

A chromosome-level genome assembly of *Vanilla planifolia* uncovers the genomic architecture underlying partial endoreplication

Received: 3 February 2026

Accepted: 4 May 2026

Published online: 13 May 2026

Cite this article as: Piet Q., Besse P., Bocs S. *et al.* A chromosome-level genome assembly of *Vanilla planifolia* uncovers the genomic architecture underlying partial endoreplication. *BMC Genomics* (2026). <https://doi.org/10.1186/s12864-026-12926-1>

Quentin Piet, Pascale Besse, Stéphanie Bocs, Olivier Bouchez, Mickael Bourge, Carine Charron, Charlotte Cravero, Amaury Dedours, Gaëtan Droc, Michel Dron, Michel Grisoni, Christophe Klopp, Fanny Lambert, William Marande, Hadrien Maté Gérando, Nathalie Rodde, Marine Sallaberry, Sonja Siljak-Yakovlev, Marilyne Summo & Cyril Jourda

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

A chromosome-level genome assembly of *Vanilla planifolia* uncovers the genomic architecture underlying partial endoreplication

Quentin Piet¹, Pascale Besse^{1,*}, Stéphanie Bocs^{2,3,4,*}, Olivier Bouchez^{5,*}, Mickael Bourge^{6,*}, Carine Charron^{7,*}, Charlotte Cravero^{8,*}, Amaury Dedours^{9,*}, Gaëtan Droc^{2,3,4,*}, Michel Dron^{10,*}, Michel Grisoni^{11,*}, Christophe Klopp^{12,*}, Fanny Lambert^{13,*}, William Marande^{8,*}, Hadrien Maté de Gérando^{13,*}, Nathalie Rodde^{8,*}, Marine Sallaberry^{5,*}, Sonja Siljak-Yakovlev^{14,*}, Marilyne Summo^{2,3,4,*} and Cyril Jourda^{7,\$}

¹Université de la Réunion, UMR PVBMT, F-97410 Saint-Pierre, La Réunion, France

²CIRAD, UMR AGAP Institut, 34398 Montpellier, France

³UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, 34398 Montpellier, France

⁴French Institute of Bioinformatics (IFB) - South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, 34398 Montpellier, France

⁵INRAE, GeT-PlaGe, Genotoul, 31326 Castanet-Tolosan, France

⁶Cytometry Facility, Imagerie-Gif, Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France

⁷CIRAD, UMR PVBMT, F-97410 Saint-Pierre, La Réunion, France

⁸INRAE, Centre National de Ressources Génomiques Végétales, 31326 Castanet-Tolosan, France

⁹Eurovanille, Rue de Maresquel, 62870 Gouy Saint André, France

¹⁰Université Paris-Saclay, CNRS, INRAE, Univ. Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91405 Orsay, France

¹¹CIRAD, UMR PVBMT, 501 Tamatave, Madagascar

¹²Plateforme Bioinformatique, Genotoul, BioinfoMics, UR875 Biométrie et Intelligence Artificielle, INRAE, Castanet-Tolosan, France

¹³Département Biotechnologie, V. Mane Fils, 06620 Le Bar Sur Loup, France

¹⁴Université Paris-Saclay, CNRS, AgroParisTech, Ecologie Société Evolution (ESE), 91190 Gif-sur-Yvette, France

*These authors are listed using an alphabetical order by name.

\$Correspondence: Cyril Jourda (cyril.jourda@cirad.fr; Tel: +262262492782 - Secr. : +262262492709)

Running title: Vanilla partial endoreplication genome landscape

Abstract

Background

Partial endoreplication is a prominent developmental feature and poses a significant challenge for whole genome assembly in orchids. This form of cell cycle results in highly unbalanced cell DNA content, with the highly

endoreplicated (P) fraction being overrepresented in sequencing data compared to the non-endoreplicated (F) fraction.

Results

Here, we report the first genome assembly of *Vanilla planifolia* into 16 chromosome pairs using axillary buds enriched in non-endoreplicated 2C-nuclei (55%) as determined by flow cytometry. The assembly was generated using a hybrid approach combining PacBio HiFi sequencing and Omni-C scaffolding generated in this study, together with a GBS-SNP genetic map and Oxford Nanopore Technologies long-read data from the literature.

For the first time, we identified P and F regions within reconstructed chromosomes, representing 20.57% and 79.43% of the genome, respectively based on DNA sequencing data from three tissues with varying levels of endoreplicated nuclei. P regions were gene-rich and located at chromosome ends, whereas F regions were SSR-rich and located at central parts of chromosomes. Remarkably, 97.24% of SSRs were found in F regions, predominantly comprising the trinucleotide AAG/CTT motif, which may contribute to the absence of endoreplication in these regions.

Protein-encoding genes overrepresented in F regions were associated with negative regulation of flower development, mitotic cycle progression, cell division and histone modification.

Conclusions

This accurate high-quality chromosome-scale *V. planifolia* genome assembly provides unprecedented insights into the structural and molecular characteristics of partial endoreplication in Orchids and represents a major step toward the characterization of this complex genome.

Keywords

Vanilla planifolia, genomics, partial endoreplication, Chromosome-scale architecture, Long-read sequencing

Background

Endoreplication is a common cell cycle variant in eukaryotes. It is a developmentally regulated process characterized by repeated S and G phases resulting from the exit from the mitotic cycle during the G2 phase or early mitotic stages, followed by a direct return to S phase. This leads to successive rounds of DNA replication without cell division, causing an exponential increase in nuclear DNA content [1]. Some orchid species exhibit a distinctive form of endoreplication, called partial endoreplication (PE), also known as progressively partial endoreplication [2-4] or strict partial endoreplication [5]. Unlike complete endoreplication (CE), cells undergoing PE replicate only a portion of their genomic sequence. Consequently, their nuclei exhibit less than a two-fold increase in DNA content after each replication cycle [6]. The nuclear DNA of PE species can be conceptualized as comprising two virtual fractions: a F fraction, which remains at the diploid level, and a P fraction, which undergoes endoreplication [5]. These fractions are considered to be fixed in

organisms of a given species, independently of the developmental stage or organ, although the relative proportions of 2C nuclei and of different classes of partially replicated nuclei vary among organs and tissues. This feature of PE, along with the fact that the resulting gametes are consistently haploid suggests a tightly regulated control of the endoreplication process. By restricting DNA replication to a subset of the genome, PE likely reduces the energetic and temporal costs associated with conventional endoreplication (CE), while preserving the functional advantages of increased nuclear DNA content. This selective replication may also limit the deleterious effects of repetitive DNA by excluding repeat-rich regions from amplification, while enhancing gene expression in endoreplicated fractions through gene dosage effects. Consequently, PE may support organ growth and metabolic activity, and contribute to plant performance under varying environmental conditions. These features suggest that PE plays a significant role in ecological adaptation and lineage diversification [3, 7, 8]. Selective genome amplification or rearrangement occurs in other biological systems. For example, polytene chromosomes result from repeated DNA replication without cell division and are found mostly in insects and also in mammals, protists and plants [9]. Plant polytene chromosomes are found restricted to some specific tissues in a few species [10]. Ciliates undergo programmed genome remodeling, amplifying somatic sequences while eliminating others [11] and some metazoans eliminate portions of their genome during development [12]. However, these

patterns involve complete or variable endoreduplication rather than the systematically structured PE characteristic of Orchids.

To date, PE is specific to Orchids and the only form of endoreplication reported in the Vanilloideae and Orchidoideae subfamilies, whereas both PE and CE occur in Epidendroideae and Cyripedioideae and only CE has been documented in the basalmost subfamily Apostasioideae [3, 5, 13]. Among the Vanilloideae, *Vanilla planifolia* is a tropical orchid primarily propagated clonally through stem cuttings, due to limited natural pollination outside its native range. This has resulted in the widespread cultivation of a restricted number of genetically similar accessions (e.g., 'Daphna', 'CR0040'). The species has undergone a relatively recent domestication process, with dissemination from its center of origin in Mesoamerica to major production areas such as Madagascar [2, 14]. It represents a model system to study PE with its P fraction encompassing 28.4% of its genomic sequence, which can be replicated up to 64 fold in cells showing the highest endoploidy level in this species [5]. Previous studies comparing the P and F fractions showed that the latter was mostly composed of repeats [4, 15]. More recently, 33% of *V. planifolia* genes were found in regions identified as non-endoreplicated based on sequencing depth comparisons among several tissues with different levels of PE [16]. From a genomics perspective, this uneven replication pattern inherent to PE introduces biases in sequencing coverage which challenges the *Vanilla* genome assembly. This may result in gaps, misassemblies, unanchored sequences, or missing regions. This highlights the need for a high-quality

chromosome-level assembly to accurately resolve chromosomal organization and the distribution of the P and F fractions. In recent years, two chromosome-scale haplotype phased assemblies have been published [16, 17], each generated from different cultivars and organs (Daphna leaves and CR0040 nodes, respectively). Unfortunately, only 14 pairs of pseudomolecules were assembled whereas the basic number of chromosomes for the *Vanilla* genus is $x=16$ [18] and two thirds of the sequence was either missing (Daphna) or unanchored (CR0040). Therefore, these assemblies did not allow for an analysis of structural characteristics including PE and the distribution of the F and P fractions along the chromosomes.

The *V. planifolia* genome assembly completeness and contiguity are highly impacted by PE [16]. This is attributable to the underrepresentation of the F fraction in endoreplicated nuclei, leading to reduced sequencing depth, especially in tissues with high proportions of strongly endoreplicated nuclei, such as leaves, challenging the assembly of these repeat rich regions [5]. Thus, achieving sufficient coverage of this fraction should be possible by selecting tissues with low proportions of endoreplicated nuclei. In addition, the first genetic map for this species, composed of 16 linkage groups (LG), was recently assembled using a genotyping by sequencing (GBS) approach on a population obtained from self-fertilization of *V. planifolia* CR0040 [19]. The mapping of the flanking sequences of SNP used to construct the genetic map on the CR0040 and Daphna assemblies showed that chromosome 1 should be

divided into three parts, thus resolving the mystery of the two missing chromosome pairs.

