An improved reference of the grapevine genome reasserts the origin of the PN40024 highly-homozygous genotype

Authors:

Amandine Velt[1°], Bianca Frommer[2°], Sophie Blanc[1], Daniela Holtgräwe[2], Éric Duchêne[1], Vincent Dumas[1], Jérôme Grimplet[3], Philippe Hugueney[1], Catherine Kim[4,5], Marie Lahaye[1], José Tomás Matus[6], David Navarro-Payá[6], Luis Orduña[6], Marcela K. Tello-Ruiz[4], Nicola Vitulo[7], Doreen Ware[4,8] and Camille Rustenholz[1]#
° equally contributed

#Corresponding author: Camille Rustenholz, camille.rustenholz@inrae.fr (C.R.)
Phone: 0033 3 89 22 49 81
Fax: 0033 3 89 22 49 33

Affiliations:

[1]SVQV, INRAE - University of Strasbourg, 68000 Colmar, France; amandine.velt@inrae.fr (A.V.); sophie.blanc@inrae.fr (S.B.); eric.duchene@inrae.fr (É.D.); vincent.dumas@inrae.fr (V.D.); philippe.hugueney@inrae.fr (P.H.); marie.lahaye@inrae.fr (M.L.); camille.rustenholz@inrae.fr (C.R.)

[2]Genetics and Genomics of Plants, CeBiTec & Faculty of Biology, Bielefeld University, 33615 Bielefeld, Germany; daniela.holtgraewe@cebitec.uni-bielefeld.de (D.H.); frommer@cebitec.uni-bielefeld.de (B.F.)

[3] Unidad de Hortofruticultura, Centro de Investigación y Tecnología Agroalimentaria de Aragón (CITA), 50059 Zaragoza, Spain; jgrimplet@cita-aragon.es (J.G.)

[4] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA; ckim@cshl.edu (C.K.); telloruiz@cshl.edu (M.K.T.-R.); ware@cshl.edu (D.W.)

[5] Haverford College, 370 Lancaster Avenue, Haverford, PA 19041, USA; ckim@cshl.edu (C.K.)

[6] Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, Paterna, 46908, Valencia, Spain; tomas.matus@uv.es (J.T.M.); david.navarro.paya@gmail.com (D.N-P.); luis.orduna.rubio@gmail.com (L.O.)

[7] Dipartimento di Biotecnologie, Università degli Studi di Verona, 37134, Verona, Italy; nicola.vitulo@univr.it (N.V.)

[8] USDA ARS NEA Robert W. Holley Center for Agriculture and Health, Agricultural Research Service, Ithaca, NY 14853, USA; ware@cshl.edu (D.W.)

ORCID:

Amandine Velt: 0000-0003-2368-839X

Bianca Frommer: 0000-0002-5792-0102

Sophie Blanc: 0000-0003-2501-548X

Daniela Holtgräwe: 0000-0002-1062-4576

Éric Duchêne: 0000-0003-2712-1892

Vincent Dumas: 0000-0001-6089-3048

Jérôme Grimplet: 0000-0002-3265-4012

Philippe Hugueney: 0000-0002-1641-9274

José Tomás Matus: 0000-0002-9196-1813

David Navarro-Payá: 0000-0003-1623-3039

Luis Orduña: 0000-0003-2756-4028

Marcela K. Tello-Ruiz: 0000-0002-7499-5368

Nicola Vitulo: 0000-0002-9571-0747

Doreen Ware: 0000-0002-8125-3821

Camille Rustenholz: 0000-0001-5355-3408

Running Head: PN40024.v4, an improved grapevine reference genome

Keywords: *Vitis vinifera*, genotype PN40024, reference genome, long reads, improved annotation

Abstract

The genome sequence of the diploid and highly homozygous *V. vinifera* genotype PN40024 serves as the reference for many grapevine studies. Despite several improvements to the PN40024 genome assembly, its current version PN12X.v2 is quite fragmented and only represents the haploid state of the genome with mixed haplotypes. In fact, being nearly homozygous, this genome contains several heterozygous regions that are yet to be resolved. Taking the opportunity of improvements that long-read sequencing technologies offer to fully discriminate haplotype sequences, an improved version of the reference, called PN40024.v4, was generated. Through incorporating long genomic sequencing reads to the assembly, the continuity of the 12X.v2 scaffolds was highly increased with a total number decreasing from 2,059 to 640 and a reduction in N bases of 88%. Additionally, the full alternative haplotype sequence was built for the first time, the chromosome anchoring was improved and the number of unplaced scaffolds was reduced by half. To obtain a high-quality gene annotation that outperforms previous versions, a liftover approach was complemented with an optimized annotation workflow for *Vitis.* Integration of the gene reference catalogue and its manual curation have also assisted in improving the annotation, while defining the most reliable estimation of 35,230 genes to date. Finally, we demonstrated that PN40024 resulted from nine selfings of cv. 'Helfensteiner' (cross of cv. 'Pinot noir' and 'Schiava grossa') instead of a single 'Pinot noir'. These advances will help maintain the PN40024 genome as a gold-standard reference, also contributing towards the eventual elaboration of the grapevine pangenome.

Introduction

Cultivated grapevine (*Vitis vinifera* ssp. *vinifera*) was the fourth plant whose genome was sequenced and assembled (Jaillon *et al.* 2007). Because of the grapevine's high level of heterozygosity (one Single Nucleotide Polymorphism (SNP) per 100 bp and one Indel per 450 bp, (Velasco *et al.* 2007)), the genotype selected for sequencing was PN40024, whose ~475 Mb genome (Lodhi and Reisch 1995) is nearly homozygous (estimated at ~93%). PN40024 was indeed generated through nine rounds of selfing and supposedly originated from 'Pinot noir', hence its identification as 'PN'. This unique genome characteristic allowed a high-quality whole-genome shotgun assembly based on 8X coverage Sanger reads (Jaillon *et al.* 2007). In 2009, a 4X coverage was added, which improved the overall coverage of the genome

(from 68.9% for the 8X version to 91.2% for the 12X.v0) (http://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences; FN597015-FN597047 at EMBL, release 102; File S1 Figure S1). In 2017, a third assembly version, named 12X.v2, was published as the result of a large anchoring effort using six dense parental genetic maps (Canaguier *et al.* 2017). Despite these advances, no additional sequencing efforts have been made and although it's of very high quality, the 12X.v0 Sanger contigs are numerous (14,642), the 12X.v2 scaffolds are composed of large N gaps (3.1% of the cumulative scaffold size) and the 19 pseudomolecules are quite fragmented (19.3 scaffolds on average per pseudomolecule).

In recent years, the advent of third generation sequencing technologies, especially those from the Pacific Biosciences (PacBio) platform, have allowed the assembly of grapevine diploid genomes with a higher level of contiguity compared to the 12X.v2 version of the PN40024 genome (for example, cv. 'Cabernet Sauvignon' genome assembly (Massonnet *et al.* 2020)).

Along with the versions of each genome assembly, several versions of gene annotations were made available (File S1 Figure S1). The first version of the grapevine genome assembly, 8X, was published along with the prediction of 30,434 gene models based on the GAZE software (Howe *et al.* 2002; Jaillon *et al.* 2007). For the 12X.v0, three different versions of gene predictions were made available: the v0 version (26,346 gene models), based on the GAZE software (Howe *et al.* 2002), the CRIBIv1 version (29,971 gene models), based on the JIGSAW software (Allen and Salzberg 2005), and the CRIBIv2 version (31,845 gene models), with an effort made on the discovery of splicing variants (Vitulo *et al.* 2014). For the 12X.v2, the International Grapevine Genome Program (IGGP) led the initiative of merging annotations from NCBI Refseq, CRIBIv1 and VCost, which was based on the Eugene software (Sallet *et al.* 2019) and was generated in the frame of the COST Action FA1106. This version, called VCost.v3, resulted in an exhaustive view of the PN40024 grapevine gene content with its 42,413 gene models (Canaguier *et al.* 2017). However, after several years as the reference annotation by the grapevine scientific community, it appeared that the great increase in number of gene models for VCost.v3 compared to all the previous annotation versions was caused by many small and fragmented predictions that were probably erroneous.

By combining the top-quality Sanger contigs from the 12X version and long reads generated here by Single-Molecule Real-Time (SMRT) sequencing (PacBio), we provide an improved version of the PN40024 genome sequence assembly, referred to as PN40024.v4. Along with

this new assembly, we also provide a new version of the gene annotation, PN40024.v4.2, based on a newly-developed annotation workflow, RNA-Seq datasets and an exhaustive manual curation of a set of catalogued genes of functional interest to the community. Finally, we demonstrate that PN40024 originates from selfings of the 'Helfensteiner' cultivar instead of 'Pinot noir'.

Methods & Materials

Plant material, DNA extractions and sequencing

DNA extractions of young leaves of cv. 'Pinot noir' clone 162 (ID code FRA038-193.Col.162), cv. 'Schiava grossa' (synonymous 'Trollinger', ID code FRA038-2525.Col.1) and cv. 'Helfensteiner' (ID code FRA038-2744.Col.1) were performed as described by Merdinoglu *et al.* (Merdinoglu *et al.* 2005). Illumina DNA PCR-Free Prep kit was used to prepare the resequencing libraries according to provider procedures. Paired-end Illumina HiSeq 4000 sequencing at about 15x coverage was performed for 'Pinot noir' and 'Schiava grossa', respectively. Paired-end Illumina NovaSeq 6000 sequencing at about 15x coverage was performed for 'Helfensteiner'.

One gram of young leaves (1cm²) of PN40024 (ID code FRA038-40024.Col.1) was collected and DNA was extracted using QIAGEN Genomic-tips 100/G kit. SMRT sequencing on a Sequel I machine (3 SMRTCells; PacBio) and dedicated library preparation were performed according to provider procedures.

Genotyping-by-sequencing (GBS) was performed on the population 'Riesling' x 'Gewurztraminer' (exhaustively described by Duchêne *et al.* (Duchêne *et al.* 2020) using the procedure described by Girollet *et al.* (Girollet *et al.* 2019).

All data generated in the frame of this study were submitted under the ENA Study Accession PRJEB45423.

