified repeat elements. Finally, the repetitive sequence of the complete reference genome was annotated by RepeatMasker.

Genome comparison between different versions of the grapevine reference genome

To compare previous versions of grapevine genome with PN_T2T, we align the genomes using minimap2 and index the alignment BAM file using samtools(minimap2 -ax asm5 -t 4 --eqx A.fa B.fa | samtools sort -O BAM - > A_B.bam, samtools index A_B.bam). Next, to detect structural variations between genomes, we need to find synteny and structural rearrangements between the genomes. For this, we use SyRI: (syri -c A_B.bam -r A.fa -q B.fa -F B --prefix A_B). Finally, Plotsr were used to generate the graph: plotsr --sr A_Bsyri.out --sr B_Csyri.out --sr C_Dsyri.out --genomes genomes.txt -o output_plot.pdf, https://github.com/schneebergerlab/plotsr). MUMmer (v.4.0.0) was used to compare the 12X.v2 genome with the reference genome PN_T2T using whole-genome alignments (Marçais et al., 2018). First, we aligned the two genome sequences using nucmer (nucmer --mum) and then filtered one-to-one alignments with a minimum alignment length of 10,000 bp (delta-filter -i 95 -l 10000).

Samtools (v.1.7) were used to extract the sequence of chromosome 18: 25.0-26.0 Mb in 12X.v2 and aligned the sequence in PN_T2T. The gap information was detected with a python script (getgaps.py) and finally used LINKVIEW2 (https://github.com/YangJianshun/LINKVIEW2) to visualize the alignment results.

The identification of telomeres and centromeres

The telomere repeat units were explored by using the TIDK (v.0.2.0) (https://github.com/tolkit/telomeric-identifier) with options: tidk explore -f genome.fa --minimum 5 --maximum 12 -o tidk_explore -t 2 --log --dir telemere_find --extension TSV. Then the whole genome was searched using the parameter: tidk search -f genome.fa -s TTTAGGG -o tidk_search --dir telemere_find. Finally, we completed the rapid statistics of telomere based on the tidk plot and used R script to visualize the telomere peak.

For centromere annotation, the TRF (v.4.09) (Benson, 1999) was used to finish tandem repeats annotation with the parameter: trf genome.fa 2 7 7 80 10 50 500 -f -d -m, and then we merged the results of annotation by using TRF2GFF (https://github.com/Adamtaranto/TRF2GFF). To complete data statistics and visualization, we performed information extracted by using awk command in the linux system and analyzed the results in IGV (v.2.12.3) (Thorvaldsdóttir et al., 2013). We used four software to show more details about the centromeric region: Iqtree (v. 2.1.4-beta) (Minh et al., 2020) was used to achieve the phylogenetic tree (options: -m GTR+I+G -bb 1000 -bnni -alrt 1000); itol (v.6) (Letunic and Bork, 2021) was used to visualize the phylogenetic tree; GeneDoc (v.2.7.0) (https://github.com/karlnicholas/genedoc) was to achieve multiple sequence alignment; R script was used to plot the data statistics and typeset details respectively.

To detect the functions of the genes captured in the centromeric regions, we downloaded the protein sequence library of Swissprot (2022/08/30, https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/） for a local blast. After this, we extracted all the protein sequences of PN_T2T blasted by diamond (v.2.0.15) (parameter: - k 1 - e 0.00001, https://github.com/python-diamond/Diamond). We further uploaded the SwissProt ID to DAVID (https://david.ncifcrf.gov/tools.jsp) and completed Gene Ontology enrichment and annotation. Finally, Data visualization was completed by our R scripts.

The identification of gene clusters

To define the clustered genes in the reference genome, protein sequences were extracted using gffread and then filtered by e-value less than 1e-5 and similarity greater than 30% using blastp for all-vs-all alignments. The filtered alignment results were combined with functional annotations to filter out alignment results that did not share the same structural domains. Finally, we determined the presence of gene clusters by identifying three consecutive identical PF numbers, using such PF numbers as seeds, and going up and down 30 genes to find genes with the same PF number. In total, 377 gene clusters were found (Table S8).

The heterozygosity in selfed PN40024 clones

Four resequencing samples were downloaded from NCBI database (SRR6156373, SRR8835144, SRR8835157, SRR8835168) and mapped to newly assembled PN_T2T genome for SNP calling. Quality-controlled reads were mapped to the genome using bwa (v.0.7.15) with the default parameters. SAMtools (v.1.4) and GATK (v.4.1.8) were used for sorting and indexing the bam file with no duplicates. The gvcf files were combined in GATK and were used to join calling SNPs across all samples. To obtain high-quality SNPs, we performed strict filtering of the SNP calls based on the following criteria: (1) the SNPs with more than two alleles were removed in all samples in vcftools with parameters --min-alleles 2 --max-alleles 2; (2) we removed the SNPs with quality scores (GQ) less than 30 (--minGQ 30) and the missing rate is 0 (--max-missing 1); (3) SNPs had minor allele frequencies (MAFs) ≥ 0.01 to remove the invariable sites.

Data availability

All PacBio sequence data have been deposited to the NCBI Sequence Read Archive under the project number: PRJNA882193 and the National Genomics Data Center (NGDC) Genome Sequence Archive (GSA) (https://ngdc.cncb.ac.cn/gsa/), with BioProject number PRJCA012093. The assembly and annotation have been deposited to zenodo: https://zenodo.org/record/7751391#.ZBgVmcJBy3A.

Code availability

All the scripts and pipelines used in this study have been achieved in GitHub: https://github.com/zhouyflab

Reference

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology **215**:403-410. 10.1016/s0022-2836(05)80360-2.

Belser, C., Baurens, F.C., Noel, B., Martin, G., Cruaud, C., Istace, B., Yahiaoui, N., Labadie, K., Hřibová, E., Doležel, J., et al. (2021). Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Communications biology **4**:1047. 10.1038/s42003-021-02559-3.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research **27**:573-580. 10.1093/nar/27.2.573.

Canaguier, A., Grimplet, J., Di Gaspero, G., Scalabrin, S., Duchêne, E., Choisne, N., Mohellibi, N., Guichard, C., Rombauts, S., Le Clainche, I., et al. (2017). A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genomics data **14**:56-62. 10.1016/j.gdata.2017.09.002.

Castro, C., Carvalho, A., Gaivão, I., and Lima-Brito, J. (2021). Evaluation of copper-induced DNA damage in Vitis vinifera L. using Comet-FISH. Environmental science and pollution research international **28**:6600-6610. 10.1007/s11356-020-10995-7.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature methods **18**:170-175. 10.1038/s41592-020-01056-5.