In the present study, we address the previously identified challenges of PE in genome assembly by combining flow cytometry, PacBio HiFi long-read sequencing, Omni-C scaffolding, and genetic map anchoring to generate a phased assembly of *V. planifolia*. This assembly is more complete than previous versions and successfully represents all 16 chromosome pairs. This high-quality genome and its annotation enabled us to delineate the chromosomal regions corresponding to the F and P fractions and to compare their composition in terms of genomic features.

Methods

Plant material

A cultivated *Vanilla planifolia* specimen from CIRAD, collected in La Reunion in 2003 by Michel Grisoni, was used in this study. CIRAD confirmed that this material is legally available for research purposes and is not subject to a Material Transfer Agreement. It was deposited under the voucher number CR-VA-00040 (also named “CR0040”) in the VATEL Biological Resource Center collection (Saint Pierre, La Réunion, France) on 10 March 2004. It complies with phytosanitary regulations (LOA permit n°03/RE2/01555). Formal taxonomic identification was made by the collector, and was further confirmed by molecular technologies including AFLP [2] and Genotyping-By-Sequencing [20]. It is maintained and publicly available by cuttings under shade-house and

by *in vitro* culture established from axillary buds. The vitroplants were propagated by stem cuttings and grown *in vitro* in Murashige & Skoog medium under controlled conditions. The environment was maintained at 24-26 °C with a 12-hour light/dark cycle.

Flow cytometry

The total nuclear DNA amount was assessed by flow cytometry as described in [21]. Leaves, roots and axillary buds from young *in vitro* *V. planifolia* (CR0040 cultivar) were chopped with a razor blade in a plastic Petri dish with 1 ml of Gif nuclei-isolation buffer (45 mM MgCl₂, 30 mM sodium citrate, 60 mM MOPS, 1% (w/v) polyvinylpyrrolidone 10,000, pH 7.2) containing 0.2% (w/v) Triton X-100, supplemented with 5 mM sodium metabisulphite. The resulting suspension was filtered through 50-µm nylon mesh. Nuclei were stained with 5-10 µg/ml DAPI (or 50-100 µg/ml propidium iodide, PI) and kept 5 min at 4°C. DNA content of 5,000–10,000 stained nuclei were determined for each sample using a cytometer (CytoFLEX S, Beckman Coulter. DAPI: Excitation 405 nm; emission through a 450/45 nm band-pass filter. PI: Excitation 561 nm; emission through a 610/20 nm band-pass filter). Nuclei were identified using a gate on Side-Scatter and DNA content-Area and the cytogram of DNA content-Area versus DNA content-Height signals served to select singlets, eliminating doublets and detecting any degradation. Frequencies of different ploidy levels were identified by their fluorescence intensity both in DAPI and PI to confirm ploidy levels. Endoreplication levels were determined from measurements obtained from five clonal vitropants of the cultivar CR0040

following the approach previously described [5]. Briefly, in tissues exhibiting PE, DNA fluorescence intensity profiles show a series of peaks that differ from those expected in conventional endoreplication (CE). Successive peaks follow a strict twofold progression in tissues with CE whereas in tissues with PE, the ratios between adjacent peaks are lower than two. This indicates that only a fraction of the genome is selectively replicated. These distinct peak patterns allow the endoreplication state of the cells to be accurately determined [2, 3, 4, 5, 7, 8].

High molecular weight (HMW) DNA extraction

Meristematic tissues from axillary buds were harvested from 44 CR0040 *in vitro* plantlets under binocular lenses. The harvesting process yielded approximately 180-190 mg of meristematic tissue, subsequently stored at -80°C until DNA extraction. HMW DNA extraction was performed on the total amount of sampled axillary bud meristems using QIAGEN Genomic-tips 100/G kit (Qiagen, MD, USA). The tissues were mechanically grinded at 30 Hz for two intervals of 45 seconds each and extraction was performed following manufacturer's instructions. Briefly, after 3 hours of lysis and one centrifugation step, the DNA was immobilized on the column. After two washing steps, DNA was eluted from the column, then desalted and concentrated by isopropyl alcohol precipitation. A final wash in 70% ethanol was performed before resuspending the DNA in the buffer. Analyses of DNA quantity and quality were performed using NanoDrop and Qubit (Thermo

Fisher Scientific, MA, USA). DNA integrity was also assessed using the Agilent FP-1002 Genomic DNA 165 kb on the Femto Pulse system (Agilent, CA, USA).

DNA sequencing

Library preparation and sequencing were performed according to the manufacturer's instructions "Procedure & Checklist – Preparing whole genome and metagenome libraries using SMRTbell® prep kit 3.0". At each step, DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA purity was tested using the nanodrop (Thermofisher) and size distribution and degradation assessed using the Femto pulse Genomic DNA 165 kb Kit (Agilent). Purification steps were performed using AMPure PB beads (PacBio) and SMRTbell cleanup beads (PacBio). A repair stage was performed on 10 µg of DNA using SMRTbell Damage Repair Kit SPV3 (PacBio) then sheared at 15 kb (speed V33) using the Megaruptor3 system (Diagenode). After a nuclease treatment, the library was size-selected using a 6 kb cutoff on the Pippin HT Size Selection system (Sage Science) with "0.75% Agarose, 6-10 kb High Pass, 75E" protocol. To obtain sufficient material for sequencing, this procedure was repeated on another 10 µg of DNA. The only difference is that the fragmentation was performed this time on the Megaruptor1 system at 20 kb. Libraries were sequenced using binding kit 3.2 by Adaptive Loading onto 12 SMRTcells on the PacBio Sequel II System instrument at 90 pM with a 2 hours pre-extension and a 30 hours movie.

Evaluation of representativity bias in the sequencing data

PacBio HiFi CCS reads from *V. planifolia* CR0040 nodes [15] and newly generated PacBio HiFi CCS were aligned to the *V. planifolia* CR0040 V1 assembly [15] using Minimap2 v2.2.24 [22]. Mean read depths across 20 kb non-overlapping windows were calculated using mosdepth v0.3.3 [23] with '--fast_mode' option. For each tissue, the mean depth of both chromosomal fraction and the unanchored sequences of the assembly was computed.

Omni-C library preparation and sequencing

The Omni-C library (Dovetail Genomics®) was produced according to the manufacturer's instructions. Briefly, close to 50 mg of frozen axillary bud meristems were grinded in liquid nitrogen and resuspended in PBS. Then, DNA was fixed with formaldehyde and digested using 0.3 µl of a nuclease enzyme mix. After binding 350 ng of the digested DNA to chromatin capture beads, a proximity ligation was performed and the crosslinks are reversed to produce the linked DNA. Finally, 54 ng of the linked DNA was used to produce an Illumina® library. An initial run of 1 Gb data was produced to assess the library quality then 65X coverage data (150 Gb) was produced (paired-end reads 2 x 150 bp on an Illumina® NovaSeq system).

Genome assembly and pseudomolecules reconstruction

Axillary buds HiFi reads from *V. planifolia* CR0040, generated in the present study, were assembled using Hifiasm v0.19.5 [24], incorporating *V. planifolia* Daphna ONT long-read data [17] via the --ul option to improve the assembly contiguity. Contigs were checked for foreign organism contamination using

NCBI's FCS-GX v0.4.0 [25] and for mitochondrial and chloroplastic sequences using MitoHiFi v3.2 [26] with MitoFinder v1.4.1 [27]. Omni-C reads were then mapped to each haplotype using Juicer v1.5.7 [28] using default parameters. A single iteration of 3D de novo assembly v529ccf4 [29] with -r 0 parameter to limit automatic contig splitting was done before manual curation and scaffolding validation in Juicebox v2.17 [30]. The manual curation result was validated using additional information including alignment with *V. planifolia* CR0040 V1 scaffolds using D-GENIES [31], mapping of GBS-SNPs markers [19] using BWA-MEM2 v2.0 [32] and telomeric repeat locations using Seqtk v1.4-r130 [33] with the telomeric motif CCCTAAT and its reverse complement ATTAGGG [34]. The chromosome assembly completeness was evaluated using BUSCO v5.0.0 [35] with the viridiplantae_odb10 and liliopsida_odb10 lineages datasets. The k-mer analysis was performed with KAT v2.4.2 using the comp tool [36].

Structural and functional annotation of genes

An automatic gene annotation was performed on contigs using the Nextflow version of EuGene Eukaryotic Pipeline v2.1.1 [37]. To train its internal statistical model, the *V. planifolia* Daphna proteome [17] and the *V. planifolia* CR0040 transcriptome were used in addition to protein evidence provided by the software. The transcriptome was produced using ten RNA-Seq datasets [16] mapped separately onto the first version of the *V. planifolia* CR0040 assembly with HISAT2 v2.2.1 [38] in an annotation guided fashion by providing splice sites and exons information from CR0040 V1 annotation. Exons and

splice sites positions were first extracted from gene annotations using python scripts provided with the aligner and were used to build the reference genome index with the `--ss` and `--exon` options. Then, the ten RNA-Seq samples were aligned to the *V. planifolia* V2 assembly using HISAT2 with the `--dta` option. A first run of StringTie v2.0.3 [39] was executed by providing CR0040 V1 assembly gene annotation information. Finally, the ten tissue-specific transcriptomes were merged with '*stringtie --merge*' to obtain a non-redundant set of transcripts.

Functional annotation of protein-encoding genes was performed using Funcannotate [40] with default filtering parameters. Briefly, protein sequences were confronted to several biological databases including InterPro (InterProScan v5.67-99.0, [41]), KEGG (KofamScan v1.3.0, [42]), HMM profiles (Gag retrotransposon and Gypsy-like retrotransposon), Repbase v23.12 [43], BUSCO (liliopsida_odb10), RefSeq proteomes of *Arabidopsis thaliana* (GCF_000001735.4), *Oryza sativa* (GCF_001433935.1) and *Phalaenopsis equestris* (GCF_001263595.1). UniProtKB/Swiss-Prot and TrEMBL databases were also queried, with TrEMBL entries filtered to retain only proteins with "protein existence" annotation score going from 1 to 3. Genic content completeness assessment was then carried out on this final set using BUSCO v5.0.0 with the viridiplantae_odb10 and liliopsida_odb10 lineages datasets.