Genome assembly

Raw SMRT reads (ERR7997743) were self-corrected using CANU (v.1.6) (Koren *et al.* 2017), followed by a correction with PN40024 Illumina reads (SRR8835144) using LORDEC

(v.0.5.3) (Salmela and Rivals 2014). The corrected reads were mapped on PN12X scaffolds (https://urgi.versailles.inra.fr/download/vitis/VV_12X_embl_102_Scaffolds.fsa.zip) using minimap2 (v2.17-r954-dirty) (Li 2018). A total of 163,446 reads (15%) were aligned on less than 80% of their length and/or with less than 80% identity and were thus considered as missing from PN12X scaffolds. These unmapped reads were assembled using Flye (v2.4-gc9db046) (Kolmogorov *et al.* 2019). We aligned these new contigs on the Uniprot *Arabidopsis* database (release 2019_01) using blastx (Altschul *et al.* 1990). Contigs longer than 5 kb and having hit(s) with *Arabidopsis* proteins with >60% identity and >60% length coverage, were selected for the next step. The fasta files of these new contigs and the PN12X scaffolds were concatenated to generate the new assembly. Firstly, the repeats were masked using Red (v05/22/2015) (Girgis 2015). Then, Haplomerger2 (v20180603) (Huang *et al.* 2017) was used following three steps according to developer procedures: i) break the misjoins and output the new diploid assembly; ii) separate/merge two haplotypes and output haploid assemblies (REF and ALT); and iii) remove tandem errors from haploid assemblies. Some scaffolds / contigs were deleted by Haplomerger2 during the assembly process but sequences longer than 10 kb were retrieved and added to the REF scaffolds. The two haploid assemblies (REF/ALT) were then scaffolded with the OPERA-LG tool (v2.0.6) (Gao *et al.* 2016), which uses both, corrected SMRT reads and Illumina reads. A first gap-filling step (two rounds) was carried out with Illumina reads using GapCloser (v1.12) (Luo *et al.* 2012) and a second gap-filling step (three rounds) was carried out with corrected SMRT reads using LR_Gapcloser (v1.0) (Xu *et al.* 2019). A final polishing step was performed with the Illumina reads using PILON (v1.23-1-g41e0b8e) (Walker *et al.* 2014) (Figure 1A and 1B).

The anchoring of the new haploid scaffolds was performed using the six genetic maps used for the same purpose by Canaguier *et al.* (Canaguier *et al.* 2017) and two new genetic maps from cv. 'Riesling' and cv. 'Gewurztraminer' derived from GBS. To transfer the markers from Canaguier *et al.* (Canaguier *et al.* 2017) from PN12X.v2 to the scaffolds of PN40024.v4, BLAST (v2.2.28) (Altschul *et al.* 1990) or ipcress (ipcress from exonerate v2.2.0) (Slater and Birney 2005) was used to align the markers and find the position of each on the scaffolds of PN40024.v4 REF and ALT. A total of 2,333 markers for REF and 2,326 markers for ALT were used from these six maps to anchor the scaffolds. For the two new genetic maps from 'Riesling' and 'Gewurztraminer', 5,884 ('Riesling') and 5,840 ('Gewurztraminer') SNP markers were available for REF and 5,866 ('Riesling') and 5,832 ('Gewurztraminer') for ALT. The SNP markers were derived from GBS data (ERR8657388 to ERR8657647) and were analyzed with

Fast-GBS (Torkamaneh *et al.* 2017) with modifications to allow paired-end read analysis (https://forgemia.inra.fr/sophie.blanc/gbs). The two genetic maps were built using R ASMap package with the "kosambi" parameter (Taylor and Butler 2017). A first run of Allmaps (v0.9.13) (Tang *et al.* 2015) was performed with the "merge" command to merge all genetic maps and then "split", "gaps", "refine" and "build" commands to create breakpoints (58 for REF scaffolds and 47 for ALT scaffolds), with default parameters. Subsequently, all maps were recreated for new scaffolds and then orientation and anchoring of new haploid scaffolds on the 19 pseudochromosomes were performed using Allmaps with the "merge" command to merge all maps and "path" command to anchor, with default parameters (Figure 1C).

Quality assessment of the PN40024.v4 genome sequence assembly

A quality analysis of the genome assembly was done with Merqury v1.3 (Rhie *et al.* 2020). Since PN40024 is a 'Helfensteiner' selfing (demonstrate below) and since 'Helfensteiner' originated from a cross between 'Pinot noir precoce' and 'Schiava grossa', 'Schiava grossa' was used as the maternal parent. The run was carried out on the scaffolds using genomic paired-end short reads of PN40024 as the child data (SRR8835144), short reads of 'Pinot noir' as the paternal data (ERR8014965) and short reads of cv. 'Schiava grossa' as the maternal data (ERR8014964). A k-mer database was built for the three read datasets with k=19, the Merqury hap-mer databases were computed and the PN40024.v4 genome assembly was evaluated using 'num_switch 100' and 'short_range 20,000'. For comparison reasons, the Merqury quality analysis was carried out on PN12X.v2 using the same k-mer databases.

The "*Flowering locus T*" (*FT*) and the "*Adenine phosphoribosyltransferase 3*" (*APRT3*) genes are absent and truncated in PN12X.v2, respectively. To check whether these genes could be retrieved in the new genome assembly, cDNA sequences of *FT* (NM_001280978.1) and *APRT3* (GSVIVT00007310001, PN8X version) were used to perform blastn (Altschul *et al.* 1990) against PN8X, PN12X, PN12X.v2 and PN40024.v4 genome assemblies. High Scoring Pairs (HSPs) were then accumulated for each analysis and the mean percentage identity, query overlap, hit query start and end were calculated.

PN40024 (SRR8835144), and the cultivars 'Silvaner Gruen' (SRR5891620), 'Cabernet Franc' (SRR5891774), 'Cabernet Sauvignon' (SRR5891776), 'Chardonnay' (SRR5891778), 'Muscat Hamburg' (SRR5891787), 'Semillon' (SRR5891866), 'Pinot Noir' (SRR5891886), 'Merlot'

(SRR5891890), 'Sauvignon Blanc' (SRR5891893), 'Muscat of Alexandria' (SRR5891985) and 'Riesling' (SRR5891989) genomic paired-end resequencing datasets were aligned against PN40024.v4 REF, PN12X.v2 and 'Cabernet Sauvignon' haplotype 1 (Massonnet *et al.* 2020) pseudomolecule assemblies (without chrUn or unplaced contigs / scaffolds) using bwa-mem2 (v2.0) (Vasimuddin *et al.* 2019) "mem" command with default parameters. Samtools' (v1.9) (Li *et al.* 2009) "flagstat" command was used with default parameters to compute alignment statistics.

PN12X scaffolds were mapped against PN40024.v4 REF pseudomolecules using NUCmer (MUMmer v3.1) (Kurtz *et al.* 2004) with "-maxmatch -l 100 -c 500" parameters. The output file was filtered using MUMmer show-coords command with "-l -g -I 99.5" parameters. The resulting file was formatted into BED format and merged with the bed file corresponding to N gap regions in the PN40024.v4 assembly. Pseudomolecule regions over 100 bp that did not correspond to either PN12X scaffolds or N gap regions were identified as 'newly assembled' PacBio long read-based regions.

The identification of variants between PN40024 paired-end Illumina resequencing (SRR8835144) and PN40024.v4 REF and ALT pseudomolecules was performed as described in section "Origin of PN40024". The homozygous calls "1/1" were considered as assembly errors. The densities of the heterozygous calls "0/1" along the REF and ALT pseudomolecules were used to define seven heterozygous regions of the PN40024 genome.

Origin of PN40024

PN40024 (SRR8835144), 'Pinot noir' (ERR8014965), 'Schiava grossa' (ERR8014964), 'Helfensteiner' (ERR8014963) and 'Araklinos' (SRR8835172) paired-end resequencing datasets were all analyzed using the same pipeline. Datasets were aligned against PN40024.v4 REF assembly using bwa-mem2 (v2.0) (Vasimuddin *et al.* 2019) "mem" command with default parameters. Samtools (v1.9) (Li *et al.* 2009) "view" and "sort" commands with default parameters were used to convert and sort the output BAM files. GATK (v4.1.4.0) (McKenna *et al.* 2010) "MarkDuplicatesSpark", "HaplotypeCaller" and "GenotypeGVCFs" commands with default parameters were used to generate variant files in VCF format. The GATK "VariantFiltration" command was used to filter out variants meeting at least one of the following criteria: QD < 8.0, QUAL < 100.0, FS > 60.0, SOR > 3.0, DP < 3, DP > 30, AD < 2. The final variant files were obtained using GATK "SelectVariants" command with "--exclude-filtered --exclude-non-variants" parameters. The homozygous SNP calls "1/1" were

selected for each analyzed genotype. All SNPs corresponding to a homozygous call in PN40024 genotypes were excluded from the analysis as they represent assembly errors. The remaining homozygous SNPs were used to draw density plots on the PN40024.v4 pseudomolecules. The regions that are rich in homozygous SNPs for a given genotype correspond to regions for which this genotype does not share a haplotype with PN40024.

The haplotypic blocks were defined after segmentation of homozygous SNP densities along the chromosomes using the R package changepoint (v2.2.2) (Killick and Eckley 2014) with command "cpt.mean" and the parameters method="PELT" and penalty="AIC". Some manual curation of the segments was performed to join directly adjacent segments of the same origin ('Pinot noir' or 'Schiava grossa'). The size of the segments was used to calculate the proportion of 'Pinot noir', 'Schiava grossa' and common haplotypes.

Gene prediction

Before performing gene prediction, the PN40024.v4 genome assembly was repeat masked with RepeatMasker v4.1.2 (Smit *et al.* 2013) using crossmatch as search engine. Predictions with a SW-Score <1,000 were filtered out and predictions with a Smith-Waterman (SW)-Score between 1,000 and 2,000 were only kept if the reported percentage of substitutions were <20%. The PN40024.v4 genome assembly was softmasked with BEDTool (v2.26.0) (Quinlan and Hall 2010).

To annotate the PN40024.v4 genome assembly, publicly available *V. vinifera* stranded (File S2 Table S1) and unstranded (File S2 Table S2) paired-end RNA-Seq datasets of different tissues and treatments were collected. RNA-Seq data were trimmed with Trimmomatic (v0.39) (Bolger *et al.* 2014). The annotation pipeline was first tested on the PN40024 12X.v0 genome assembly using VCost.v3 gene annotation as quality reference. The gene predictors SNAP (Korf 2004) and BRAKER2 (Hoff *et al.* 2016, 2019; Brůna *et al.* 2021) were trained and tested on the softmasked 12X.v0 genome assembly. The RNA-Seq data was mapped on 12X.v0 and on PN40024.v4 REF and ALT sequences with GMAP/GSNAP v2020-09-12 setting "-B 5 --novelsplicing 1" (Wu and Watanabe 2005). Primary mappings were extracted with SAMTools v1.9 (Li *et al.* 2009). Based on the primary mappings, stranded and unstranded reference-guided transcriptome assemblies were computed with PsiCLASS v1.0.1 using default parameters (Song *et al.* 2019).