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature methods **13**:1050-1054. 10.1038/nmeth.4035.

Cochetel, N., Minio, A., Massonnet, M., Vondras, A.M., Figueroa-Balderas, R., and Cantu, D. (2021). Diploid chromosome-scale assembly of the Muscadinia rotundifolia genome supports chromosome fusion and disease resistance gene expansion during Vitis and Muscadinia divergence. G3 (Bethesda, Md.) **11**10.1093/g3journal/jkab033.

Coulon, S., and Vaurs, M. (2020). Telomeric Transcription and Telomere Rearrangements in Quiescent Cells. Journal of molecular biology **432**:4220-4231. 10.1016/j.jmb.2020.01.034.

Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., Xu, B., Tian, Y., Sun, Y., Li, B., et al. (2022). A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provides important resources for gene discovery and breeding. Molecular plant **15**:1268-1284. 10.1016/j.molp.2022.06.010.

Engin, A.B., and Engin, A. (2021). The Connection Between Cell Fate and Telomere. Advances in experimental medicine and biology **1275**:71-100. 10.1007/978-3-030-49844-3_3.

Fajkus, J., Sýkorová, E., and Leitch, A.R. (2005). Telomeres in evolution and evolution of telomeres. Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology **13**:469-479. 10.1007/s10577-005-0997-2.

Fernandes, J.B., Wlodzimierz, P., and Henderson, I.R. (2019). Meiotic recombination within plant centromeres. Current opinion in plant biology **48**:26-35. 10.1016/j.pbi.2019.02.008.

Fitzkee, N.C., Fleming, P.J., and Rose, G.D. (2005). The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB. Proteins

58:852-854. 10.1002/prot.20394.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for the automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences of the United States of America **117**:9451-9457. 10.1073/pnas.1921046117.

Fournier-Level, A., Le Cunff, L., Gomez, C., Doligez, A., Ageorges, A., Roux, C., Bertrand, Y., Souquet, J.M., Cheynier, V., and This, P. (2009). Quantitative genetic bases of anthocyanin variation in grape (Vitis vinifera L. ssp. sativa) berry: a quantitative trait locus to quantitative trait nucleotide integrated study. Genetics **183**:1127-1139. 10.1534/genetics.109.103929.

Giani, A.M., Gallo, G.R., Gianfranceschi, L., and Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and structural biotechnology journal **18**:9-19. 10.1016/j.csbj.2019.11.002.

Grassi, F., and De Lorenzis, G. (2021). Back to the Origins: Background and Perspectives of Grapevine Domestication. International journal of molecular sciences **22**10.3390/ijms22094518.

Guo, X., Su, H., Shi, Q., Fu, S., Wang, J., Zhang, X., Hu, Z., and Han, F. (2016). De Novo Centromere Formation and Centromeric Sequence Expansion in Wheat and its Wide Hybrids. PLoS genetics **12**:e1005997. 10.1371/journal.pgen.1005997.

Hofstatter, P.G., Thangavel, G., Lux, T., Neumann, P., Vondrak, T., Novak, P., Zhang, M., Costa, L., Castellani, M., Scott, A., et al. (2022). Repeat-based holocentromeres influence genome architecture and karyotype evolution. Cell **185**:3153-3168.e3118. 10.1016/j.cell.2022.06.045.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics **12**:491. 10.1186/1471-2105-12-491.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**:463-467. 10.1038/nature06148.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics (Oxford, England) **30**:1236-1240. 10.1093/bioinformatics/btu031.

Kawabe, A., Forrest, A., Wright, S.I., and Charlesworth, D. (2008). High DNA sequence diversity in pericentromeric genes of the plant Arabidopsis lyrata. Genetics **179**:985-995. 10.1534/genetics.107.085282.

Kille, B., Balaji, A., Sedlazeck, F.J., Nute, M., and Treangen, T.J. (2022). Multiple genome alignment in the telomere-to-telomere assembly era. Genome biology **23**:182. 10.1186/s13059-022-02735-6.

Kobayashi, T. (2011). How does genome instability affect lifespan?: roles of rDNA and telomeres. Genes to cells : devoted to molecular & cellular mechanisms **16**:617-624. 10.1111/j.1365-2443.2011.01519.x.

Korf, I. (2004). Gene finding in novel genomes. BMC bioinformatics **5**:59. 10.1186/1471-2105-5-59.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature **409**:860-921. 10.1038/35057062.

Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic acids research **49**:W293-w296. 10.1093/nar/gkab301.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England) **22**:1658-1659. 10.1093/bioinformatics/btl158.

Liao, Y., Zhang, X., Li, B., Liu, T., Chen, J., Bai, Z., Wang, M., Shi, J., Walling, J.G., Wing, R.A., et al. (2018). Comparison of Oryza sativa and Oryza brachyantha Genomes Reveals Selection-Driven Gene Escape from the Centromeric Regions. The

Plant cell **30**:1729-1744. 10.1105/tpc.18.00163.

Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. Nature reviews. Genetics **21**:597-614. 10.1038/s41576-020-0236-x.

Logsdon, G.A., Vollger, M.R., Hsieh, P., Mao, Y., Liskovykh, M.A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A., et al. (2021). The structure, function and evolution of a complete human chromosome 8. Nature **593**:101-107. 10.1038/s41586-021-03420-7.

Maestri, S., Gambino, G., Lopatriello, G., Minio, A., Perrone, I., Cosentino, E., Giovannone, B., Marcolungo, L., Alfano, M., Rombauts, S., et al. (2022). 'Nebbiolo' genome assembly allows surveying the occurrence and functional implications of genomic structural variations in grapevines (Vitis vinifera L.). BMC genomics **23**:159. 10.1186/s12864-022-08389-9.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLoS computational biology **14**:e1005944. 10.1371/journal.pcbi.1005944.

Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C.S., Ens, J., Gundlach, H., Boston, L.B., Tulpová, Z., et al. (2021). Long-read sequence assembly: a technical evaluation in barley. The Plant cell **33**:1888-1906. 10.1093/plcell/koab077.

Massonnet, M., Cochetel, N., Minio, A., Vondras, A.M., Lin, J., Muyle, A., Garcia, J.F., Zhou, Y., Delledonne, M., Riaz, S., et al. (2020). The genetic basis of sex determination in grapes. Nature communications **11**:2902. 10.1038/s41467-020-16700-z.

Magris, G., Di Gaspero, G., Marroni, F., Zenoni, S., Tornielli, G.B., Celii, M., De Paoli, E., Pezzotti, M., Conte, F., Paci, P., et al. (2019). Genetic, epigenetic and genomic effects on variation of gene expression among grape varieties. The Plant journal : for cell and molecular biology 99:895-909. 10.1111/tpj.14370.