Transposable elements (TEs) were detected, classified and localized using combined approaches. EDTA v2.2.0 [44], LTR_STRUC [45] and Inpactor2 [46] were run in parallel on the CR0040 genome sequence. PASTECClassifier [47]

of REPET v3.0 [48] was used to filter on the Simple Sequence Repeats (SSRs, corresponding to sequences of 1–6 nt motifs repeated in tandem) coverage ($\leq 50\%$) and to remove the Potential Host Genes. CD-HIT-EST [49] was used to remove identical TE between the three sets and to help during (i) the visual inspection of the longest TE dot plots using bl2seq [50] to remove those with a chimeric structure and (ii) the manual validation of PASTECClassifier results to remove redundancy results. TEannot from REPET v3.0 was run in order to keep only TE with at least one full length fragment and a combined total of ten fragments across both haplotypes. TEannot was used again with TE consensi identified in *Ophrys sphegodes* genome [51] in order to evaluate their conservation in *V. planifolia* CR0040 genome and to incorporate the most abundant ones in the final TE library. A last round of CD-HIT-EST clustering was done on the TE set enriched with RepeatModeler [52] and consensi previously identified in *V. planifolia* CR0040 genome [16] with RepeatScout [53] and cleaned using PASTECClassifier, TEannot and visual inspection of bl2seq dot plot of the longest consensi. The RepeatScout and RepeatModeler consensi were added to consolidate EDTA ones when their structure seemed valid (visual inspection of dotplots with Gepard [54]). At the end, *V. planifolia* CR0040 genome was annotated using TEannot and the non-redundant TE library newly produced.

SSRs (1–6 nt motifs) were identified with MISA v2.1 [55]. The minimum number of motif repetitions for a SSR to be considered was defined as ten for mononucleotides, seven for dinucleotides, six for trinucleotides and five for

other repeat units (tetra- to hexanucleotides). SSRs separated by less than 100 nucleotides were considered as compound SSRs. Only those composed of motifs whose length was between 1-6 nucleotides were considered for statistics summary and further analyses.

Synten analysis

Synten blocks and homologous gene pairs between haplotypes A and B of the CR0040 V2 assembly were computed by aligning identified protein sequences using a reciprocal best hit approach with NCBI BlastP and running MCSanX [56] with default parameters on the results.

Detection and molecular characterization of P and F genomic fractions

The distinction between P and F chromosomal fractions was inferred from sequencing depth profiles. In species exhibiting PE, some genomic regions are not uniformly replicated during endocycle progression, leading to relative differences in DNA copy number across chromosomal regions. Normalized read depth is proportional to DNA copy number and so it can be used to infer genomic dosage variation, an approach originally developed for aneuploid plant cells [57]. Consequently, these dosage differences are directly reflected in sequencing depth variations. Regions assigned to the P fraction exhibit read depths proportional to the overall ploidy level, whereas F fraction regions are relatively underrepresented consistent with incomplete endoreplication. Comparative analysis of sequencing depth along the genome therefore

enables the identification and quantification of these two chromosomal fractions. PacBio HIFI (axillary buds and nodes, CR0040 cultivar) and Illumina short reads (leaves, Daphna cultivar) were mapped on the final assembly with Minimap2 v2.2.24 [22] using the *map-hifi* preset and with BWA-MEM2 v2.0 [32], respectively. Secondary alignments were discarded using SAMTOOLS v1.14 [58] with the command 'samtools view -bF 260 -q 1'. Depth values were computed at each genomic position with the command 'samtools depth -@ 4 -aa -Q 1 -J -s'. For each tissue type, the distribution of the mapped reads along the chromosomes was segmented into 200 kb non-overlapping genomic windows. Each window was assigned a value corresponding to the median mapping depth calculated across all the positions within that window. The assembly was then partitioned into two fractions (P and F) using k-means clustering (k=2) based on mapping results obtained exclusively from Illumina Daphna leaves reads, since they exhibited the greatest sequencing-depth contrast between high- and low-coverage regions. The clustering was applied to the normalized median sequencing depth values computed over 200 Kb windows across the genome. For each chromosome, depth values were first normalized by the chromosome's mean sequencing depth prior to clustering. The midpoints between the centroids of P and F clusters, used as thresholds for window assignment to each fraction, ranged from 53.23x to 60.80x raw sequencing depth, reflecting chromosome-specific variation in coverage. Genes, TEs and SSRs densities were calculated in 200-kb genomic windows

based on the annotation results using BEDOPS v2.4.39 [59] with the command 'bedmap --echo --count'. Overlapping TEs were counted only once.

Gene ontology (GO) annotations of genes located in F and P fractions were used to perform a GO enrichment analysis for each fractions using the 'enricher' function from clusterProfiler v4.8.3 [60]. This function conducts an overrepresentation analysis based on a hypergeometric test comparing a gene set of interest to a background gene set. In our study, the background set comprised all genes associated with at least one GO term. P-values were adjusted using the Benjamini-Hochberg correction, with a significance threshold of 0.05.

Results

The first high-quality genome sequence of *Vanilla planifolia* anchored to all 16 chromosome pairs

Endoreplication rates in different tissues of *in vitro* plants (leaves, axillary buds and roots) were measured by flow cytometry. Fig. 1 shows that axillary buds contained the highest proportion of 2C-nuclei ($55.12 \pm 9.52\%$) compared to roots ($17.36 \pm 3.80\%$) and leaves ($3.64 \pm 2.85\%$). Based on these results and to limit the bias of representativeness of genomic fragments induced by PE, axillary buds from *in vitro* plants were selected for subsequent sequencing efforts.

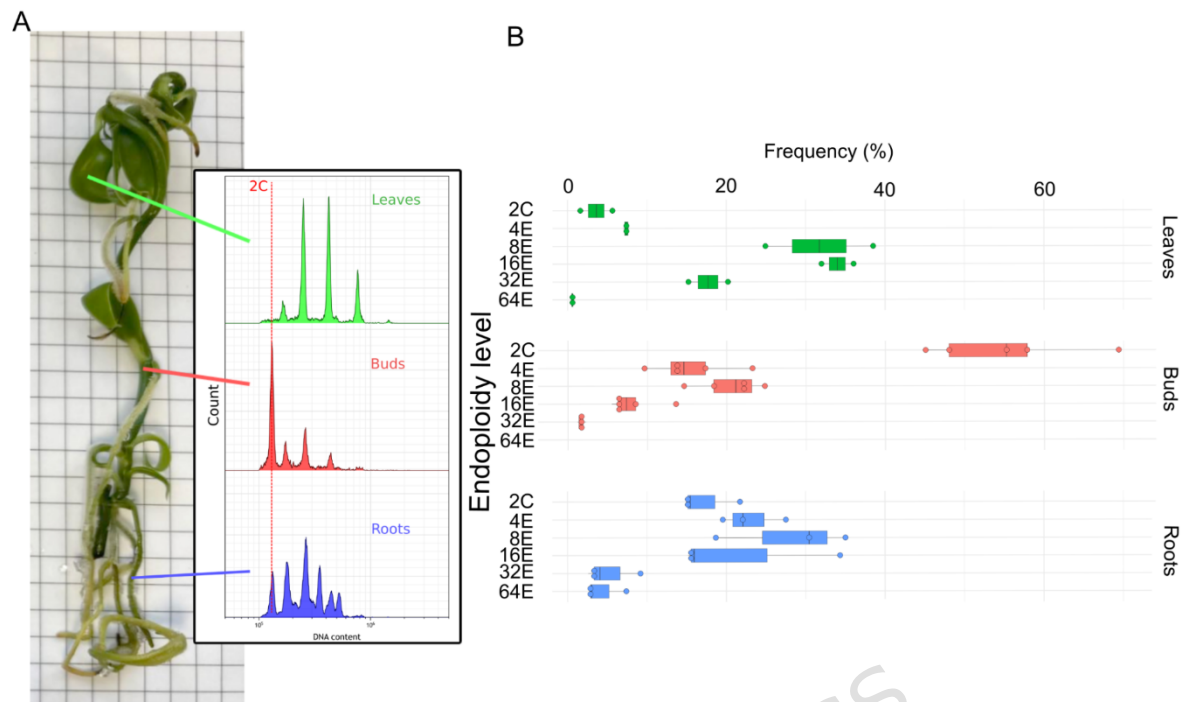


Figure 1. Flow cytometry DNA content distribution in leaves, axillary buds and roots of *Vanilla planifolia* in vitro plantlets. (A) The proportions of nuclei of various endoploidy states were measured. In the DNA content histograms from different tissues, the dotted line represents the expected position of the 2C-peak. (B) Boxplot representation summarizing all measurements of DNA content made in nuclei isolated from the different tissues.

The analysis of the primary alignments of the PacBio HiFi reads from axillary buds (8,240,084 reads) and from nodes (4,659,912 reads) against the CR0040 V1 assembly revealed that, on average, comparable sequencing depths were achieved across the 14 chromosome pairs of this previous assembly (Supplementary Table S1). Conversely, the previously unanchored scaffolds, associated with F regions, exhibited an increased average mapping depth when analysed using the newly generated axillary buds data. These results were reflected in the ratio between the average sequencing depth of

chromosomes and unanchored regions, which decreased from 4.59 for nodes HIFI reads to 1.67 for axillary buds HIFI reads (Supplementary Table S1), indicating a more accurate representation of the F fraction of the *V. planifolia* genome in reads obtained from axillary buds DNA.