Additionally, *A. thaliana* protein sequences (UniProt/SwissProt release 2020_02), eudicotyledone protein sequences (UniProt/SwissProt release 2020_02, OrthoDB10 v1), Viridiplantae and Vitales sequences (UniProt/SwissProt release 2020_02) were aligned on 12X.v0 and on PN40024.v4 REF and ALT with pBLAT v1.9 (Wang and Kong 2019), a parallel implementation of the original blat algorithm (Kent 2002). The genome regions on which the protein data mapped were extracted and the protein sequences were aligned to these regions with exonerate v2.4.0 (Slater and Birney 2005). Only the proteins that aligned on the reference genome with an identity of 25%, a similarity of 50% and with a sequence alignment coverage of at least 80%, were retained and included in the gene prediction.

The gene predictor GlimmerHMM v3.0.4 (Majoros *et al.* 2004) was trained on 12X.v0 and on PN40024.v4 REF and ALT using 7,500 (12X.v0) and 15,000 (PN40024.v4) random PsiCLASS transcripts of the 12X.v0 or PN40024.v4 REF or ALT stranded transcriptome assembly, respectively. The training was followed by gene prediction with GlimmerHMM with default settings.

Moreover, the gene predictor SNAP v2006-07-28 was trained on the 12X.v0 genome assembly. For this, the 12X.v0 genome assembly, the stranded transcriptome assembly, the Viridiplantae protein sequences and the eudicotyledone protein sequences were given to MAKER2 v3.01.03 (Holt and Yandell 2011; Campbell *et al.* 2014) and initial data alignment with BLAST (ncbi-blast-2.10.1+) (Altschul *et al.* 1990; Camacho *et al.* 2009) and exonerate was performed followed by MAKER2 *ab initio* gene prediction. MAKER2 was run with "max_dna_len=300000" and "split_hit=20000". A SNAP hmm file was generated with the MAKER2 gff file and a second MAKER2 run was performed with enabled SNAP gene prediction and the SNAP hmm file as input. Hmm file generation and SNAP gene prediction with MAKER2 and the new hmm file was repeated. The hmm file generated with the 12X.v0 assembly was used to run SNAP gene prediction on the PN40024.v4 REF and ALT genome sequences.

An AUGUSTUS species model was computed with BRAKER2 v2.1.5-master_20200915 and the 12X.v0 genome assembly. BRAKER2 was run with enabled softmasking and in *etpmode* calling GeneMark-ETP+ v4.61 (Lomsadze *et al.* 2005, 2014; Brůna *et al.* 2020) for initial gene prediction followed by AUGUSTUS training and gene prediction (AUGUSTUS version master_v3.3.3_20200914) (Stanke *et al.* 2006, 2008). With BRAKER2, the programs DIAMOND v0.9.24.125 (Buchfink *et al.* 2015), SAMtools v1.9-180-gf9e1caf (Li *et al.* 2009),

SPALN version 2.3.3f (Gotoh 2008; Iwata and Gotoh 2012), ProtHint version 2.5.0 and BamTools v2.5.1 (Barnett *et al.* 2011) were called. The stranded RNA-Seq primary mappings, the eudicotyledon protein sequences (OrthoDB10 v1) and the Viridiplantae protein sequences were used as input. The gene prediction on PN40024.v4 REF and ALT was performed with BRAKER2 v2.1.5-master_20210218, the generated AUGUSTUS species model and AUGUSTUS version master_v3.4.0_20210218. Again, the stranded RNA-Seq mappings and the same protein sequences were used as input. The BRAKER2 parameter settings were left the same as above.

The last *ab initio* gene prediction was done on the PN40024.v4 genome assembly with GeneID v1.4.5-master-20200916 and the publicly available *V. vinifera* parameter set using default settings. To add the VCost.v3 gene annotation to the set of predictions, an annotation liftover was performed with liftoff v1.5.1 (Shumate and Salzberg 2021) with default parameters onto the PN40024.v4 genome assembly.

To combine the predictions and evidence data into an overall gene model set, the GlimmerHMM, SNAP, BRAKER2 and GeneID *ab initio* gene prediction as well as the lifted VCost.v3 annotation, the stranded and unstranded transcriptome assemblies, the GFF file with the aligned protein data, the repeat annotation GFF file and the PN40024.v4 genome assembly was given to EvidenceModeler v1.1.1 (Haas *et al.* 2008). The used weights are listed in File S2 Table S3.

Subsequently, the raw gene models were quality filtered. Gene models only supported by *ab initio* predictors were kept if at least two gene prediction programs predicted them, if the start and stop codon was present and if the gene length was equal or larger than 300 bp. However, *ab initio* supported gene models not matching these constraints were kept if they had a database hit with the UniProt/SwissProt or NCBI non-redundant database. To obtain that, a blastp search of the protein sequences against the two databases was run, allowing hits with an e-value $<1e^{-6}$. Of the gene models only supported by evidence data or by VCost.v3 lifted annotation, those gene models with missing start and stop and a gene length <300 bp were discarded.

The gene models generated by EvidenceModeler were finally processed by PASA (v2.4.1, default parameters) using the stranded transcriptome assembly as a reference to add UTR regions and to calculate alternatively spliced models. Genes with overlapping UTRs were

shortened. tRNAs were predicted with tRNAscan-SE-2.0 (Chan *et al.* 2021) on the PN40024.v4 genome assembly.

To retain gene naming of VCost.v3 gene models, a reciprocal best blast hit (RBH) search between protein sequences of PN40024.v4.1 gene models and protein sequences of VCost.v3 gene models was carried out. For the RBH search, only the longest protein sequence per gene was used, the e-value was set to $1e^{-4}$ and the query coverage and identity was set to 70%. Moreover, only RBHs with genes on the same pseudochromosome and showing collinearity with other genes were considered valid. Thus, genes with a valid RBH were named according to the VCost.v3 gene, novel genes received the prefix '04' at the start of the gene number and genes predicted for alternative heterozygous sequence regions received the suffix '_alt' (File S2 Table S4).

The PN40024.v4.1 gene models were functionally annotated with Blast2GO (v1.5.1) (Conesa *et al.* 2005; Götz *et al.* 2008). For this, protein domains of the PN40024.v4.1 proteins were identified with InterProScan (v5.52-86.0) (Jones *et al.* 2014) with options "--goterms --pathways -dp" using the databases/tools CDD-3.18 (Lu *et al.* 2020), Coils-2.2.1 (Lupas *et al.* 1991), Gene3D-4.3.0 (Sillitoe *et al.* 2019), Hamap-2020_05 (Pedruzzi *et al.* 2013), MobiDBLite-2.0 (Necci *et al.* 2017), PANTHER-15.0 (Mi *et al.* 2021), Pfam-33.1 (Mistry *et al.* 2021), PIRSF-3.10 (Wu *et al.* 2004), PIRSR-2021_02, PRINTS-42.0 (Attwood *et al.* 2012), ProSitePatterns-2021_01, ProSiteProfiles-2021_01 (Sigrist *et al.* 2013), SFLD-4 (Akiva *et al.* 2014), SMART-7.1 (Letunic and Bork 2018), SUPERFAMILY-1.75 (Gough *et al.* 2001; Wilson *et al.* 2009), TIGRFAM-15.0 (Haft *et al.* 2013). PN40024.v4.1 protein sequences were aligned with diamond "blastp" (v2.0.11) (Buchfink *et al.* 2015) to the NCBI nr database (nr.07_07_2021.fasta) with options "--sensitive --top 5 -e 1e-5 -f 5". InterProScan and diamond results were used as input for Blast2GO.

Quality assessment of the PN40024.v4.1 gene annotation

To estimate completeness of the PN40024.v4.1 gene model set, plant core genes were predicted with BUSCO v5.1.2 (Simão *et al.* 2015; Waterhouse *et al.* 2018) using database eudicots_odb10.

Samples previously analyzed by Palumbo *et al.* (Palumbo *et al.* 2014) were used to perform differential gene expression analysis by using either PN12X.v2 assembly with VCost.v3 annotations or PN40024.v4 assembly with PN40024.v4.1 annotations. We analyzed cv.

'Sangiovese' (SRR1631822; SRR1631823; SRR1631824), cv. 'Barbera' (SRR1631834; SRR1631835; SRR1631836) and cv. 'Refosco' samples (SRR1631858; SRR1631859; SRR1631860) for the stage "Berries beginning to touch" (~EL35 according to Eichhorn and Lorenz phenological scale (Eichhorn and Lorenz 1977)). The RNA-Seq data were downloaded from the SRA with SRA Toolkit (v2.10.8) (SRA Toolkit Development Team; https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) and analyzed with an *in-house* pipeline using FASTQC (v0.11.5) (Andrews 2010), STAR (v2.5.3a) (Dobin *et al.* 2013), Samtools (v1.4.1) (Li *et al.* 2009), Bamtools (v2.4.0) (Barnett *et al.* 2011), featureCounts (v1.5.3) (Liao *et al.* 2014) and SARTools (v1.7.3) (Varet *et al.* 2016).

Manual gene model curation

For manual gene model curation, an Apollo Webserver v2.6.4 (https://github.com/GMOD/Apollo/blob/master/README.md) (Dunn *et al.* 2019) was set up for the PN40024.v4 genome assembly and provided with different data tracks such as PN40024.v4.1 and previous gene annotations, RNA-Seq mappings and exonerate protein mappings (see section Material and Methods - Gene prediction). By these means, gene models were manually inspected and curated if needed or also new genes were added following dedicated guidelines offered to the community (https://integrape.eu/resources/data-management/). Using Apollo, the plant core genes classified as fragmented or missing by BUSCO were manually curated and adapted if necessary. In the frame of this study, we also began to manually curate genes present in the grape reference catalogue ((Navarro-Payá *et al.* 2022); https://grapedia.org/genes/). A home-made python script was used to generate the PN40024.v4.2 version of gene annotations including those manually curated (https://gitlab.com/MSVteam/pn40024-visualization-tools/-/tree/master/update_gff3_script).

Results and Discussion

Improved metrics for the genome assembly of PN40024

A hybrid strategy was developed to assemble the genome of PN40024 genotype using 27X of SMRT long reads along with the PN12X scaffolds and 15X PN40024 Illumina paired-end resequencing data (Figure 1). This new assembly was named PN40024.v4. Six hundred and

forty scaffolds were produced with a N50 size of 6.5 Mb for a cumulative size of 474.5 Mb for the PN40024.v4 REF haplotype (Table 1). Compared to the former PN12X.v2, the number of scaffolds was reduced by a factor of three and the N50 was doubled. Moreover, the number of unknown bases, marked as N in the new scaffold sequences, represents 1.8 Mb and 0.4% of the assembly size versus 15.0 Mb and 3.1% for PN12X.v2 scaffolds. Thus, PN40024.v4 REF is more contiguous and has more informative sequences than PN12X.v2. Also, the PN40024.v4 assembly size is closer to the grapevine genome size of 475 Mb, estimated using flow cytometry by Lodhi and colleagues (Lodhi and Reisch 1995). Phasing efforts on the partially heterozygous genotype resulted in the reconstruction of the second PN40024 haplotype (PN40024.v4 ALT) with 485 scaffolds and a total genome assembly size of ~463 Mb (File S2 Table S5). Thus, the PN40024.v4 genome assembly now represents both haplotypes of the diploid PN40024 genome.