McKinley, K.L., and Cheeseman, I.M. (2016). The molecular basis for centromere identity and function. Nature reviews. Molecular cell biology **17**:16-29. 10.1038/nrm.2015.5.

Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, J.G., Sebra, R., Peluso, P., Eid, J., Rank, D., et al. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome biology **14**:R10. 10.1186/gb-2013-14-1-r10.

Miga, K.H., and Sullivan, B.A. (2021). Expanding studies of chromosome structure and function in the era of T2T genomics. Human molecular genetics **30**:R198-r205. 10.1093/hmg/ddab214.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature **585**:79-84. 10.1038/s41586-020-2547-7.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular biology and evolution **37**:1530-1534. 10.1093/molbev/msaa015.

Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A., and Cantu, D. (2019a). Diploid Genome Assembly of the Wine Grape Carménère. G3 (Bethesda, Md.) **9**:1331-1337. 10.1534/g3.119.400030.

Minio, A., Cochetel, N., Vondras, A.M., Massonnet, M., and Cantu, D. (2022). Assembly of complete diploid-phased chromosomes from draft genome sequences. G3 (Bethesda, Md.) **12**10.1093/g3journal/jkac143.

Minio, A., Massonnet, M., Figueroa-Balderas, R., Vondras, A.M., Blanco-Ulate, B., and Cantu, D. (2019b). Iso-Seq Allows Genome-Independent Transcriptome Profiling of Grape Berry Development. G3 (Bethesda, Md.) **9**:755-767. 10.1534/g3.118.201008.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al.

(2021). Pfam: The protein families database in 2021. Nucleic acids research **49**:D412-d419. 10.1093/nar/gkaa913.

Morales-Cruz, A., Aguirre-Liguori, J. Massonnet, M.,Minio. A., Zaccheo, M., Cochetel, Noe., Walker, A., Riaz, S., Zhou, Y.F., Cantu, D., Gaut, B.S. (2022). Multigenic resistance to Xylella fastidiosa in wild grapes (Vitis sps.) and its implications within a changing climate. bioRxiv, doi: https://doi.org/10.1101/2022.10.08.511428

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A.J., Abramson, B.W., Schmücker, A., Mandáková, T., Jamge, B., Lambing, C., Kuo, P., et al. (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. Science (New York, N.Y.) **374**:eabi7489. 10.1126/science.abi7489.

Navarro-Payá, D., Santiago, A., Orduña, L., Zhang, C., Amato, A., D'Inca, E., Fattorini, C., Pezzotti, M., Tornielli, G.B., Zenoni, S., et al. (2021). The Grape Gene Reference Catalogue as a Standard Resource for Gene Selection and Genetic Improvement. Frontiers in plant science **12**:803977. 10.3389/fpls.2021.803977.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science (New York, N.Y.) **376**:44-53. 10.1126/science.abj6987.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols **11**:1650-1667. 10.1038/nprot.2016.095.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology **33**:290-295. 10.1038/nbt.3122.

Podlevsky, J.D., and Chen, J.J. (2016). Evolutionary perspectives of telomerase RNA structure and function. RNA biology **13**:720-732. 10.1080/15476286.2016.1205768.

Ramu, P., Esuma, W., Kawuki, R., Rabbi, I.Y., Egesi, C., Bredeson, J.V., Bart,

R.S., Verma, J., Buckler, E.S., and Lu, F. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nature genetics **49**:959-963. 10.1038/ng.3845.

Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature communications **11**:1432. 10.1038/s41467-020-14998-3.

Riaz, S., Tenscher, A.C., Rubin, J., Graziani, R., Pao, S.S., and Walker, M.A. (2008). Fine-scale genetic mapping of two Pierce's disease resistance loci and a major segregation distortion region on chromosome 14 of grape. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik **117**:671-681. 10.1007/s00122-008-0802-7.

Rice, E.S., and Green, R.E. (2019). New Approaches for Genome Assembly and Scaffolding. Annual review of animal biosciences **7**:17-40. 10.1146/annurev-animal-020518-115344.

Roach, M.J., Johnson, D.L., Bohlmann, J., van Vuuren, H.J.J., Jones, S.J.M., Pretorius, I.S., Schmidt, S.A., and Borneman, A.R. (2018). Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. PLoS genetics **14**:e1007807. 10.1371/journal.pgen.1007807.

Rudd, M.K., Wray, G.A., and Willard, H.F. (2006). The evolutionary dynamics of alpha-satellite. Genome research **16**:88-96. 10.1101/gr.3810906.

Sasaki, M., and Kobayashi, T. (2021). Gel Electrophoresis Analysis of rDNA Instability in Saccharomyces cerevisiae. Methods in molecular biology (Clifton, N.J.) **2153**:403-425. 10.1007/978-1-0716-0644-5_28.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics (Oxford, England) **31**:3210-3212. 10.1093/bioinformatics/btv351.

Song, J., Logeswaran, D., Castillo-González, C., Li, Y., Bose, S., Aklilu, B.B., Ma, Z., Polkhovskiy, A., Chen, J.J., and Shippen, D.E. (2019). The conserved structure of plant telomerase RNA provides the missing link for an evolutionary pathway from ciliates to humans. Proceedings of the National Academy of Sciences of the United States of America **116**:24542-24550. 10.1073/pnas.1915312116.

Song, J.M., Xie, W.Z., Wang, S., Guo, Y.X., Koo, D.H., Kudrna, D., Gong, C., Huang, Y., Feng, J.W., Zhang, W., et al. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. Molecular plant **14**:1757-1767. 10.1016/j.molp.2021.06.018.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research **34**:W435-439. 10.1093/nar/gkl200.

Steiner, F.A., and Henikoff, S. (2014). Holocentromeres are dispersed point centromeres localized at transcription factor hotspots. eLife **3**:e02025. 10.7554/eLife.02025.

Talbert, P.B., and Henikoff, S. (2020). What makes a centromere? Experimental cell research **389**:111895. 10.1016/j.yexcr.2020.111895.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics **14**:178-192. 10.1093/bib/bbs017.

Turner, K.J., Vasu, V., and Griffin, D.K. (2019). Telomere Biology and Human Phenotype. Cells **8**10.3390/cells8010073.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science (New York, N.Y.) **291**:1304-1351. 10.1126/science.1058040.

Vondras, A.M., Lerno, L., Massonnet, M., Minio, A., Rowhani, A., Liang, D., Garcia, J., Quiroz, D., Figueroa-Balderas, R., Golino, D.A., et al. (2021). Rootstock influences the effect of grapevine leafroll-associated viruses on berry development and metabolism via abscisic acid signalling. Molecular plant pathology **22**:984-1005. 10.1111/mpp.13077.

Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., Wang, S., Xu, T., Zhao, X., Gao, S., et al. (2022a). High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long Reads. Genomics, proteomics & bioinformatics **20**:4-13. 10.1016/j.gpb.2021.08.003.

Wang, N., Song, X., Ye, J., Zhang, S., Cao, Z., Zhu, C., Hu, J., Zhou, Y., Huang, Y., Cao, S., et al. (2022b). Structural variation and parallel evolution of apomixis in citrus during domestication and diversification. National Science Review 10.1093/nsr/nwac114.

Xu, Y., Wu, Y., Wang, L., Qian, C., Wang, Q., and Wan, W. (2020). Identification of curcumin as a novel natural inhibitor of rDNA transcription. Cell cycle (Georgetown, Tex.) **19**:3362-3374. 10.1080/15384101.2020.1843817.

Yan, H., Bombarely, A., and Li, S. (2020). DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics (Oxford, England) **36**:4269-4275. 10.1093/bioinformatics/btaa519.

Yuan, X., Dai, M., and Xu, D. (2020). Telomere-related Markers for Cancer. Current topics in medicinal chemistry **20**:410-432. 10.2174/1568026620666200106145340.

Yue, J., Chen, Q., Wang, Y., Zhang, L., Ye, C., Wang, X., Cao, S., Lin, Y., Huang, W., Xian, H., et al. (2022). Telomere-to-telomere and gap-free reference genome assembly of the kiwifruit Actinidia chinensis. Horticulture Research uhac264. 10.1093/hr/uhac264.

Zhang, C., Yang, Z., Tang, D., Zhu, Y., Wang, P., Li, D., Zhu, G., Xiong, X., Shang, Y., Li, C., et al. (2021). Genome design of hybrid potato. Cell **184**:3873-3883.e3812. 10.1016/j.cell.2021.06.006.

Zhang, Y., Fu, J., Wang, K., Han, X., Yan, T., Su, Y., Li, Y., Lin, Z., Qin, P., Fu, C., et al. (2022). The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. Plant biotechnology journal **20**:1642-1644. 10.1111/pbi.13880.

Zhou, Y., Massonnet, M., Sanjak, J.S., Cantu, D., and Gaut, B.S. (2017). 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. 10.1073/pnas.1709257114.

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., and Gaut, B.S. (2019). The population genetics of structural variants in grapevine domestication. Nature plants **5**:965-979. 10.1038/s41477-019-0507-8.

Gil, P., Dewey, E., Friml, J., Zhao, Y., Snowden, K.C., Putterill, J., Palme, K., Estelle, M., and Chory, J. (2001). BIG: a calossin-like protein required for polar auxin transport in Arabidopsis. Genes & development **15**:1985-1997. 10.1101/gad.905201.

Saibo, N.J., Vriezen, W.H., De Grauwe, L., Azmi, A., Prinsen, E., and Van der Straeten, D. (2007). A comparative analysis of the Arabidopsis mutant amp1-1 and a novel weak amp1 allele reveals new functions of the AMP1 protein. Planta **225**:831-842. 10.1007/s00425-006-0395-9.

Shi, H., Ye, T., Wang, Y., and Chan, Z. (2013). Arabidopsis ALTERED MERISTEM PROGRAM 1 negatively modulates plant responses to abscisic acid and dehydration stress. Plant physiology and biochemistry : PPB **67**:209-216. 10.1016/j.plaphy.2013.03.016.

Zubimendi, J.P., Martinatto, A., Valacco, M.P., Moreno, S., Andreo, C.S., Drincovich, M.F., and Tronconi, M.A. (2018). The complex allosteric and redox regulation of the fumarate hydratase and malate dehydratase reactions of Arabidopsis thaliana Fumarase 1 and 2 gives clues for understanding the massive accumulation of fumarate. The FEBS journal **285**:2205-2224. 10.1111/febs.14483.

Hölscher, C., Lutterbey, M.C., Lansing, H., Meyer, T., Fischer, K., and von Schaewen, A. (2016). Defects in Peroxisomal 6-Phosphogluconate Dehydrogenase Isoform PGD2 Prevent Gametophytic Interaction in Arabidopsis thaliana. Plant physiology **171**:192-205. 10.1104/pp.15.01301.

Zou, C., Massonnet, M., Minio, A., Patel, S., Llaca, V., Karn, A., Gouker, F., Cadle-Davidson, L., Reisch, B., Fennell, A., et al. (2021). Multiple independent recombinations led to hermaphroditism in grapevine. Proceedings of the National

Academy of Sciences of the United States of America 11810.1073/pnas.2023548118.

Hua Xiao, Zongjie Liu, Nan Wang, Shuo Cao , Guizhou Huang , Wenwen Liu , Yanling Peng , Qiming Long, Summaira Riaz , Andrew M. Walker , Brandon S. Gaut, and Yongfeng Zhou. (2023). The adaptive and maladaptive introgression in grapevine domestication. PNAS, in press

Acknowledgments

This work was supported by the National Natural Science Fund for Excellent Young Scientists Fund Program (Overseas) to Yongfeng Zhou, the National Key Research and Development Program of China(grant2019YFA0906200), the Agricultural Science and Technology Innovation Program (CAAS-ZDRW202101), the Shenzhen Science and Technology Program (grant KQTD2016113010482651), the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI). We thank Bianca Frommer, Marie Lahaye, David Navarro-Payá, Marcela K. Tello-Ruiz and Kapeel Chougule for their help in analyzing the RNA-Seq data and in running the gene annotation pipeline. This study is also based upon work from COST Action CA17111 INTEGRAPE and form COST Innovators Grant IG17111 GRAPEDIA, supported by COST (European Cooperation in Science and Technology).

Author contributions

Y.Z. conceived and designed the project with H.X., Z.C. and C.R.. Z.C. provided the PN40024 sample. X.S. W.L., X.X and Z.M. performed the tissue culture of the sample in the greenhouse. X.S., X.W., H.X., N.W., F.Z., H.X. and Y.W. performed the bioinformatic analyses. A.V., K.A., D.H., J. G., J.T., D.W. Z.L., X.L. and W.L. performed the gene annotation. Y.P., S.H. Z.L., W.L., X.W., Y.F., Y.W and C.L. assisted in bioinformatics analyses. X.S., S.C., X.W., H.X. and Y.Z. wrote the manuscript with comments and inputs from all authors.

Conflict of interests

The authors declare no conflict of interest.