Secondly, a comparative analysis of the 27-mer distributions in sequences from axillary buds and nodes, as well as in the assembly, was performed (Fig. 2). HiFi reads obtained from nodes did not clearly reveal the two expected distributions (heterozygous and homozygous fractions, shown in red and purple respectively) for a diploid genome. Blue arrows in Fig. 2A and 2C show intermediate 27-mers depth levels between the expected heterozygous and homozygous distributions at $\sim 30\times$ and $\sim 70\times$, which may result from a mixture of Gaussian distributions in the k-mers profiles of the P fraction of the genome at various endoploidy levels. Note that a typical diploid pattern should show two well separated peaks corresponding to heterozygous and homozygous distributions. Fig. 2B and 2D show that the distributions corresponding to the expected heterozygous (single-copy 27-mers at $\sim 37\times$) and homozygous (two copy 27-mers at $\sim 77\times$) regions were better separated (two distinct peaks) in axillary buds reads than in nodes reads (Fig. 2A and 2C). Additionally, the area corresponding to low-depth k-mers (5x-15x) was considerably smaller in axillary bud reads than in node reads. Some 27-mers were overrepresented in copy number in both the heterozygous and homozygous distributions, with heterozygous 27-mers present in more than one copy and homozygous 27-mers in more than two copies in the assembly (Fig. 2A and 2B). The orange

arrow in Fig. 2C shows that a fraction of 27-mers observed in the nodes sequencing data (V1 assembly) were absent from the V2 assembly.

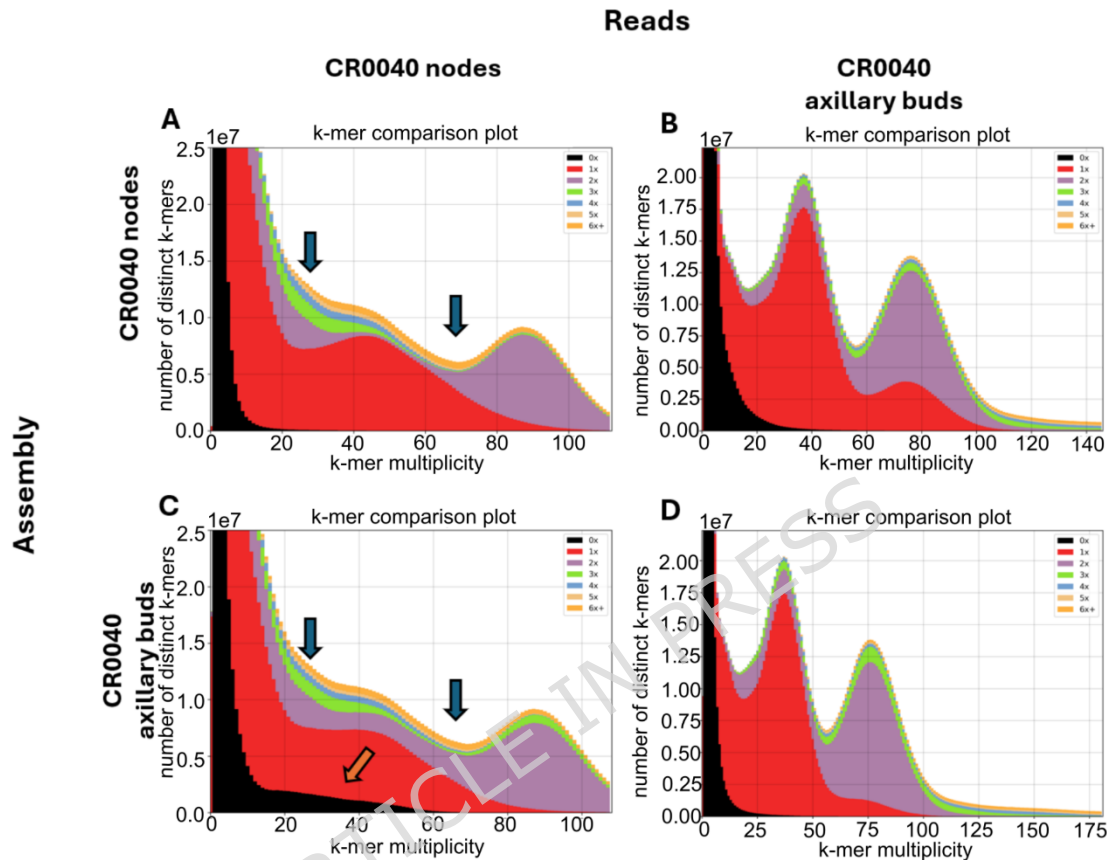


Figure 2. K-mers content comparison in sequencing reads and assemblies produced from DNA extracted from nodes and axillary buds of *V. planifolia* CR0040. Four spectra-cn plots are presented using $k=27$: (A) CR0040 nodes reads versus CR0040 V1 assembly, (B) CR0040 axillary buds reads versus CR0040 V1 assembly, (C) CR0040 nodes reads versus CR0040 V2 assembly, (D) CR0040 axillary buds reads versus CR0040 V2 assembly. Each histogram represents the comparison between the 27-mers found in one of the two PacBio HiFi sequencing datasets versus the 27-mers found in one of the two CR0040 assemblies. The histogram presents for each coverage the number of distinct 27-mers (y-axis) present n times in the sequencing dataset (x-axis). Cumulative distributions are colored based on the number of times the corresponding 27-mers are found in the assembly (black : 0x, red:

1x, purple: 2x, green: 3x, blue: 4x, beige: 5x and orange: 6x). Blue arrows point towards k-mers showing intermediate depth levels compared to the expected heterozygous and homozygous distributions which correspond to the two peaks in each panel. The orange arrow points toward a missing fraction (as shown by the black-colored distribution) of k-mers in the present assembly but found in CR0040 nodes sequencing data.

GBS-SNPs markers combined with an Omni-C approach allowed to scaffold the 16 pairs of chromosomes (85.89% of the total assembly) from a total of 2,708 contigs with 1,794 and 914 contigs for haplotypes A and B respectively (Fig. 3A and 3B). The final assembly had a cumulative size of 3.53 Gb with a N50 of 105.76 MB organized into 16 pairs of chromosomes and 286 and 158 unanchored sequences for haplotype A and B, respectively (Table 1). The assembly completeness was estimated at 96% for haplotype A and 95% for haplotype B using the 425 conserved single-copy orthologous genes from the Viridiplantae (viridiplantae_odb10) BUSCO lineage. Additionally, the completeness was estimated at 91.3% and 90.2% using 3,236 conserved single-copy orthologous genes in Liliopsida for haplotype A and B, respectively (Fig. 3C). Following previously reported recommendations [61], pairs of chromosomes were labeled from chr1 to chr14 following previous nomenclatures [16, 17] except for the pair of chromosome 1 that was separated into three pairs of chromosomes following the information provided by the genetic map [19]. The largest portion was kept as chr1 while the two others were labeled chr15 and chr16. A total of 27 and 20 telomeric regions were identified in haplotype A and B, respectively (Supplementary Table S2).

Five assembled chromosomes (chr1, 3, 4, 14 and 15) are still lacking the identification of a second telomere (combining haplotype A and B information).

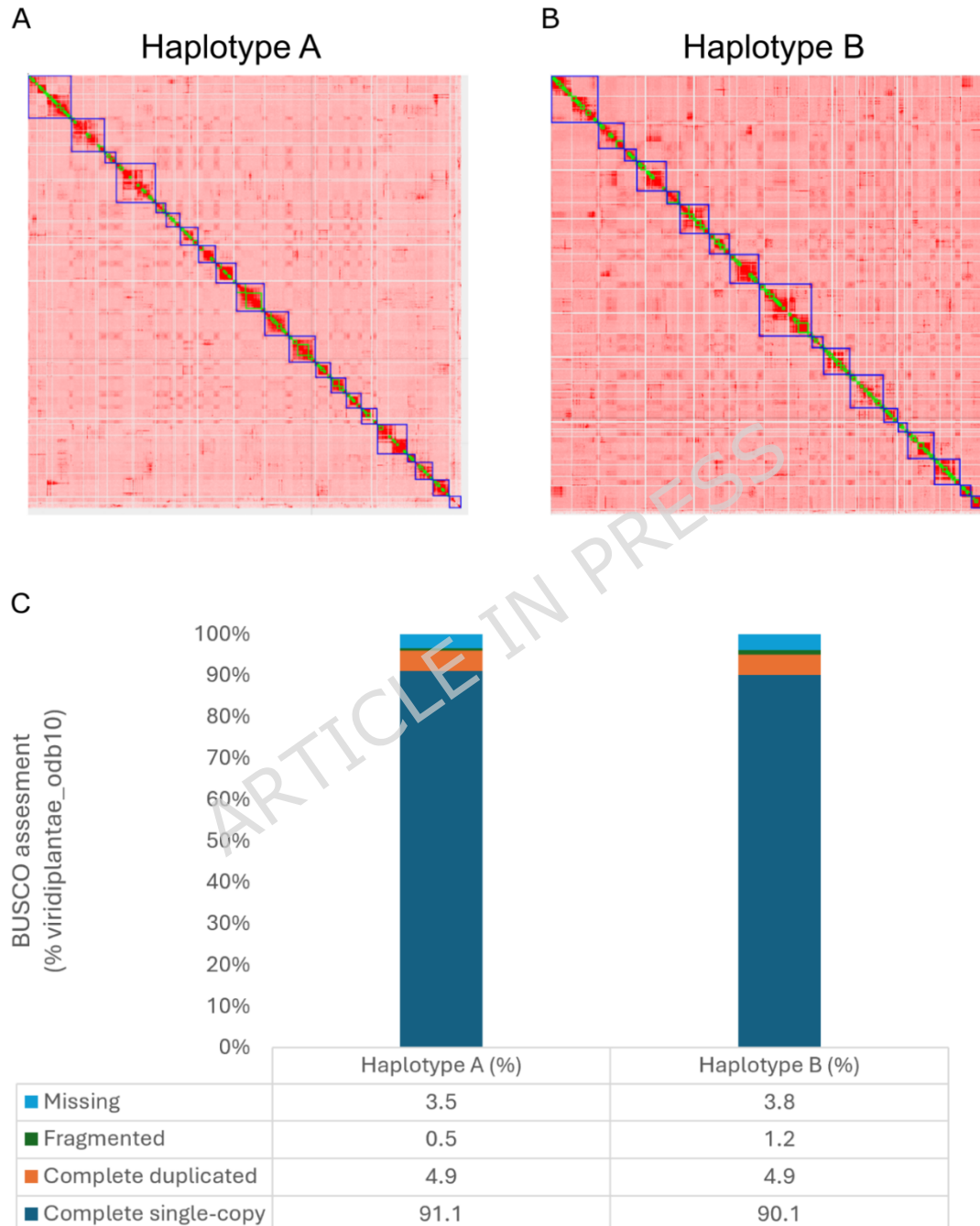


Figure 3. Omni-C contact maps from axillary buds of *V. planifolia* CR0040 haplotypes. (A) haplotype A and (B) haplotype B. The intensity of chromatin contacts

between different genomic regions are indicated by the intensity of red in the Omni-C matrix. Blue frames: the 16 pairs of chromosomes (identified using the genetic map) and seven additional large unanchored scaffolds (positioned at the right end of the map); green frames: contigs from hifiasm assembly. (C) Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment of the assembly completeness.