There are 7,640 newly assembled PacBio long read-based regions that were identified as missing from PN12X.v2 scaffolds. Their cumulative size is 24.1 Mb, *i.e.* 5.1% of the total PN40024.v4 genome assembly size (average = 3,152 bp; median = 558 bp; max = 32,650 bp).

A total of 2,333 markers were used from the six Canaguier's maps (Canaguier *et al.* 2017), in addition to 5,884 and 5,840 SNP markers from cv. 'Riesling' and cv. 'Gewurztraminer' GBS maps, respectively, to anchor these scaffolds. We were able to anchor 165 PN40024.v4 REF scaffolds to the 19 pseudochromosomes, for a cumulative size of ~462 Mb (97.4%) (Table 1). The 19 PN40024.v4 REF pseudomolecules are composed of 8.7 scaffolds on average (min = 3; max = 26; median = 6) whereas 19.3 scaffolds on average composed the PN12X.v2 pseudomolecules (min = 5; max = 82; median = 13). The remaining unplaced scaffolds were ordered according to their size to generate "chrUn" sequence representing 12.5 Mb (-47% compared to PN12X.v2 unplaced scaffolds). Thus, PN40024.v4 anchoring was improved as the pseudomolecules are less fragmented and as the size of chrUn has almost been halved.

At the chromosome scale, 10 pseudomolecules became shorter in PN40024.v4 compared to PN12X.v2 (average loss = ~448 kb; median = ~255 kb; min = 2,961 bp; max = 1,133,439 bp). Chromosome 6 showed the biggest reduction as the location of a large fragment has been correctly assigned to chromosome 9 (File S1 Figure S2). Nine pseudomolecules became larger (average gain = ~869 kb; median = ~582 kb; min = 15,118 bp; max = 2,045,414 bp), notably chromosome 9, 7 and 15, which gained 1.5 Mb, 1.9 Mb and 2.0 Mb, respectively.

By aligning PN40024 Illumina paired-end reads against PN40024.v4 genome assembly, we identified 101,778 heterozygous variations. Using their density along the chromosomes, we were able to identify seven well-defined heterozygous regions in PN40024.v4 genome assembly as it was the first time that a software dedicated to diploid assembly (Haplomerger2) was used to assemble the PN40024 genome. These regions were located on chromosomes 2, 3, 4, 7, 10, 11 and 15 with the two largest regions being on chromosome 7 and 10 (11.4 Mb and 5.5 Mb, respectively) (Figure 3). Their overall cumulative size of 20.6 Mb represents 4.3% of PN40024.v4, which is less than the residual heterozygosity size of 7%, estimated by Jaillon and colleagues based on genetic markers (Jaillon *et al.* 2007). Using the same procedure, we identified six heterozygous regions in PN12X.v2 assembly on the same chromosomes as PN40024.v4 except for the one on chromosome 15. Their overall cumulative size of 16.6 Mb represents 3.4% of PN12X.v2 and 4 Mb less than the heterozygous regions anchored on the PN40024.v4 chromosomes. These sequences were badly resolved and mostly located in the unanchored fraction of PN12X.v2 assembly (File S1 Figure S2). Thus, we conclude that PN40024.v4 is a better diploid assembly compared to PN12X.v2.

Quality of the PN40024.v4 genome assembly

The BUSCO analysis performed on the PN40024.v4 genome assembly confirmed that the gene space was more complete with 98.1% of the 2,326 total searched Eudicots BUSCO genes being complete, compared to PN12X.v2 with 97.6% (Figure 6). The *FT* (*Flowering locus T*) gene is conserved among all flowering plants as it promotes transition from vegetative growth to flowering. However, its sequence could only be found on an unanchored scaffold in the PN8X version and was totally missing in PN12X.v0 and PN12X.v2. It is now present on chromosome 7 of the PN40024.v4 assembly and also on its allelic region, chromosome 7_ALT sequence. Similarly, the *APRT3* gene, located in the sex determination locus of grapevine, was present on chromosome 2 in the PN8X version and was truncated in PN12X.v0 and PN12X.v2. It is now fully retrieved on chromosome 2 of PN40024.v4 assembly and on its allelic region, chromosome 2_ALT sequence. These two examples, along with the BUSCO analysis, show that the PN40024.v4 assembly is more complete, especially in the residual heterozygous regions that are now more accurately exposed.

The alignment metrics of PN40024 genomic Illumina paired-end reads have always been better against PN40024.v4 compared to PN12X.v2, either for overall percentage of mapped reads (97.58% versus 96.58%) or for properly mapped pairs of reads (85.81% versus 82.82%)

(Figure 2). This confirms that the PN40024.v4 assembly is more complete and with a more accurate structure than PN12X.v2. Moreover, we compared alignments of 11 genomic Illumina paired-end read datasets from various cultivars against PN40024.v4 and PN12X.v2 assemblies, but also against 'Cabernet Sauvignon' (Massonnet *et al.* 2020) haplotype 1, whose assembly metrics and technology were similar to PN40024.v4. Again, PN40024.v4 performs best for each dataset, even when 'Cabernet Sauvignon' was aligned against its own assembly (Figure 2). These results confirm that PN40024.v4 shows a quality suitable to become the new grapevine reference genome assembly, as it performs well with aligning genomic reads of various *V. vinifera* cultivars.

The error rate at nucleotide level was assessed by calling homozygous variations between PN40024 genomic Illumina paired-end reads aligned against the PN40024.v4 genome assembly. We identified 28.7 errors / Mb compared to 8.4 errors / Mb in the PN12X.v2 genome assembly. However, they are unevenly distributed along the chromosomes and they mostly co-localize with the newly assembled long read-based regions and the seven heterozygous regions (Figure 3). A higher density of errors was also detected in the heterozygous regions of the PN12X.v2 genome assembly (File S1 Figure S3). We detected 284.4 errors / Mb in PN40024.v4 heterozygous regions of and 83.1 errors / Mb in PN12X.v2 heterozygous regions, which is, respectively, about 10 times denser than their average error rate. Thus, the overall increase of error rate in the PN40024.v4 assembly is mostly due to the use of SMRT long reads to improve the completeness of the reference genome assembly.

Using Merqury, the base level quality value (QV) of the PN40024.v4 genome assembly was estimated to be 36.02, which is slightly worse than QV of 37.43 of the PN12X.v2 genome assembly (Table 2). This result confirms that additional SMRT sequences are not as accurate as Sanger-based sequences and they slightly decrease overall accuracy of the assembly. Also, the error rate of the PN40024.v4 genome assembly was increased by 0,00006964% compared to PN12X.v2, but still represents an accuracy of 99.999749801%, a metric associated with high-quality genome assemblies.

Nevertheless, the k-mer completeness was raised from 96.79% to 96.96% for the PN40024.v4 assembly. Based on k-mer profiles of PN40024 and its parents (see "The origin of the PN40024 genotype" section for details), Merqury computed the inheritance spectrum (File S1 Figure S4) showing a low portion of read-only missing k-mers that are unique for the child read set (paired-

end short reads of PN40024). The few missing sequences are probably due to sequencing errors, k-mers of novel variations or contamination from microbiome in PN40024 short reads, indicating an almost fully-complete PN40024.v4 genome sequence assembly. Also, as the spectrum shows a single 2-copy peak around 12x and that no 1-copy peak was observed at half the size, the k-mer analysis supports the assumptions of an almost homozygous grapevine genotype.

The origin of the PN40024 genotype

So far, the PN40024 genotype was supposed to be originally derived from cv. 'Pinot noir' (Jaillon *et al.* 2007). However, we found 1,415,200 homozygous variants between 'Pinot noir' and PN40024.v4 (versus 17,696 homozygous variants of PN40024 against its own assembly), meaning that 'Pinot noir' haplotypes were completely missing at these locations. These homozygous 'Pinot noir' variants were unevenly distributed along the chromosomes and formed blocks (Figure 4). We identified that the haplotypes of unknown origins could be assigned to 'Schiava grossa' (synonyms: 'Trollinger' and 'Frankenthal') as already suspected by Jaillon and colleagues (Jaillon *et al.* 2007). There were 953,735 homozygous variants found between cv. 'Schiava grossa' and PN40024.v4 and the formed haplotype blocks were highly complementary to 'Pinot noir' haplotype blocks (Figure 4). As a negative control, the same analysis was performed with cv. 'Araklinos' and 2,273,888 homozygous variants were identified, evenly distributed along the chromosomes (File S1 Figure S5).

Using Merqury, only a small portion of hap-mer specific k-mers (parental specific k-mers of the assembled F1) were found in the PN40024.v4 genome assembly (File S1 Figure S4). With the use of read data from both parents and child, Merqury was able to compute haplotype blocks by using the parental specific k-mers as anchors. A total of 1,454 haplotype blocks were computed for PN40024.v4 sequences with additional 289 haplotype blocks for alternative heterozygous sequence regions and 2,575 haplotype blocks for the 12X.v2 genome assembly (Table 3). The N50 was measured to 2.05 Mb (REF), 0.25 Mb (ALT) and 1.76 Mb (PN12X.v2). Compared to the PN12X.v2 genome assembly, PN40024.v4 presented less haplotype blocks, but comprised almost all bases showing a higher N50 value, *i.e.* its haplotype blocks are more contiguous.

A greater amount of paternal ('Schiava grossa') than maternal ('Pinot Noir') specific k-mers were identified. After identifying the origin of each haplotype block using segmentation, it is estimated that 41% of the genome harbours a 'Schiava grossa'-specific haplotype and 27% a 'Pinot noir'-specific haplotype. It is estimated that 32% of the genome shares a common haplotype between the two parents, *i.e.*, that these regions could originate either from 'Pinot noir' or 'Schiava grossa' indicating that ~57% could originate from 'Schiava grossa' and ~43% from 'Pinot noir'.

The switch error rate was determined to 0.96% (REF), to 4.76% (ALT) and to 0.77% (PN12X.v2). Some of the switches are probably due to sequencing errors in the additional long read-based sequences. Moreover, as the error rate of ALT sequences was measured to ~4.76%, portions of the alternative sequences are a mixture of the maternal and paternal haplotype, confirming that despite the improved separation of the two haplotypes in PN40024.v4, phasing is still not perfect.