Table 1. Comparison of genomic features of 12X.v2 and PN_T2T assembly

	12X.v2	**PN_T2T**
Total sequence length (bp)	426,176,009	494,873,210
Number of chromosomes	19	19
Contig N50 (bp)	102,700	26,899,771
Max length	30,274,277	36,684,271
Number of gaps	9429	0
Centromere	-	19/19
Telomere	-	36/38
Bases masked (bp)	303,719,475	328,929,883
Retroelements (bp)	217,819,122	241,027,616
LTR (bp)	212,117,752	235,245,099
The number of genes	28,516	37,534
The number of TE	942,096	935,783
BUSCO	93.70%	98.50%

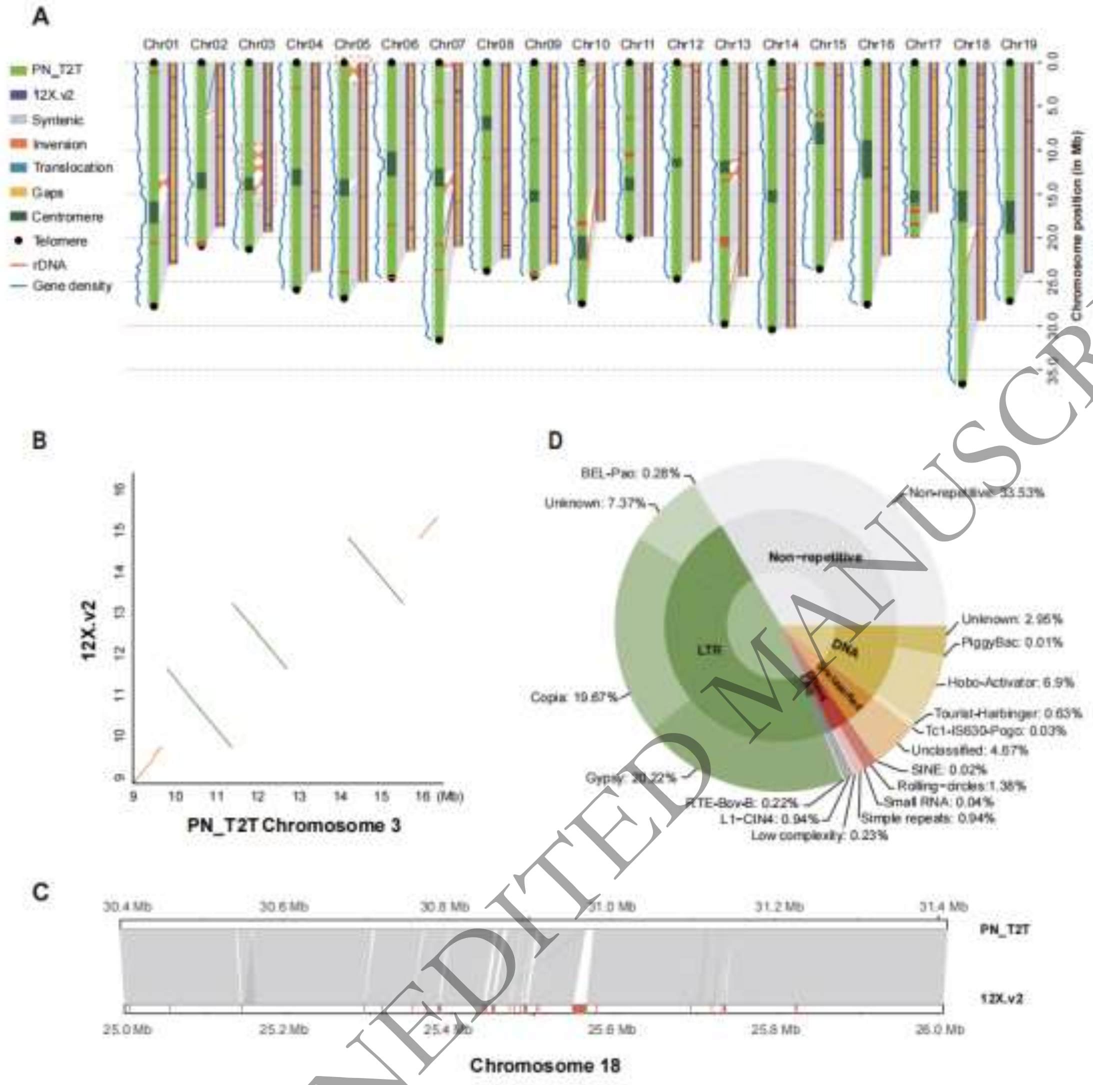

Figure 1. The T2T gap-free assembly of the grapevine reference genome. (A) An overview of the genome assemblies (12X.v2 right, PN_T2T left). The red dashed boxes on Chromosome 3 and Chromosome 5 indicated differences in large inversions between the two versions of genomic assemblies. (B) A zoomed-in portion of the red dashed box region on Chromosome 3 in A. (C) Plot showing 1 Mb syntenic region between the 12X.v2 and PN_T2T assemblies on Chromosome 18. Grey bands connected corresponding collinear regions, and red boxes at the bottom showed the gaps in 12X.v2. (D) Types and percentages of different TE families detected in the PN_T2T genome.