A total of 63,872 protein encoding genes were annotated and the 16 pairs of chromosomes contained the majority of these genes with 29,828 for haplotype A and 29,293 for haplotype B, representing 92.56% of the total annotated protein-encoding genes (Fig. 4 and Supplementary Table S3-S4). The remaining 4,751 protein-encoding genes were not anchored to the chromosomes.

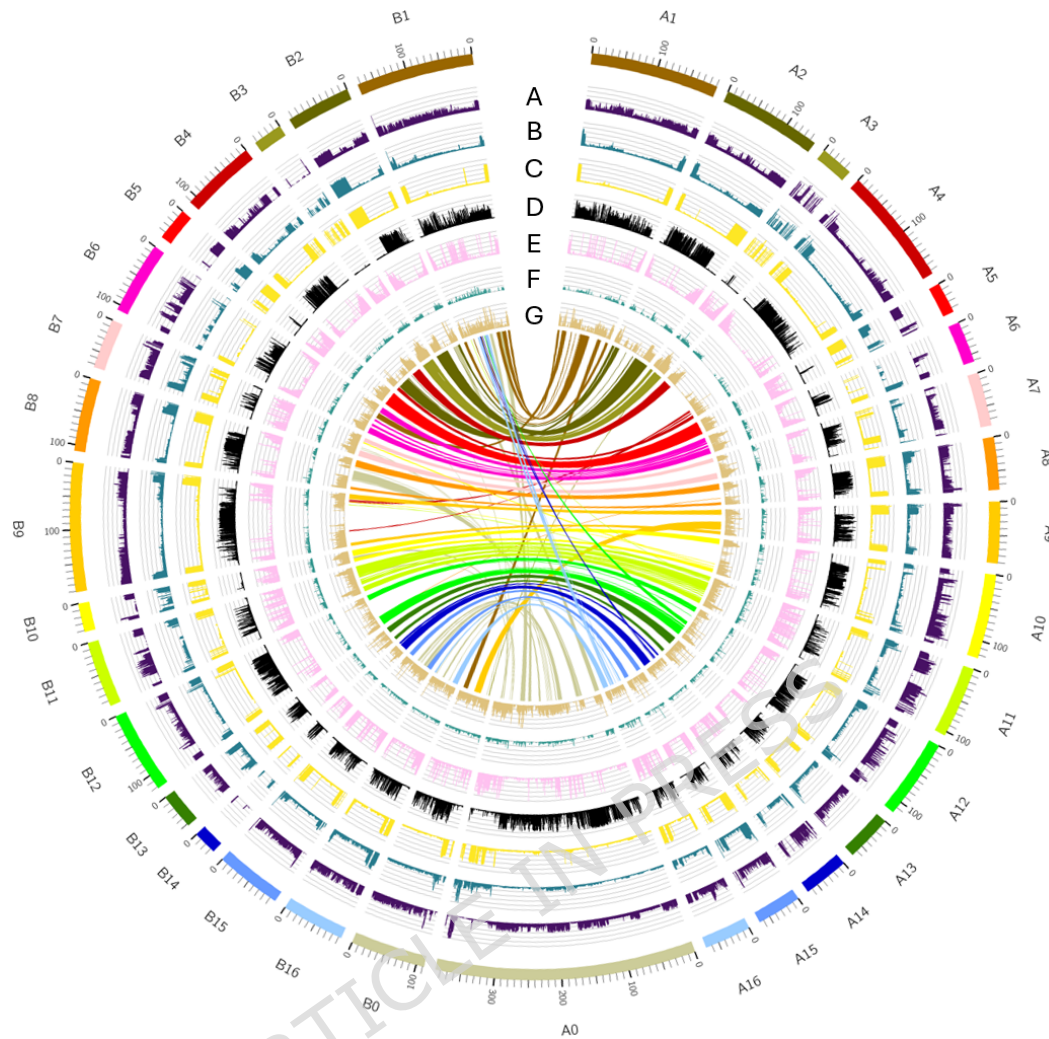


Figure 4. Circular representation of *V. planifolia* CR0040 phased assembly and its characteristics. The outer track represents the 16 chromosome pairs (with units in megabases). Unanchored sequences were concatenated and represented as two mosaic chromosomes (A0 and B0). Each colored link in the center represents a block of collinear genes between chromosomes (a total of 40,656 genes). Histograms in interior tracks represent genomic windows of 200 Kb (A) median values of mapping depth of CR0040 axillary buds PacBio HIFI reads, (B) median values of mapping depth of CR0040 nodes PacBio HIFI reads, (C) median values of mapping depth of Daphna leaves Illumina reads, (D) 1-6 nucleotides SSR density, (E) gene density, (F) DNA transposons density, (G) retrotransposons density.

The gene annotation completeness was estimated at around 93% for both haplotypes using the 425 conserved single-copy orthologous genes present in the Viridiplantae (viridiplantae_odb10) BUSCO lineage dataset (Supplementary Table S5). Collinearity analysis between both haplotypes showed that 40,656 genes were collinear. We identified a total of 2,078 repeats, including one satellite DNA, 1,228 transposable elements (TEs) and 849 unknown repeats (Supplementary Table S6) covering 39.85% of the genome. Excluding unknown repeats and ambiguously classified elements, LTR retrotransposons are the most abundant TE order (5.9% of the genomic sequence) with the predominant Gypsy superfamily (2.97%), followed by Copia (2.07%). LINE, SINE and other retrotransposons spanned 2.63%, 0.64% and 1.19% of the assembly, respectively. DNA transposon class II was rarer, representing 1.15% of the assembly. A total of 10,584,834 SSRs were identified, covering 6.87% of the diploid assembly sequence (Supplementary Table S4 and Supplementary Table S7). Low proportions of SSR were observed for chromosomes A3/B3, A5/B5 and B10. Additionally, the most prevalent class was trinucleotide SSR that comprised over 90% of the SSR identified for each haplotype (Supplementary Table S7). The motif AAG/CTT was the most common and represented 95.4% and 95.3% of the trinucleotides identified in haplotype A and B, respectively (Supplementary Table S8).

Table 1. Final assembly and annotation statistics of the diploid *V. planifolia* CR0040 genome

Total assembly size (Gb)	3.53
Total sequences number	476
N50 (Mb)	105.76
Maximum sequence length (Mb)	191.23
Sequence in chromosomes (% sequence)	86.32
GC content (%)	31.87
Number of protein-encoding genes	63872
BUSCO completeness	96.7
Total interspersed repeats content (% sequence)	32.98
Total SSR content (% sequence)	6.87

Gb, Giga bases; Mb, Mega bases

Distribution of F and P genomic fractions along the 16 pairs of chromosomes

Sequencing depth variations along the 16 chromosome pairs were analysed to identify F and P genomic fractions. The differential in sequencing depth between the two fractions varied by tissue (Fig. 4 and 5B), with the highest difference observed in *Daphna* leaves, followed by CR0040 nodes and CR0040 axillary buds. A conserved pattern was observed for most chromosomes, regardless tissue type. K-means clustering on the normalized depth median values per 200 Kb windows across the whole genome allowed the inference of F and P regions in the assembly. The normalized depth centroids for *Daphna* leaves median depth values were 0.26X for the F fraction and 4.28X for the P

fraction, with a midpoint of 2.27X (corresponding to 59.21X in raw depth for the dataset). Low sequencing depth regions (F fraction) were predominantly located in the central part of chromosomes whereas high sequencing depth regions (P fraction) were mostly found at chromosome ends (Fig. 4) with the exception of the A3/B3 and A5/B5 pairs (Fig. 5A).