By exploring the VIVC database (www.vivc.de), the 'Helfensteiner' cultivar was found to originate from a cross between 'Pinot noir precoce' (a clone of 'Pinot noir') and 'Schiava grossa'. By performing the same variant calling analysis, 53,671 homozygous variants were found between cv. 'Helfensteiner' and PN40024.v4, with 543 homozygous variants / Mb in the heterozygous regions and 93 homozygous variants / Mb in the homozygous regions (Figure 5). As a negative control, 'Araklinos' showed 3,967 homozygous variants / Mb in the heterozygous regions and 4,818 homozygous variants / Mb in the homozygous regions). Thus, the 'Helfensteiner' homozygous variants are almost six times denser in error-prone regions of the PN40024.v4 assembly, which makes them probable "false positive" homozygous variants. Apart from heterozygous regions, no blocks of homozygous variants could be identified, meaning that one of the two 'Helfensteiner' haplotypes is always present in the PN40024 genome. This confirms that the 'Helfensteiner' variety is the true parent of the first selfing, from which the PN40024 genotype was created after eight more selfings.

PN40024.v4.1 gene prediction, functional annotation and manual curation

The PN40024.v4.1 gene annotation of REF haplotype comprises 35,922 gene models of which 35,197 are protein-coding and 725 encode for tRNAs (Table 4). In particular, 1,572 novel protein-coding genes were annotated in the newly assembled long read-based regions. For heterozygous regions, 1,855 and 1,809 protein-coding genes were predicted for REF and ALT

haplotypes, respectively (Table 5). Most genes were predicted on the ~11 Mb heterozygous region on chromosome 7 with 830 on the reference sequence and 792 on the alternative sequence followed by the ~5 Mb region on chromosome 10 with 650 and 623 protein-coding genes.

To check for completeness of the gene models, the plant core genes of the database eudicots_odb10 were predicted with BUSCO (Figure 6). Of the 2,326 searched plant core genes, 2,296 or 98.7% were classified as complete in the PN40024.v4.1 gene annotation. Only 16 were predicted as fragmented and only 14 were not found.

Compared to PN12X.v2 VCost.v3 gene annotation, PN40024.v4.1 counts less predictions (41,182 versus 35,197) but their size is longer on average (4,485 bp versus 4,742 bp) (Table 4). Also, the BUSCO analysis performed on VCost.v3 showed that 2,257 or 97.0% were classified as complete (Figure 6). Thus, PN40024.v4.1 gene annotation represents PN40024 gene space in a more exhaustive and less fragmented manner compared to VCost.v3.

To help the community in the transfer of information across versions (*i.e.*, correspondences), we retained as many gene names from VCost.v3 in PN40024.v4.1 as possible. We adopted a strategy based on RBHs followed by some filtering steps which allowed us to transfer names for 66% (23,206) of PN40024.v4.1 gene models with the nomenclature VitviXXg0YYYY (XX being the chromosome number and YYYY a sequential number below 4,000). One third (11,991) of PN40024.v4.1 gene models could not be named with a VCost.v3 identifier and were named with the nomenclature VitviXXg0**4**ZZZ (XX being the chromosome number and ZZZ a sequential number below 1,000). The detailed nomenclature for PN40024.v4.1 gene annotations is given in File S2 Table S4.

The functional annotation of PN40024.v4.1 was performed using Blast2GO and resulted in at least one Gene Ontology term for 87% (30,689) of the genes and one Enzyme Code for 41% (14,512) of them. The main classes and ontologies are detailed in File S1 Figure S6 and Figure S7.

A subset of the RNA-Seq data published by Palumbo *et al.* (Palumbo *et al.* 2014) was used to compare the results of a differential gene expression analysis performed with PN12X.v2/VCost.v3 and PN40024.v4/PN40024.v4.1. In terms of mapping, the percentage of aligned reads was equivalent or slightly better when using PN40024.v4 genome assembly compared to PN12X.v2 (File S2 Table S6). Additionally, the percentage of assigned reads, *i.e.*,

the percentage of reads aligned under an annotated gene, was 2.4 to 3% better with PN40024.v4/PN40024.v4.1 compared to PN12X.v2/VCost.v3, which confirms the improved quality of PN40024.v4.1 gene annotation. Moreover, after differential gene expression analysis, the use of PN40024.v4/PN40024.v4.1 allowed the identification of more differentially expressed genes than PN12X.v2/VCost.v3 (File S1 Figure S8). This result along with the exhaustive functional annotation of PN40024.v4.1 shows that this new version of the PN40024 reference genome and annotation is a very efficient resource to perform transcriptomics and functional enrichment analyses.

Despite marked improvement of the PN40024.v4.1 automated annotation with respect to the previous VCost.v3 annotation, some recently expanded gene families have not been comprehensively annotated, such as the stilbene synthase (STS) gene family. Therefore, 1,641 genes (1,579 edited and 62 deleted) were manually curated using a purpose-built Apollo server (http://138.102.159.70:8080/apollo) providing a wide range of transcriptomic and genomic data for PN40024.v4. In an effort to preserve previous VCost.v3 manual curation and functional annotation efforts, a particular focus was given to genes present in the reference catalogue (Navarro-Payá et al, 2022). The PN40024.v4.1 automated annotation including the manually curated features was called PN40024.v4.2, which metrics are presented in (Table 4). An automated annotation from PN40024.v4.1 that was manually curated was deleted and replaced by its curated version in PN40024.v4.2. Also, the same rules were applied for gene name transfer and nomenclature for PN40024.v4.1 and PN40024.v4.2. The BUSCO analysis performed on PN40024.v4.2 shows that the fragmented plant core genes were reduced to six and the missing genes to eight (Figure 6). Thus, PN40024.v4.2 gene models comprise 2,308 or 99.2% complete plant core genes.

In conclusion, the here provided PN40024.v4 assembly is the most suitable grapevine reference genome sequence assembly as it notably outperforms PN12X.v2. In terms of genomic and transcriptomic read mapping, the assembly also outperforms other high-quality *V. vinifera* genome assemblies, something that occurs even when reads from these recently sequenced cultivars are used. Having a fully resolved alternative haplotype sequence, more continuous sequences and resolving many up-to-now unknown bases, PN40024.v4 represents the near complete diploid genome of the PN40024 genotype. Despite many improvements and advances in PN40024.v4, the genome sequence is still not perfect in regard to haplotype switching and newly introduced errors by implementation of long genomic reads. Further improvements

should focus on these regions. Nevertheless, the gene annotation of PN40024.v4 should be used as the most updated resource for transcriptomics and functional enrichment analyses, while the genes of heterozygous regions that are likely represented on both haplotypes will allow exploration of heterozygous genetic traits.

Data availability statement

Supplemental files are provided with the manuscript. File S1 contains additional figures and File S2 additional tables. Raw sequencing data and the PN40024.v4 genome assembly are available at ENA under BioProject PRJEB45423. Also, the PN40024.v4 genome assembly with structural and functional gene annotation is available on the INTEGRAPE website (https://integrape.eu/resources/genes-genomes/genome-accessions), on the Grape Genomics Encyclopedia portal (http://grapedia.org/) and under the DOI number doi:10.57745/F9N2FZ (https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/F9N2FZ). A Sequence Server v2.0.0 interface (http://138.102.159.70:4567/) was set up to perform BLAST analyses. A JBrowse interface (http://138.102.159.70/jbrowse/) was set up to visualize PN40024.v4 assembly and PN40024.v4.1 and v4.2 annotations, but also some previous annotation versions that were transferred, some RNA-Seq alignments and miscellaneous tracks. An Apollo interface (http://138.102.159.70:8080/apollo; training and account mandatory) was set up to manually curate gene annotations according to the dedicated guidelines (https://integrape.eu/resources/data-management/). Code used to analyze GBS data can be found at https://forgemia.inra.fr/sophie.blanc/gbs and code used to generate the PN40024.v4.2 version can be found at https://gitlab.com/MSVteam/pn40024-visualization-tools/-/tree/master/update_gff3_script.

Acknowledgements

We thank INRAE department “Biologie et Amélioration des Plantes” for funding; experimental unit UEAV (INRAE, Colmar) for plant maintenance; Anne-Marie Digby (University of Verona) for English correction; Emilce Prado for plant DNA extractions; EPGV (INRAE, Evry) for library prep and DNA sequencing; CNRGV (INRAE, Toulouse) for long read DNA extractions; Gentyane platform (INRAE, Clermont-Ferrand) for SMRT sequencing; Dr.

Timothée Flutre and Amandine Launay for their help in coordinating the FruitSelGen project and in acquiring GBS data; Mario Pezzotti, Anne-Françoise Adam-Blondon, Michele Morgante, Gabriele Di Gaspero and Gabriele Magris for helpful discussions throughout the project; Pablo Carbonell-Bejerano (Instituto de Ciencias de la Vid y el Vino - ICVV) for critically reviewing the manuscript. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI). This article is based upon work from COST Action CA17111 INTEGRAPE, supported by COST (European Cooperation in Science and Technology).

Contributions

A.V. performed genome assembly and functional annotation. B.F. performed gene prediction. B.F., N.V. and D.H. built the annotation pipeline. S.B. performed GBS analysis and built the genetic maps. É.D. built the genetic map pipeline. A.V., B.F. and C.R. performed the analysis for quality assessments. C.R. performed the analysis on PN40024's origin. V.D. performed plant material management and sampling. J.G. and B.F. worked on the gene name transfer. A.V., M.L. and D.N.-P. built the online tools. A.V., B.F., D.H., J.G., C.K., J.T.M, D.N.-P., L.O., M.K.T.-R., D.W. and C.R. worked on gene manual curation and writing of the dedicated guidelines. É.D., P.H. and C.R. looked for funding. C.R. supervised the project. A.V., B.F., D.H. and C.R. drafted and formatted the manuscript. All the authors read and helped improve the manuscript.

Conflict of interests

We declare no conflict of interests.

Funder Information

This work was supported by INRAE department "Biologie et Amélioration des Plantes", by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and by COST (European Cooperation in Science and Technology).

Literature Cited

Akiva, E., S. Brown, D. E. Almonacid, A. E. Barber 2nd, A. F. Custer *et al.*, 2014 The Structure–Function Linkage Database. Nucleic Acids Res. 42: D521–D530.

Allen, J. E., and S. L. Salzberg, 2005 JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21: 3596–3603.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Andrews, S., 2010 FASTQC. A quality control tool for high throughput sequence data.

Attwood, T. K., A. Coletta, G. Muirhead, A. Pavlopoulou, P. B. Philippou *et al.*, 2012 The PRINTS database: A fine-grained protein sequence annotation and analysis resource-its status in 2012. Database 2012: bas019.