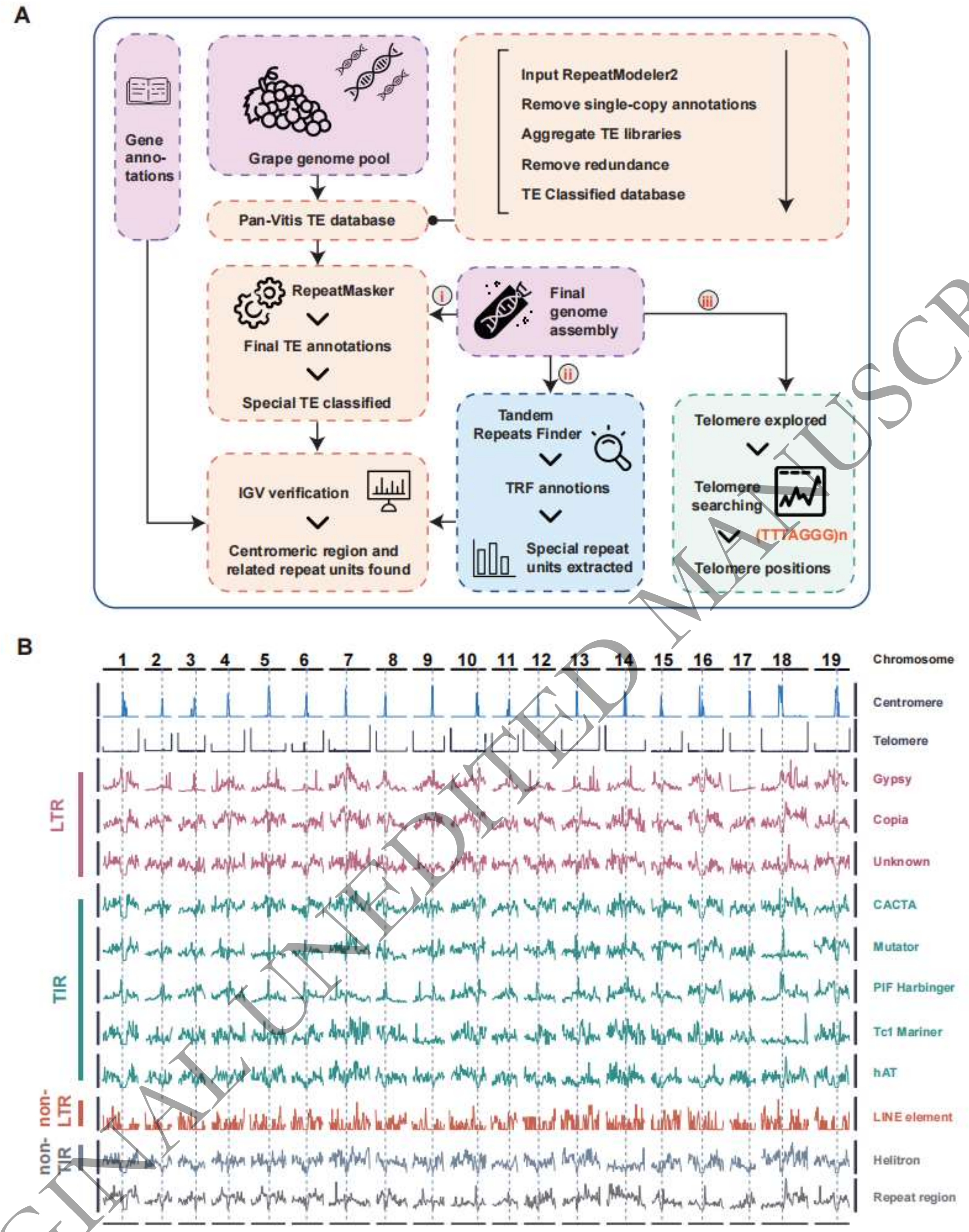

Figure 2. The repeats annotation in PN_T2T reference genome. (A) Dataflow of centromere and telomere predictions. (B) Chromosomal distribution of telomeres, centromeres and different types of TEs. Dotted vertical lines indicated the center locations of predicted centromeres.

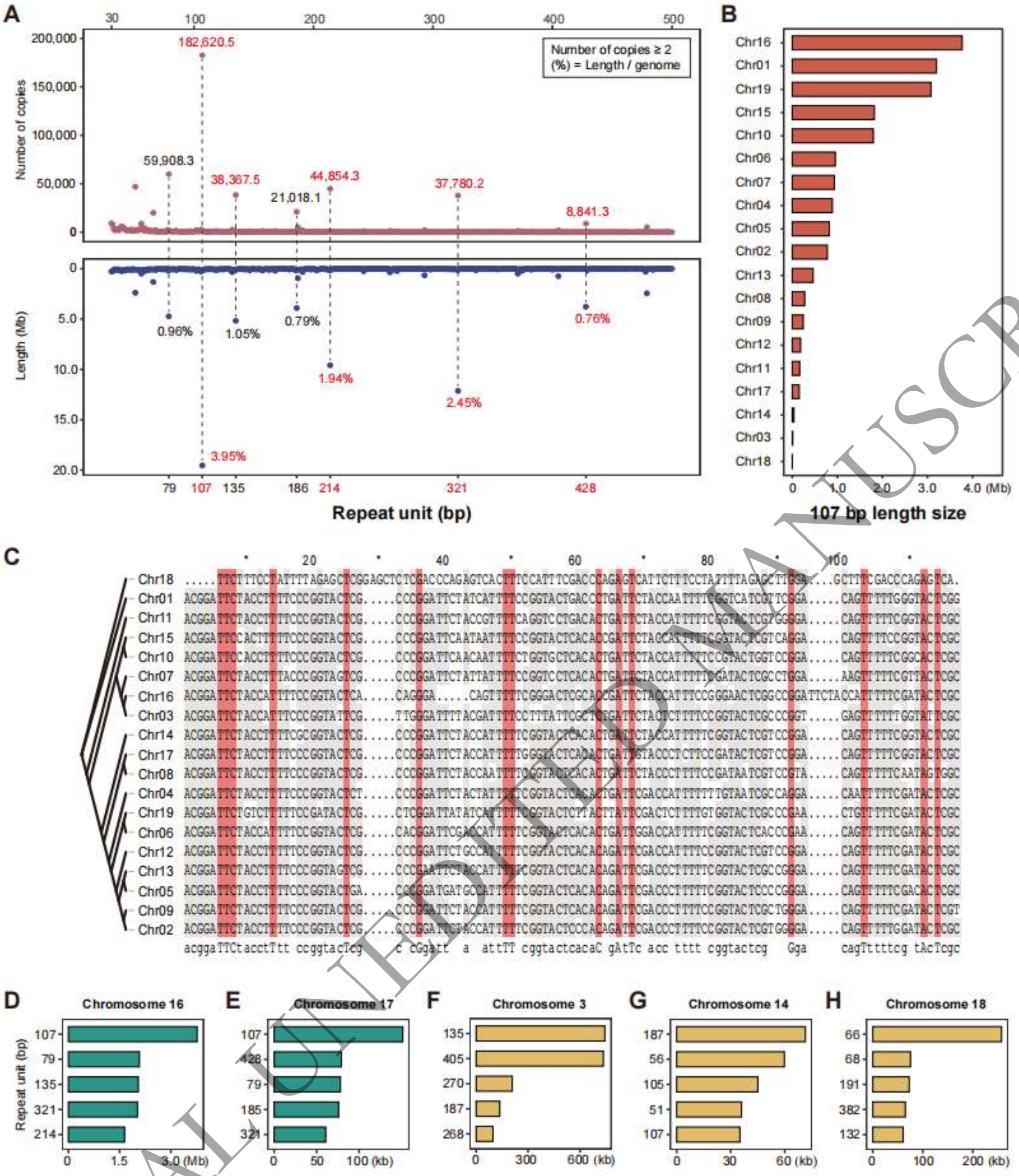

Figure 3. The schematic illustration of centromeric repeat units in the PN_T2T genome. (A) The distribution of different repeat units' lengths in the whole genome. The number of different repeat unit copies was indicated in the upper part of the graphs while the chromosomal percentage of different repeat units was shown in the lower part. (B) The total length of 107 bp repeat unit copies in each chromosome. (C) The alignment of the 107 bp repeat units among 19 chromosomes. (D-H) The total length of different repeat units in chromosomes 16, 17, 3, 14 and 18, respectively.