ARTICLE IN PRESS

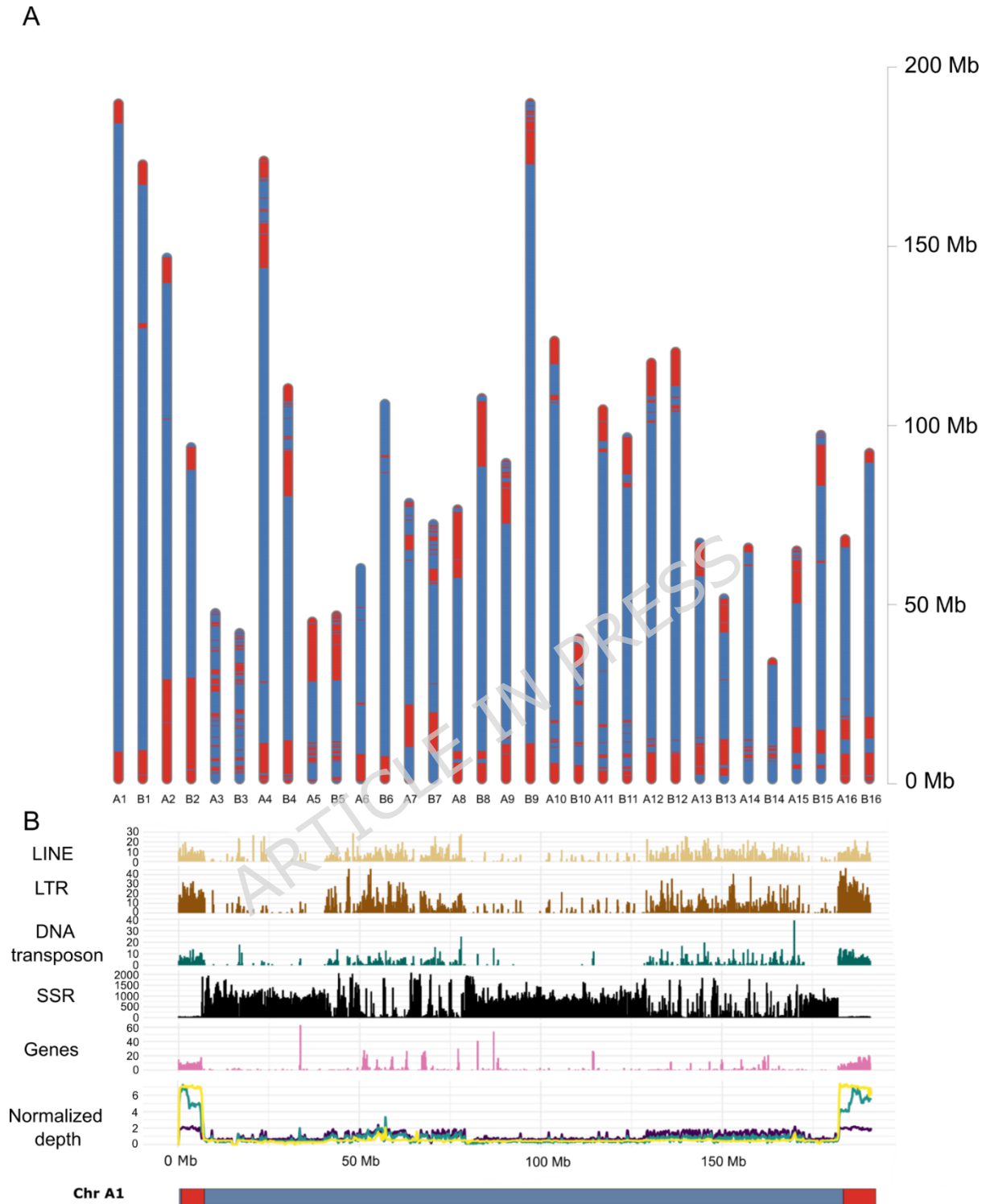


Figure 5. (A) Distribution of F and P regions along the 16 pairs of chromosomes and (B) detailed view of chromosome A1. F and P regions are indicated in blue and red respectively,

along chromosomes represented by ideograms. The size bar is shown on the right in megabases (Mb). Distributions of LINE, LTR, DNA transposons, SSR and genes counts are indicated with beige, brown, green, black and pink histograms, respectively. Normalized sequencing depth from *Daphna* leaves (Illumina), CR0040 nodes (PacBio HIFI) and CR0040 axillary buds (PacBio HIFI) reads are indicated with yellow, green and purple lines, respectively. Size of chromosome A1 is indicated above the ideogram in Mb.

A similar pattern was observed in the Omni-C data which revealed a compartmentalization between large central regions and smaller terminal regions for all the chromosomes except for pairs A5/B5 and A3/B3 (Supplementary Fig. S1). The F fraction comprised a total of 2.41 Gb (1.23 Gb in haplotype A and 1.18 Gb in haplotype B), representing 79.43% of chromosomes (Supplementary Table S9). Alignment of the 16 pairs of chromosomes of the present assembly with the 14 pairs of the V1 assembly showed that the previously assembled chromosomes corresponded entirely to the P genomic fraction in the current assembly (Fig. 5A and Supplementary Fig. S2), indicating that missing or unanchored sequences in the previous assembly corresponded to the F fraction.

Sequence composition of P and F genomic fractions

Gene and TE density were higher in the P fraction, whereas SSR were predominantly located in the F fraction, mostly in regions exhibiting lower sequencing depth across all three tissues studied (Fig. 4 and Fig. 5B). The different TE types (LINE, LTR, DNA transposons) appeared nested in similar locations along the chromosomes. Furthermore, TEs were found abundant in

the P fraction together with genes, but were also found in the F fraction although only where the density of SSR is lower, as if there were differential preferential locations for SSR versus TEs (Fig. 5B). For each type of features, their distribution between F and P genomic fractions significantly deviated from a random expectation (Table 2).

Table 2. Preferential distribution of *V. planifolia* genomic features in the two genomic fractions.

	Observed in P fraction	Observed in F fraction	Expected in P fraction	Expected in F fraction	p-value
Number of genes	27421	31877	12199	47098	***
Number of TE	383842	728146	228769	883218	***
Number of SSR	246798	8685987	1837743	7095042	***

***, Chi-2 test p-value < 0.001

SSR were more abundant in the F fraction than randomly expected whereas genes and TEs (including each major TE category, Supplementary Fig. S3) were underrepresented in this fraction and more overrepresented in the P fraction relative to random expectation.

The GO enrichment analysis of the F and P fractions highlighted the preferential retention of specific functional categories within each fraction. Fig. 6A shows the ten most significantly enriched GO terms in the F fraction, selected from a total of 58 overrepresented categories. They were related to histone methylation (GO:0016571, GO:0051569, GO:0016593, GO:0000993), cell cycle (GO:0044772), cytoskeletal activity (GO:0005872, GO:0009971, GO:0008569), negative flower regulation and regulation of cyclin-dependent kinase activity. The P fraction was enriched for a distinct set of GO categories, including those associated with defense response, protein folding and unfolded protein binding, terpenes and terpenoids production, transcription and misc enzyme activities (Fig. 6B).

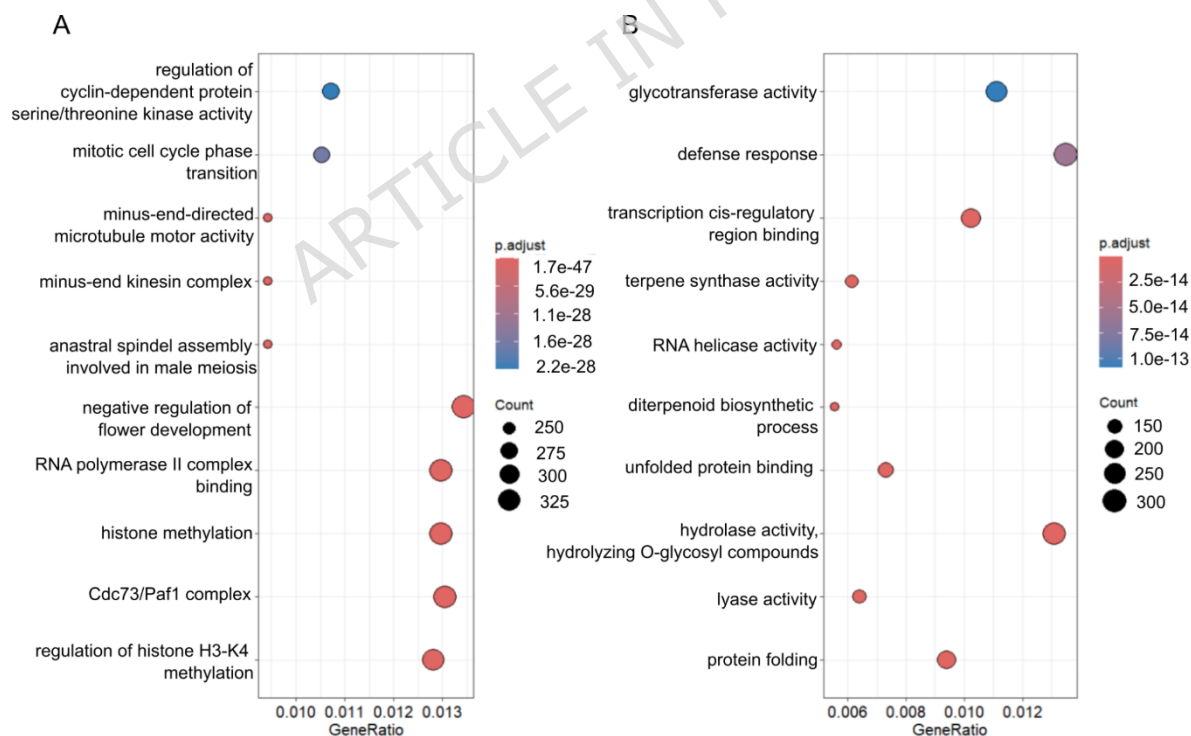


Figure 6. Functional enrichment analysis of (A) F regions and (B) P regions. The top ten significant GO enrichment results are illustrated for each fraction. Counts correspond to the

number of genes associated with each GO term. P.adjust values were calculated using Benjamini-Hochberg correction.

Discussion

Major improvements for the contiguity of *V. planifolia* assembly through PE impact consideration

The new *V. planifolia* assembly presented here encompasses over 86% of the genome size previously estimated by flow cytometry [16]. The mapping of telomeres across the 16 pairs of chromosomes revealed that five chromosomes were potentially lacking a second telomere. This observation may be explained by a combination of technical (missing telomeric regions in the current assembly) and biological (the occurrence of telocentric chromosomes in *V. planifolia*) factors. Previous metaphase observations indeed revealed that *V. planifolia* displays at least three telocentric chromosomes (corresponding to pairs 7, 11 and 13 on the karyotype in Fig. 2B in [16]) including one acrocentric pair carrying rDNA genes (pair 7). Fluorescence *in situ* hybridization (FISH) using telomeric probes could confirm the physical location of telomeres and detect telocentric chromosomes directly [62].