Barnett, D. W., E. K. Garrison, A. R. Quinlan, M. P. Strömberg, and G. T. Marth, 2011 BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinforma. Oxf. Engl. 27: 1691–1692.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Brůna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinforma. 3: lqaa108.

Brůna, T., A. Lomsadze, and M. Borodovsky, 2020 GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics Bioinforma. 2:.

Buchfink, B., C. Xie, and D. H. Huson, 2015 Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12: 59–60.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Campbell, M. S., C. Holt, B. Moore, and M. Yandell, 2014 Genome Annotation and Curation Using MAKER and MAKER-P. Curr. Protoc. Bioinforma. 48: 4.11.1-39.

Canaguier, A., J. Grimplet, G. Di Gaspero, S. Scalabrin, E. Duchêne *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.

Chan, P. P., B. Y. Lin, A. J. Mak, and T. M. Lowe, 2021 tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 49: 9077–9096.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón *et al.*, 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Duchêne, É., V. Dumas, G. Butterlin, N. Jaegli, C. Rustenholz *et al.*, 2020 Genetic variations of acidity in grape berries are controlled by the interplay between organic acids and potassium. Theor. Appl. Genet. 133: 993–1008.

Dunn, N. A., D. R. Unni, C. Diesh, M. Munoz-Torres, N. L. Harris *et al.*, 2019 Apollo: Democratizing genome annotation. PLOS Comput. Biol. 15: e1006790.

Eichhorn, K. W., and D. H. Lorenz, 1977 Phanologische Entwicklungsstadien der Rebe. Nachrichtenblatt Dtsch. Pflanzenschutzdienstes.

Gao, S., D. Bertrand, B. K. H. Chia, and N. Nagarajan, 2016 OPERA-LG: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. Genome Biol. 17: 102.

Girgis, H. Z., 2015 Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 16: 227.

Girollet, N., B. Rubio, C. Lopez-Roques, S. Valière, N. Ollat *et al.*, 2019 De novo phased assembly of the Vitis riparia grape genome. Sci. Data 6: 127.

Gotoh, O., 2008 A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res. 36: 2630–2638.

Götz, S., J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj *et al.*, 2008 High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36: 3420–3435.

Gough, J., K. Karplus, R. Hughey, and C. Chothia, 2001 Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J. Mol. Biol. 313: 903–919.

Haas, B. J., S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen *et al.*, 2008 Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9: R7.

Haft, D. H., J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu *et al.*, 2013 TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res. 41: D387-395.

Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, 2016 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinforma. Oxf. Engl. 32: 767–769.

Hoff, K. J., A. Lomsadze, M. Borodovsky, and M. Stanke, 2019 Whole-Genome Annotation with BRAKER. Methods Mol. Biol. Clifton NJ 1962: 65–95.

Holt, C., and M. Yandell, 2011 MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491.

Howe, K. L., T. Chothia, and R. Durbin, 2002 GAZE: A Generic Framework for the Integration of Gene-Prediction Data by Dynamic Programming. Genome Res. 12: 1418–1427.

Huang, S., M. Kang, and A. Xu, 2017 HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. Bioinformatics 33: 2577–2579.

Iwata, H., and O. Gotoh, 2012 Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 40: e161.

Jaillon, O., J.-M. Aury, B. Noel, A. Policriti, C. Clepet *et al.*, 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463–467.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li *et al.*, 2014 InterProScan 5: genome-scale protein function classification. Bioinforma. Oxf. Engl. 30: 1236–1240.

Kent, W. J., 2002 BLAT--the BLAST-like alignment tool. Genome Res. 12: 656–664.

Killick, R., and I. A. Eckley, 2014 changepoint: An R Package for Changepoint Analysis. J. Stat. Softw. 58: 1–19.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37: 540–546.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722–736.

Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5:.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12.

Letunic, I., and P. Bork, 2018 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 46: D493–D496.

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25: 2078–2079.

Liao, Y., G. K. Smyth, and W. Shi, 2014 featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma. Oxf. Engl. 30: 923–930.

Lodhi, M. A., and B. I. Reisch, 1995 Nuclear DNA content of Vitis species, cultivars, and other genera of the Vitaceae. Theor. Appl. Genet. 90: 11–16.

Lomsadze, A., P. D. Burns, and M. Borodovsky, 2014 Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 42: e119.

Lomsadze, A., V. Ter-Hovhannisyan, Y. O. Chernoff, and M. Borodovsky, 2005 Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33: 6494–6506.

Lu, S., J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer *et al.*, 2020 CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 48: D265–D268.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang *et al.*, 2012 SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1: 2047-217X-1–18.

Lupas, A., M. Van Dyke, and J. Stock, 1991 Predicting coiled coils from protein sequences. Science 252: 1162–1164.

Majoros, W. H., M. Pertea, and S. L. Salzberg, 2004 TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinforma. Oxf. Engl. 20: 2878–2879.

Massonnet, M., N. Cochetel, A. Minio, A. M. Vondras, J. Lin *et al.*, 2020 The genetic basis of sex determination in grapes. Nat. Commun. 11: 2902.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Merdinoglu, D., G. Butterlin, L. Bevilacqua, V. Chiquet, A.-F. Adam-Blondon *et al.*, 2005 Development and characterization of a large set of microsatellite markers in grapevine (Vitis vinifera L.) suitable for multiplex PCR. Mol. Breed. 15: 349–366.

Mi, H., D. Ebert, A. Muruganujan, C. Mills, L.-P. Albou *et al.*, 2021 PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 49: D394–D403.

Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar *et al.*, 2021 Pfam: The protein families database in 2021. Nucleic Acids Res. 49: D412–D419.

Navarro-Payá, D., A. Santiago, L. Orduña, C. Zhang, A. Amato *et al.*, 2022 The Grape Gene Reference Catalogue as a Standard Resource for Gene Selection and Genetic Improvement. Front. Plant Sci. 12:.

Necci, M., D. Piovesan, Z. Dosztányi, and S. C. E. Tosatto, 2017 MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinforma. Oxf. Engl. 33: 1402–1404.

Palumbo, M. C., S. Zenoni, M. Fasoli, M. Massonnet, L. Farina *et al.*, 2014 Integrated Network Analysis Identifies Fight-Club Nodes as a Class of Hubs Encompassing Key Putative Switch Genes That Induce Major Transcriptome Reprogramming during Grapevine Development. Plant Cell 26: 4617–4635.

Pedruzzi, I., C. Rivoire, A. H. Auchincloss, E. Coudert, G. Keller *et al.*, 2013 HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. 41: D584-589.

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. 26: 841–842.

Rhie, A., B. P. Walenz, S. Koren, and A. M. Phillippy, 2020 Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21: 245.

Sallet, E., J. Gouzy, and T. Schiex, 2019 EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. Methods Mol. Biol. Clifton NJ 1962: 97–120.

Salmela, L., and E. Rivals, 2014 LoRDEC: accurate and efficient long read error correction. Bioinformatics 30: 3506–3514.

Shumate, A., and S. L. Salzberg, 2021 Liftoff: accurate mapping of gene annotations. Bioinformatics 37: 1639–1643.

Sigrist, C. J. A., E. de Castro, L. Cerutti, B. A. Cuche, N. Hulo *et al.*, 2013 New and continuing developments at PROSITE. Nucleic Acids Res. 41: D344-347.

Sillitoe, I., N. Dawson, T. E. Lewis, S. Das, J. G. Lees *et al.*, 2019 CATH: expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res. 47: D280–D284.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Slater, G. S. C., and E. Birney, 2005 Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6: 31.

Smit, A. F. A., R. Hubley, and P. Green, 2013 RepeatMasker Open-4.0.

Song, L., S. Sabunciyan, G. Yang, and L. Florea, 2019 A multi-sample approach increases the accuracy of transcript assembly. Nat. Commun. 10: 5000.

Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler, 2008 Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24: 637–644.

Stanke, M., O. Schöffmann, B. Morgenstern, and S. Waack, 2006 Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7: 62.

Tang, H., X. Zhang, C. Miao, J. Zhang, R. Ming *et al.*, 2015 ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 16: 3.

Taylor, J., and D. Butler, 2017 R Package ASMap: Efficient Genetic Linkage Map Construction and Diagnosis. J. Stat. Softw. 79: 1–29.

Torkamaneh, D., J. Laroche, M. Bastien, A. Abed, and F. Belzile, 2017 Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics 18: 5.

Varet, H., L. Brillet-Guéguen, J.-Y. Coppée, and M.-A. Dillies, 2016 SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. PloS One 11: e0157022.

Vasimuddin, Md., S. Misra, H. Li, and S. Aluru, 2019 Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems, pp. 314–324 in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*,.

Velasco, R., A. Zharkikh, M. Troggio, D. A. Cartwright, A. Cestaro *et al.*, 2007 A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLOS ONE 2: e1326.

Vitulo, N., C. Forcato, E. C. Carpinelli, A. Telatin, D. Campagna *et al.*, 2014 A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biol. 14: 99.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9: e112963.

Wang, M., and L. Kong, 2019 pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinformatics 20: 28.

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis *et al.*, 2018 BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35: 543–548.

Wilson, D., R. Pethica, Y. Zhou, C. Talbot, C. Vogel *et al.*, 2009 SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 37: D380-386.

Wu, C. H., A. Nikolskaya, H. Huang, L.-S. L. Yeh, D. A. Natale *et al.*, 2004 PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res. 32: D112-114.

Wu, T. D., and C. K. Watanabe, 2005 GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875.

Xu, G.-C., T.-J. Xu, R. Zhu, Y. Zhang, S.-Q. Li *et al.*, 2019 LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. GigaScience 8: giy157.

Table 1: **Assembly statistics of the PN40024.v4 REF and PN12X.v2 genome assembly.** The table lists statistics of the PN40024.v4 REF and PN12X.v2 scaffolded and chromosome-anchored genome assemblies. 'N' denotes the number (No.) of unknown bases.

SCAFFOLDS	No. Scaf.	Min. size [bp]	Avg. size [kb]	Median size [kb]	L50	N50 [Mb]	Max. size [Mb]	Sum [Mb]	No. Ns	GC [%]
PN12X.v2	2,059	2,001	236	5	41	3.43	13.10	485.19	14,976,411	33.5
PN40024.v4	640	542	742	20	25	6.50	15.23	474.61	1,755,062	34.4
Anchored PN12X.v2	367	2,010	1,250	277	37	3.57	13.10	465.64	11,921,253	33.6
Anchored PN40024.v4	165	1,085	2,801	1,506	24	6.57	15.23	462.14	1,631,047	34.4

Table 2: Assembly quality values of PN40024.v4 and 12X.v2. Assembly quality values measured by Merqury for PN40024.v4 and 12X.v2 genome assemblies. QV denotes base level quality value.