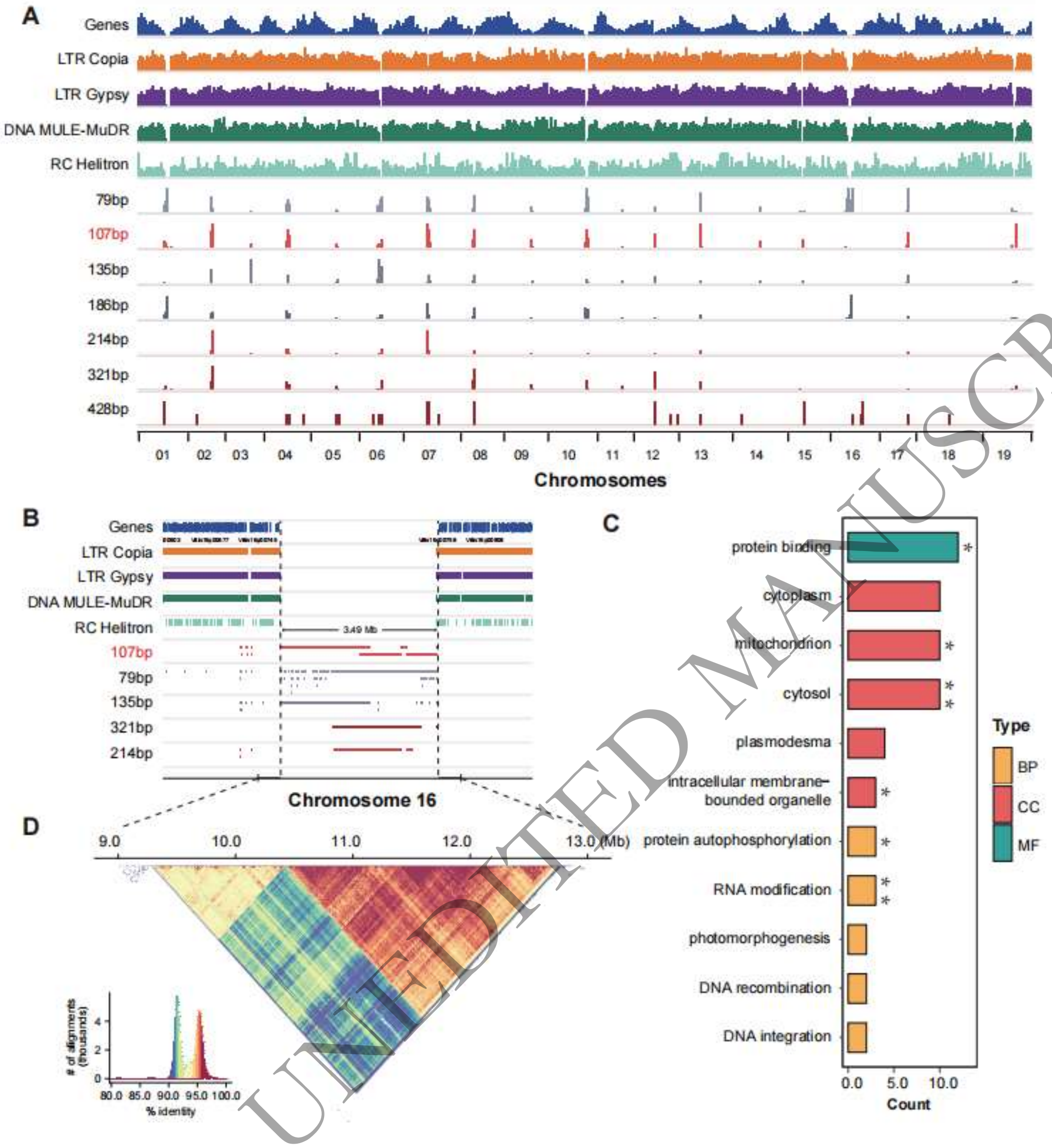

Figure 4. Characteristics and distribution of repeat unit copies in centromeres. (A) The distribution of genes, TEs and different repeat units in the whole genome. (B) Visualization of the predicted centromeric region on Chromosome 16 in IGV. (C) GO functional annotation of genes captured in centromeres. MF: molecular function, CC: cellular component, BP: biological process. Enrichment significant p-value: *, P< 0.05. **, P<0.001. (D) The triangle shows sequence similarity within each haplotype and colored by identity.