The discrepancy between the CR0040 V1 and Daphna assembly lengths was attributed to the impact of PE [16], inducing highly unbalanced DNA content within a given tissue and challenging the assembly of sequencing reads with highly unbalanced depths, particularly in low depth repetitive regions [63]. In

this study, HiFi sequencing of DNA extracted from an alternative tissue (axillary buds) confirmed the impact of PE on sequencing depth. The mapping of axillary buds and nodes HiFi reads on CR0040 V1 assembly showed a reduction in the gap in sequencing depths between anchored and unanchored sequences for axillary buds in comparison to nodes. This indicated that the representation bias affecting the unanchored scaffolds was reduced in axillary buds, a tissue that is characterized in flow cytometry by lesser levels of endoreplicated nuclei compared to nodes. This observation was also confirmed by the mapping of DNA sequencing data from Daphna leaves (Illumina), CR0040 nodes (PacBio HIFI) and CR0040 axillary buds (PacBio HIFI) on the present CR0040 assembly. The sequencing depth difference between P and F fractions was substantially lower in axillary buds sequencing data than in nodes or leaves. The differential sequencing depth between the two fractions appeared to show a gradient across the three tissues that mirrored their respective endoreplication levels. Finally, the k-mers comparison of nodes and axillary buds showed an impact of tissue selection on k-mer content profiles, with observable k-mers of intermediate sequencing depths between the expected heterozygous and homozygous distributions for nodes.

Chromosomal distribution and potential functional specialization of F and P genomic fractions

Across both fractions along the 16 pairs of chromosomes, the first type (P fraction) corresponded to high-depth regions located mainly at chromosome

ends and enriched in genes and TEs. The second type (F fraction) corresponded to low-depth regions located in central chromosome regions and enriched in microsatellite sequences. This pattern is consistent with observations in other eukaryotic species, where SSR are reported to be more abundant in non-coding regions [64]. This compartmentalization of the two fractions was also observed in the chromosome contact maps derived from Omni-C data (Supplementary Fig. S1), suggesting a heterochromatin-euchromatin organization, with central regions composed of highly repetitive heterochromatin and peripheral regions consisting of gene-rich euchromatin. The F fraction represented 79.43% of the chromosome sequences which is close to the value previously estimated by flow cytometry ($69.5\% \pm 3.2\%$) for CR0040 [16]. Our strategy efficiently revealed, for the first time, the broad distribution of PE regions along the chromosomes, although 14% of the estimated genome size remains missing and some scaffolds have yet to be anchored to chromosomes.

A total of 59,121 genes (92.56% of total predicted genes) were anchored on the 16 pairs of chromosomes with preferential positions at the chromosome ends (P fraction). The P fraction was enriched in genes, but this does not mean that genes were totally absent from the F fraction. There were, in fact, more genes annotated in the P fraction than in the F one, but this number was rather low compared to the size of the F fraction and some GO categories were significantly found overrepresented in F regions. The top ten GO categories are related to three main biological processes which are histone modification,

cell cycle and cell division and regulation of floral development. Based on GO terms, many gene families involved in epigenetic mechanisms have been retrieved in these F regions. These have a central role in heterochromatic gene silencing, either constitutive or transient. Histone methylation is known to play a fundamental role in epigenetic regulation of the expression of numerous genes in plant responses to abiotic stresses [65]. The mechanisms involved in the regulation of flowering have been particularly studied in *Arabidopsis*, in perennial and annual species [66-68]. The FLOWERING LOCUS C (FLC), or PEP1 in perennial, encodes a central repressor of the entire molecular cascade regulating the transition from a vegetative state to a reproductive state, and its strong expression from germination blocks all flowering. Thus, it maintains the plant in a vegetative state as long as it is not itself inhibited by combined exposure to low temperature and short photoperiods which will thus allow the activation by the circadian cycle of the flowering trigger factor. Epigenetic modifications such as triggering of chromatin modifications allow for a stably transcription repression of FLC or PEP1. The enrichment of genes with functions related to flowering regulation in F regions of *V. planifolia* could suggest that these regions correspond to facultative heterochromatin, i.e. a temporarily repressive chromatin state, in order to temporally regulate the expression of genes responsible for these specific cellular programs key for plant reproduction [69-71]. The enrichment of defense-related genes in the P fraction suggests a potential functional bias toward biologically active regions. In plants, endoreplication has been associated with increased transcriptional

capacity and gene amplification [1, 72], as well as with responses to environmental and stress conditions [73, 74], which could support the preferential amplification of genes involved in adaptive processes, including defense. However, to our knowledge, no direct evidence currently demonstrates a specific association between the P fraction and immune defense in *V. planifolia*, and this hypothesis remains to be further investigated.

The annotation strategy integrates a combination of transcriptomic and proteomic evidence, as well as *ab initio* methods. Recent deep learning-based tools such as Helixer [75] and ANNEVO [76] have demonstrated improved performance in gene prediction across multiple species. Incorporating such approaches in future analyses could further refine gene models and provide a more comprehensive view of the gene content in both P and F regions.

Chromatin type as a marker of PE chromosome landscape

V. planifolia chromosomes possess important portions of telomeric and pericentromeric unspecific heterochromatin resulting in many chromocenters visible in interphase nuclei after DAPI and Orcein staining [16]. Hoechst staining revealed that A-T rich heterochromatin regions were more numerous than G-C rich regions [16]. The telomeric A-T rich heterochromatin could correspond to the *V. planifolia* telomeric repeats (ATTAGGG) whereas pericentromeric A-T rich heterochromatin could correspond to the high density AAG/CTT microsatellites observed in the central part of most chromosomes (F fraction). This also confirms the first hypothesis raised in *Vanilla* [77] and later

in orchids [4, 15] stating that the F fraction likely corresponds to heterochromatin regions rich in repeated sequences. It was also shown that genome size expansion in PE and CE orchids was due to repeated element proliferation, which was limited to the F fraction for PE species [7]. Our results show that this repeated element proliferation in *V. planifolia* concerns mainly the AAG/CTT repeats and some TEs (65.48%) located in the F fraction. Finally, our results also demonstrated that the P fraction corresponds to gene-rich euchromatin regions in the *V. planifolia* genome. This is consistent with nuclei imaging experiments showing increased transcriptional activity with increased level of *Vanilla* endoreplicated nuclei, therefore confirming that the endoreplicated fraction was transcribed [5]. Interestingly, their analysis of base composition showed a trend with a possible increase in chromatin content (G-C rich) compared to A-T rich regions which tended to stay fixed [5].

AAG/CTT repeats are prone to expansion in *V. planifolia* genome and may be potential drivers of endoreplication suppression

SSR abundance in *V. planifolia* (6.87% of the genome) was higher than in various land plants [78, 79]. *V. planifolia* presents a very specific microsatellite motif AAG/CTT which has invaded the F regions of the chromosomes. This motif represents over 90% of the total SSR and 95% of the trinucleotide microsatellites identified in the genome. This non-random genomic distribution and abundance indicates a major functional role. Most SSR are neutral, but some are well known for their functional significance [64]. SSRs,

when able to adopt secondary structures, can impede DNA replication and promote chromosomal fragility, genomic instability, and DNA breakage [80, 81]. No such tri-nucleotide pattern is, to our knowledge, known in plants. In plant species, trinucleotide repeats, including AAG/CTT, are only found more frequent when considering coding regions [1, 82, 83], as in the case of human trinucleotide repeats [84]. and the motif is never unique. Such sequences known as TRS (Triplet Repeat Sequences) are well studied in the human genome. The expansion of TRS has been correlated with about 50 human genetic disorders (triplet repeat expansion diseases (TREDs)) such as fragile X syndrome and Huntington's disease [85-88]. Most early known TRS were CTG/CAG and CGG/CCG, but the intronic AAG/CTT was identified later as responsible for Friedreich's ataxia (FRDA) [86, 87, 89, 90]. These TRS are known to adopt atypical secondary structures inducing genetic instability through errors in DNA replication [85]. Indeed, biophysical and biochemical studies have revealed that five (CGG, CCG, CAG, CTG and AAG) of the six TRS that are associated with human diseases can form stable hairpin structures and are therefore likely to undergo expansion through replication slippage, responsible for the observed diseases [86]. This might also therefore explain the incredible expansion of the AAG/CTT sequence observed in the *V. planifolia* genome.

In addition to hairpin structures, some of these sequences (AAG/CTT and AGG/CCT) can also form other non-B DNA secondary structures such as intramolecular triplexes (triple-stranded DNA) *in vitro* [86, 88, 89]. The triplex

structure can stop DNA synthesis [87] and generate pausings of DNA polymerase *in vitro* [89]. Finally, AAG/CTT repeats can also form secondary R-loops (RNA:DNA structures) which were suggested to be responsible together with triplexes for the reduction of abundance of mature X25 mRNA in individuals with FRDA [88, 89]. Additionally, in *Drosophila*, heterochromatic satellite repeats have been presented as a barrier to DNA replication and correspond to underreplicated regions in endopolyploid cells, leading to formation of polytene chromosomes [91]. In a plant genome, SSR-rich regions can act as obstacles to DNA replication. Consequently, the numerous AAG/CTT repeats identified in the non-endoreplicated F fraction of *V. planifolia* may function as replication barriers. They could be directly responsible via triplex structure formation for the lack of endoreplication of the fraction containing them. It would be interesting to investigate whether similar repeats are present in the F fraction of other PE orchids.