	12X.v2	**PN40024.v4**
QV	37.4338	36.0171
Error rate [%]	0.000180559	0.000250199
k-mer completeness [%]	96.79	96.96

Table 3: Haplotype block statistics of PN40024.v4 and 12X.v2. Phasing accuracy estimation of Merqury for PN40024.v4 and 12X.v2 genome assembly. ALT denotes alternative heterozygous sequence parts of PN40024.v4.

	12X.v2	**PN40024.v4**	**ALT**

Number of blocks	2,575	1,454	289
Total bases in blocks [bp]	474,845,411	468,703,133	19,519,697
Block N50 size [kb]	1,762	2,050	250
Switch error rate [%]	0.766002	0.959042	4.75944

Table 4: VCost.v3, PN40024.v4.1 REF haplotype and PN40024.v4.2 REF haplotype gene prediction overview.

	VCost.v3		**PN40024.v4.1**		**PN40024.v4.2**	
	Number	**Mean length [bp]**	**Number**	**Mean length [bp]**	**Number**	**Mean length [bp]**
Protein-coding genes	41,182	4,485	35,197	4,742	35,230	4,735
Transcripts	47,363	1,383	41,160	1,433	41,173	1,440
Exons	239,165	273	208,581	282	208,719	283
CDS	225,869	220	199,956	231	200,059	232
5' UTRs	26,024	259	17,019	280	17,478	275
3' UTRs	26,994	327	17,873	440	18,344	433
tRNAs	19	74	725	75	725	75

Table 5: Gene numbers of heterozygous sequence regions. The abbreviation ALT denotes the alternative heterozygous sequence regions.

	Bases [bp]		**Number of genes**	
	PN40024.v4	**ALT**	**PN40024.v4.1**	**ALT**
chr02	1,610,271	1,886,900	190	214

chr03	288,001	287,774	14	13
chr04	1,049,642	929,781	123	122
chr07	11,422,405	10,851,409	830	792
chr10	5,475,057	5,100,371	650	623
chr11	733,078	630,772	43	41
chr15	60,730	52,641	5	4
TOTAL	**20,639,184**	**19,739,648**	**1,855**	**1,809**

Figure 1: **Assembly process for the PN40024.v4 genome sequence assembly. A**) Initial datasets: Sanger-based scaffolds of PN12X.v2 (red) with unknown bases ('N's'), genomic SMRT reads (green) and genomic short reads (blue). Erroneous bases are represented by a black line. **B**) Scaffold assembly steps. Same color code as A). Dark green regions represent newly incorporated SMRT sequencing regions. **C**) Pseudomolecule construction using the new scaffolds and genetic maps. Sequence regions of 12X.v2 are colored red and newly incorporated SMRT sequencing regions are colored dark green.

Figure 2: **Percentage of mapped genomic reads (A) and percentage of properly paired genomic reads (B) between PN40024.v4, PN12X.v2 and cv. 'Cabernet Sauvignon' (Massonnet et al., 2020) for 11 paired-end resequencing datasets of *V. vinifera* cultivars.** The x-axis denotes the source (cultivar) of the genomic reads and the y-axis the percentage [%] of mapped reads. Note that the PN40024 dataset was obtained with Illumina Genome Analyzer IIx sequencing and all other samples with Illumina HiSeq 4000. The PN40024 dataset is therefore of lower quality than the others.

Figure 3: **Location of regions assembled using long reads, density of errors and of heterozygous SNPs in the PN40024.v4 genome sequence assembly.** The x-axis shows the 19 main pseudochromosomes and the artificial chrUn ('Un'). The y-axis shows the base position in [bp]. 'Pacbio regions' refers to sequences derived from genomic SMRT reads. The seven heterozygous regions are squared in green.

Figure 4: **Density of 'Pinot noir' and 'Schiava grossa' homozygous SNPs compared to the PN40024.v4 genome assembly.** The x-axis shows the 19 main pseudochromosomes and the artificial chrUn ('Un'). The y-axis shows the base position in [bp]. Where density of 'Pinot noir' SNPs is high, it means PN40024.v4 carries the 'Schiava grossa' haplotype and *vice versa*. The regions where both 'Pinot noir' and 'Schiava grossa' SNP density is low correspond to regions where both genomes share a common haplotype.

Figure 5: **Density of 'Helfensteiner' homozygous SNPs compared to the PN40024.v4 genome assembly.** The x-axis shows the 19 main pseudochromosomes and the artificial chrUn ('Un'). The y-axis shows the base position in [bp]. The seven regions squared in green are the heterozygous regions.

Figure 6: **Plant core genes of the PN40024.v4 and PN12X.v2 genome assemblies and their annotations.** The 2,326 plant core genes of the database eudicots_odb10 were determined in the PN40024.v4 genome assembly, in its annotation PN40024.v4.1, in the PN12X.v2 genome assembly and in the VCost.v3 gene annotation. 'PN40024.v4.2' is the PN40024.v4 gene annotation after manual curation of the fragmented and missing plant core genes.

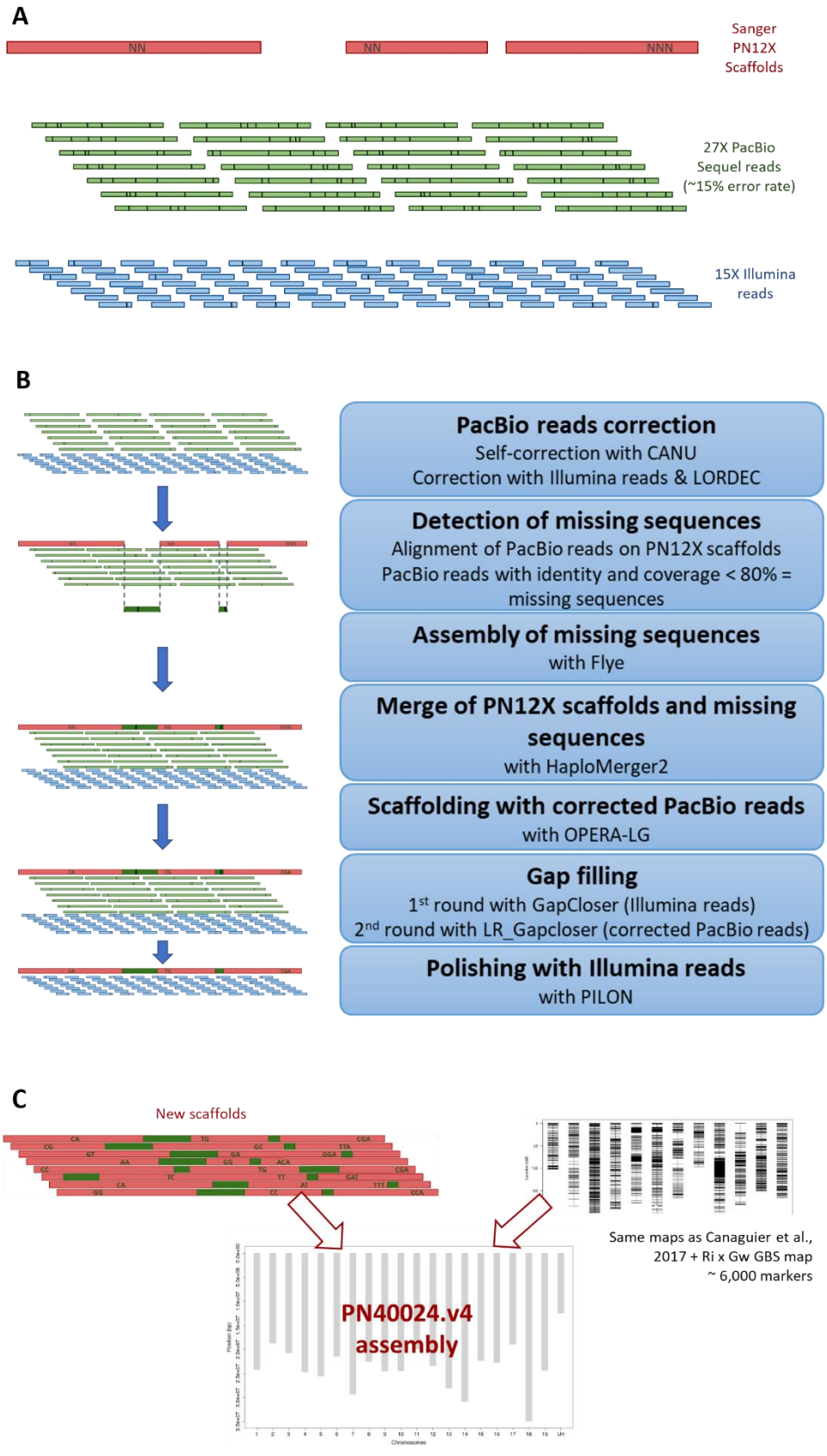

Figure 1
254x447 mm (x DPI)

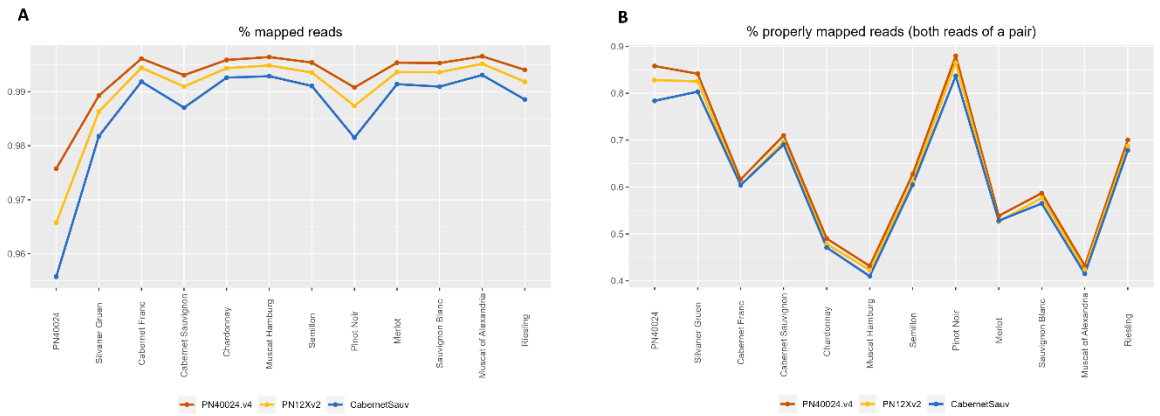

Figure 2
406x152 mm (x DPI)