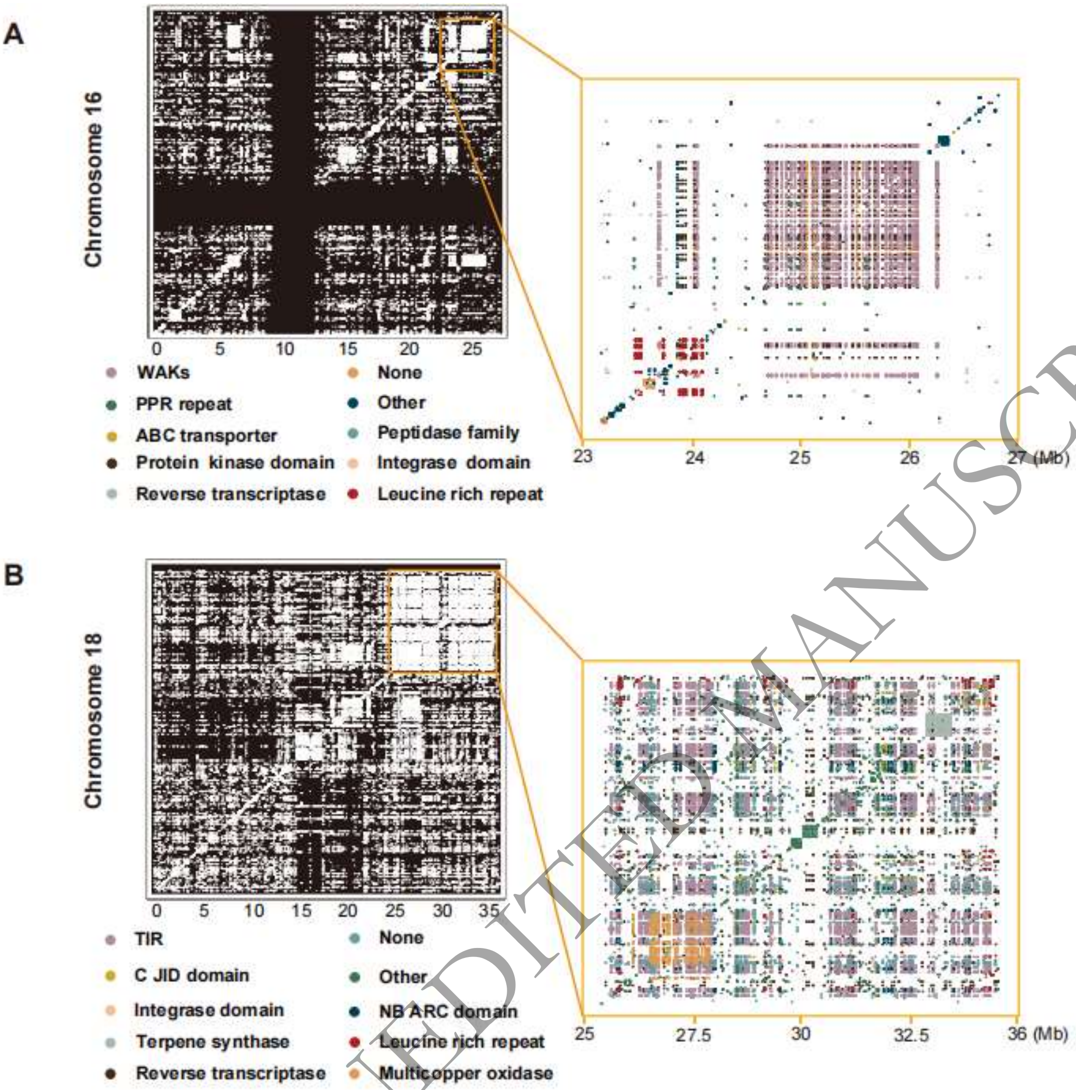

Figure 5. Schematic of identified gene clusters. (A) The gene clusters in Chromosome 16 and Chromosome 16: 22-27 Mb. **(B)** The gene clusters in Chromosome 18 and Chromosome 18:25-36 Mb. The graphs on the right were the enlargement of regions in white boxes on the left. Different color indicated the different gene clusters. Both split and compound.

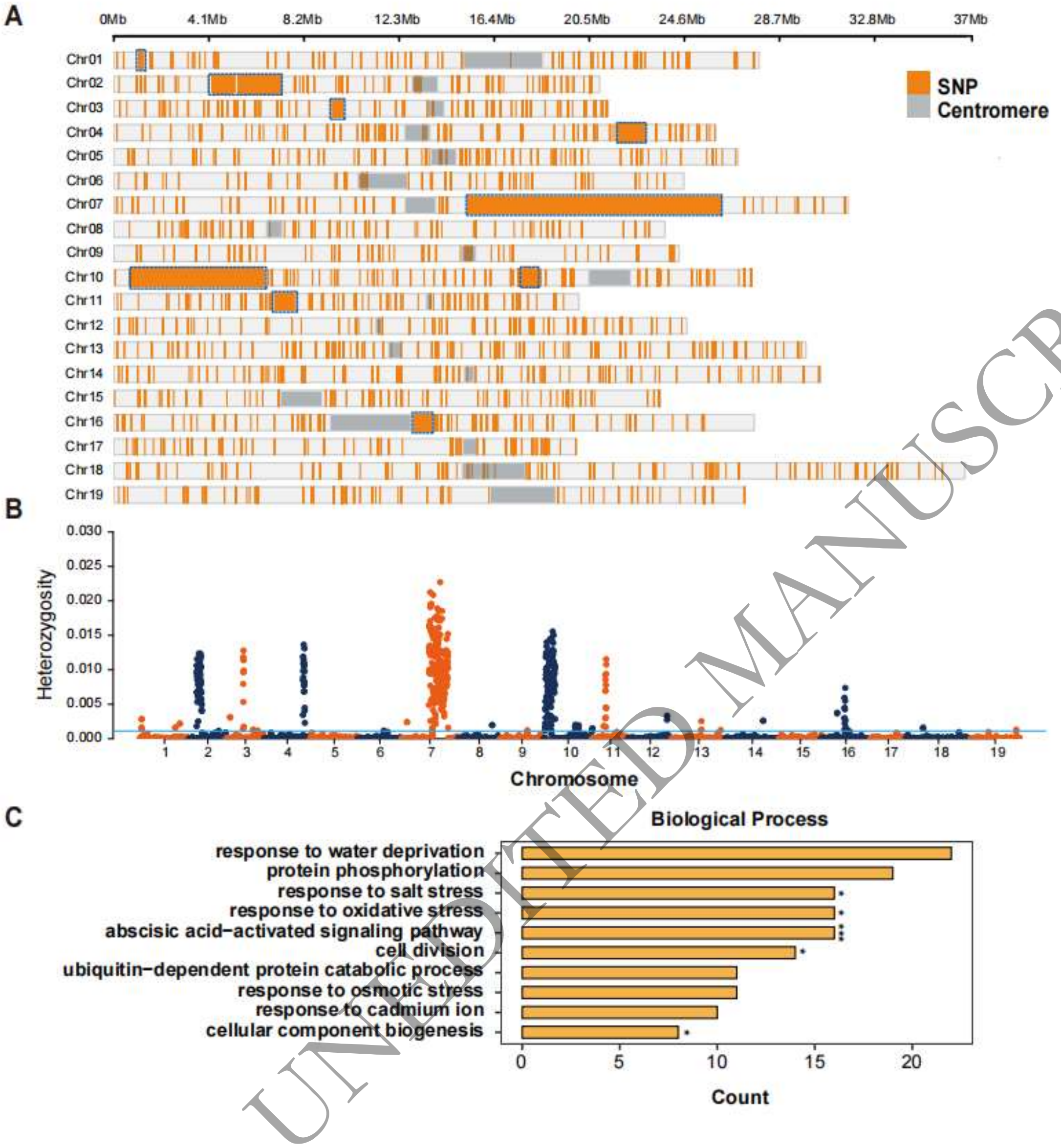

Figure 6. The characterization of heterozygous regions in PN40024. (A) The heterozygous sites were shared in all four PN40024 samples. The Grey bar indicated the centromere region while the orange lines indicated the heterozygous sites that existed in all samples. Blue boxes picked out the large heterozygous fragments. (B) The heterozygosity in PN40024 genome calculated with no overlapping 100 kb windows across four samples. (C) The GO enrichment analysis of genes contained heterozygous sites shown in A. Enrichment significant p-value: *, *P* < 0.05. **, *P* < 0.001. ***, *P* < 0.001.