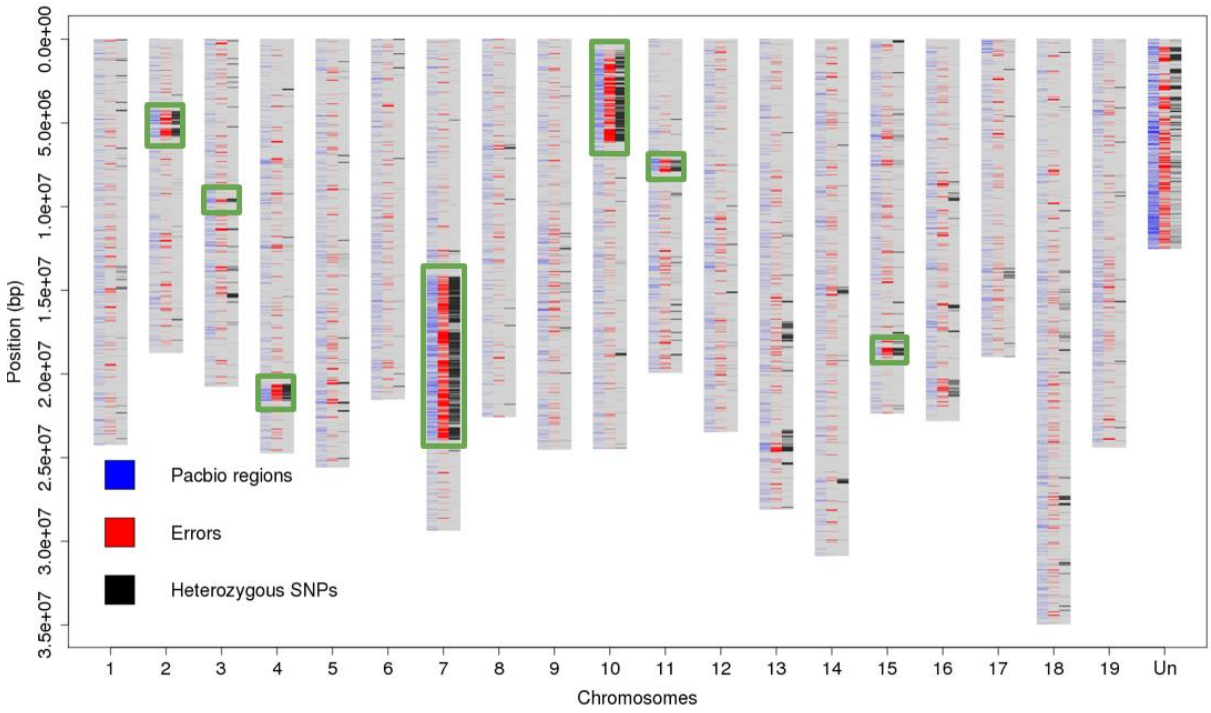

Figure 3
184x108 mm (x DPI)

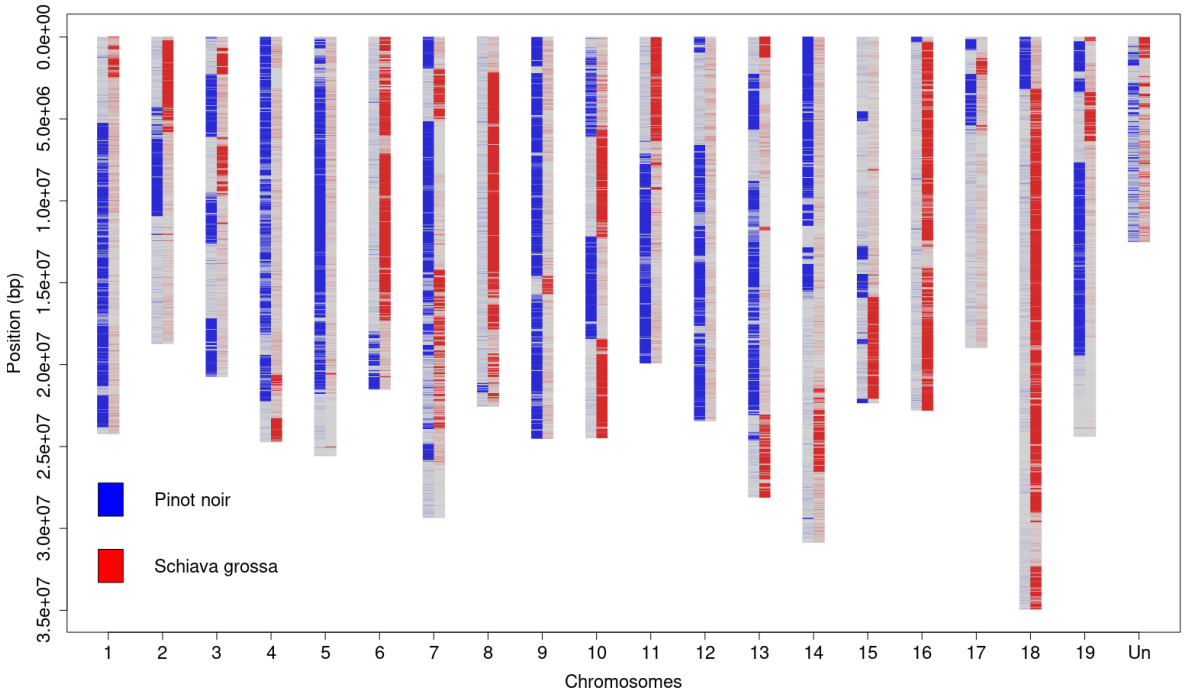

Figure 4
418x261 mm (x DPI)

1

2

3

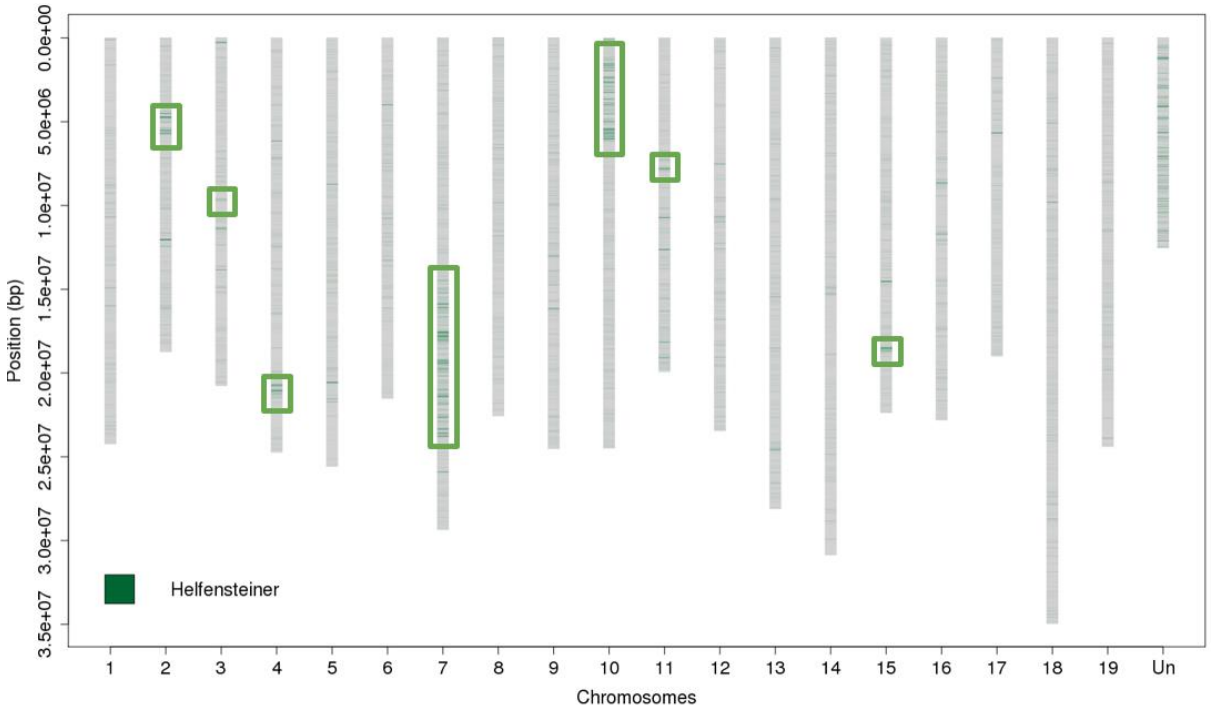

Figure 5
184x108 mm (x DPI)

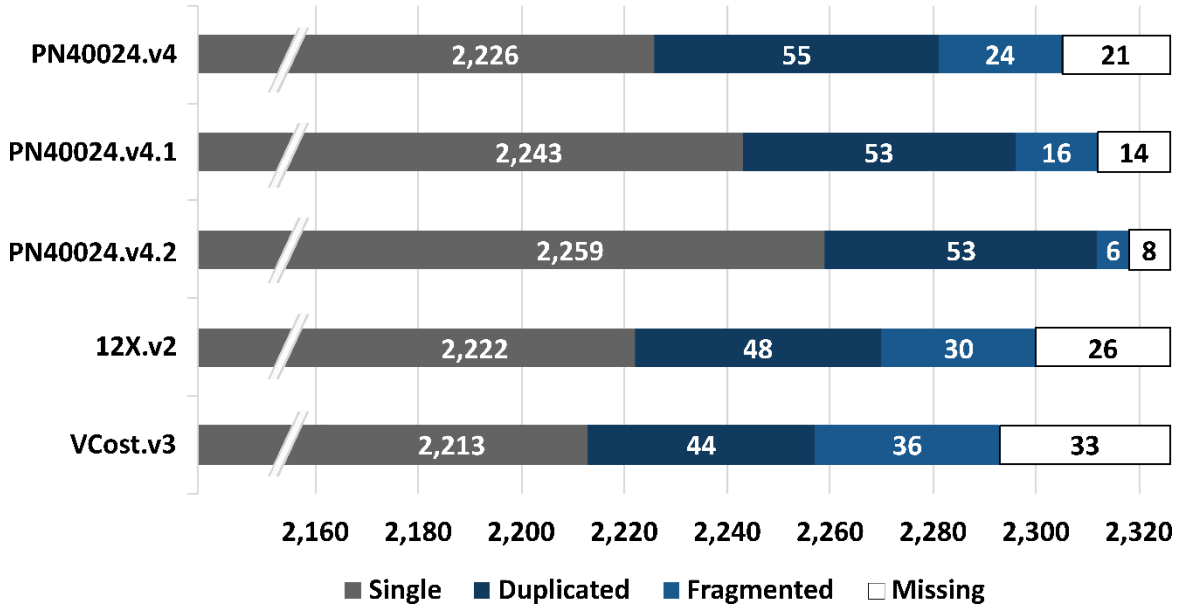

Figure 6
154x79 mm (x DPI)