Conclusions

This study reports the first chromosome-scale genome assembly of *V. planifolia*, resolved into 16 chromosome pairs, representing the most complete assembly available to date. This resource enabled new insights into the structural organization of PE and the genomic composition of its F and P fractions along the chromosomes, confirming the high gene density characteristic of the P fraction and revealing the SSR-rich nature of the F fraction. The availability of this improved reference genome paves the way to future physical characterization based on chromosome FISH (fluorescent in

situ hybridization), by designing and using synthetic oligonucleotide (oligo) probes [92, 93] specific for the P fraction on each chromosome, and by using a AAG/CTT probe to confirm the F fraction localization in situ. In addition, this work provides practical guidelines to overcome the challenges imposed by PE during genome sequencing and assembly, notably through the use of flow cytometry to select appropriate tissues. Finally, the assembled genome, together with the delineation of endoreplicated and non-endoreplicated regions, establishes a valuable foundation for future investigations of PE in orchid genomes.

Declarations

Abbreviations

Bp: basepairs; BUSCO; Benchmarking Universal Single-Copy Orthologs; CCS: circular consensus reads; Gb: gigabyte bases; Mb: million bases; NCBI: National Center for Biotechnology Information; PE: partial endoreplication; SSR: Simple sequence repeats; TE: transposable element(s)

Ethics approval and consent to participate

The authors confirmed that the experimental research on plants including the collection of plants comply with relevant institutional, national, and international guidelines and legislation. No plant seeds were used in our study. The initial cuttings of the CR0040 *Vanilla planifolia* cultivar were collected from CIRAD agricultural plots in 2003 by Michel Grisoni (CIRAD) in La Reunion,

France. This cultivated material is not endangered or protected materials and it complies with phytosanitary regulations (LOA permit n°03/RE2/01555). All plant materials used in this experiment were from in vitro tissue culture plantlets produced from cuttings in our laboratory (UMR PVBMT, Pole de Protection des Plantes, Saint Pierre, La Reunion, France). All samples were stored in the VATEL Biological Resource Center (BRC) collection (Saint Pierre, La Reunion, France), Michel Grisoni identified it. The BRC Vatel collection is recognized by the relevant authorities (with IBISA Label and ISO9001 certification).

Consent for publication

Not applicable

Availability of data and materials

The raw sequencing data generated in this study, including PacBio HiFi CCS and Omni-C reads, are available in the NCBI Sequence Read Archive under BioProject PRJNA1254293. The assemblies and associated datasets have been deposited in NCBI under BioProjects PRJNA1254293 (haplotype A) and PRJNA1254300 (haplotype B). In addition, the assemblies, gene annotations and predicted protein sequences are also publicly accessible via the Vanilla Genome Hub (<https://vanilla-genome-hub.cirad.fr/downloads>).

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by Eurovanille and V. Mane Fils companies, the Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), the Université de la Réunion (UR), the Institut National de Recherche pour l'Agriculture l'Alimentation et l'Environnement (INRAE) and the Centre National de la Recherche Scientifique (CNRS). This work was also co-funded by the European Union, the French Government and the Région Réunion. L'Europe s'engage à La Réunion avec le FEDER et le FEADER (Nos. FEDER-23-26 REU005756, 2024-1248-005756 and REU77071-1-000011). Quentin Piet is the recipient of a PhD fellowship funded by the European Union and the Région Réunion.

Authors' contributions

C.Ch., C.J., M.D., M.G. and P.B. designed the project. A.D., C.J., F.L., H.M.G. and M.D. managed the project administration. C.J. supervised the research activity. M.B. performed flow cytometry analyses. Q.P. and W.M. produced biological materials. W.M. extracted HMW DNA. O.B. and M.Sa. generated PacBio HIFI DNA sequencing. N.R. produced Omni-C data. C.K. and Q.P. analyzed raw sequencing data and assembled the genome. C.Cr., Q.P. and S.B. performed genome annotations. Q.P. performed mapping analysis. G.D. and M.Su. developed the vanilla genome hub. C.Ch., C.J., Q.P. and P.B. designed the draft manuscript. All authors contributed to write, review and edit the final manuscript.

Acknowledgments

We thank the Plant Protection Platform (3P, IBISA) for laboratory facilities and access to plant resources (BRC Vatel) as well as the GeT-PlaGe platform (INRAE, Toulouse, France) for providing sequencing facilities. This work also benefited from the support of the ISDM-MESO Platform at the University of Montpellier and the technical assistance of the bioinformatics team at the UMR AGAP Institute, a member of the French Institute of Bioinformatics - South Green Bioinformatics Platform. The study made use of the Imagerie-Gif core facility, supported by the Agence Nationale de la Recherche (FBI ANR-24-INBS-0005 (BIOGEN); SPS ANR-17-EUR-0007, EUR SPS-GSR).

Supplementary information

Supplementary Table S1. Comparison of the anchored and unanchored *V. planifolia* CR0040 V1 assembly regions sequencing depths computed from primary alignments of CR0040 nodes HIFI reads and CR0040 axillary buds HIFI reads.

Supplementary Table S2. Structural metrics of *V. planifolia* CR0040 chromosomes.

Supplementary Table S3. Summary statistics of protein coding genes per haplotype.

Supplementary Table S4. Distribution of protein-encoding genes, TEs and SSR (1-6 nucleotides) per sequence in each haplotype of CR0040 assembly.

Supplementary Table S5. Completeness assessment of *V. planifolia* CR0040 final annotation using BUSCO.

Supplementary Table S6. Repeat annotation statistics.

Supplementary Table S7. Proportions of motif lengths in the SSR found in CR0040 V2 assembly relative to the total number of SSR.

Supplementary Table S8. Proportions of trinucleotide motifs in CR0040 V2 haplotypes A and B relative to the number of trinucleotide SSR.

Supplementary Table S9. Sizes of F and P fraction in haplotype A and haplotype B.

Supplementary Fig. S1. Omni-C contact map of chromosomes A3, B5, A1 and A9 displayed with Knight-Ruiz normalization.

Supplementary Fig. S2. Dotplot of the alignment between the 14 chromosomes from CR0040 V1 haplotype A (y-axis) and the 16 chromosomes from CR0040 V2 haplotype A (x-axis).

Supplementary Fig. S3. Preferential distribution of *Vanilla planifolia* transposable elements between F and P fractions across all 16 pairs of assembled chromosomes. (***) p-value < 0.001.

References

1. Edgar BA, Orr-Weaver TL. Endoreplication cell cycles: more for less. Cell. 2001;105(3):297-306. doi:10.1016/s0092-8674(01)00334-8

2. Bory S, Catrice O, Brown S, et al. Natural polyploidy in *Vanilla planifolia* (Orchidaceae). *Genome*. 2008;51(10):816-826. doi:10.1139/G08-068
3. Trávníček P, Ponert J, Urfus T, et al. Challenges of flow-cytometric estimation of nuclear genome size in orchids, a plant group with both whole-genome and progressively partial endoreplication. *Cytometry A*. 2015;87(10):958-966. doi:10.1002/cyto.a.22681
4. Hřibová E, Holušová K, Trávníček P, et al. The enigma of progressively partial endoreplication: new insights provided by flow cytometry and next-generation sequencing. *Genome Biol Evol*. 2016;8(6):1996-2005. Published 2016 Jul 2. doi:10.1093/gbe/evw141
5. Brown SC, Bourge M, Maunoury N, et al. DNA remodeling by strict partial endoreplication in Orchids, an original process in the Plant Kingdom. *Genome Biol Evol*. 2017;9(4):1051-1071. doi:10.1093/gbe/evx063
6. Leitch IJ, Dodsworth S. Endopolyploidy in plants. Hetherington A, ed. *Encyclopedia of life sciences*. 2017:1-10.
7. Trávníček P, Čertner M, Ponert J, Chumová Z, Jersáková J, Suda J. Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. *New Phytol*. 2019;224(4):1642-1656. doi:10.1111/nph.15996

8. Trávníček P, Chumová Z, Záveská E, et al. Integrative Study of Genotypic and Phenotypic Diversity in the Eurasian Orchid Genus *Neotinea*. *Front Plant Sci.* 2021;12:734240. Published 2021 Oct 13. doi:10.3389/fpls.2021.734240
9. Stormo BM, Fox DT. Polyteny: still a giant player in chromosome research. *Chromosome Res.* 2017;25(3-4):201-214. doi:10.1007/s10577-017-9562-z
10. Carnevalheira GMG. Plant polytene chromosomes. *Genet Mol Biol.* 2000;23:1043-1050
11. Rzeszutek I, Maurer-Alcalá XX, Nowacki M. Programmed genome rearrangements in ciliates. *Cell Mol Life Sci.* 2020;77(22):4615-4629. doi:10.1007/s00018-020-03555-2
12. Wang J, Davis RE. Programmed DNA elimination in multicellular organisms. *Curr Opin Genet Dev.* 2014;27:26-34. doi:10.1016/j.gde.2014.03.012
13. Bateman RM, Guy JJ, Rudall PJ, Leitch IJ, Pellicer J, Leitch AR. Evolutionary and functional potential of ploidy increase within individual plants: somatic ploidy mapping of the complex labellum of sexually deceptive bee orchids. *Ann Bot.* 2018;122(1):133-150. doi:10.1093/aob/mcy048
14. Soto Arenas, M.A, Dressler, R.L. A revision of the Mexican and Central American species of *Vanilla* Plumier ex Miller with a characterization of their ITS region of the nuclear ribosomal DNA. *Lankesteriana.* 2010;9 (3), 285-354

15. Chumová Z, Záveská E, Hloušková P, et al. Repeat proliferation and partial endoreplication jointly shape the patterns of genome size evolution in orchids. *Plant J.* 2021;107(2):511-524. doi:10.1111/tpj.15306
16. Piet Q, Droc G, Marande W, et al. A chromosome-level, haplotype-phased *Vanilla planifolia* genome highlights the challenge of partial endoreplication for accurate whole-genome assembly. *Plant Commun.* 2022;3(5):100330. doi:10.1016/j.xplc.2022.100330
17. Hasing T, Tang H, Brym M, Khazi F, Huang T, Chambers AH. A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. *Nat Food.* 2020;1(12):811-819. doi:10.1038/s43016-020-00197-2
18. Felix LP, Guerra M. Basic chromosome numbers of terrestrial orchids. *Plant Systematics and Evolution.* 2005;254(3-4):131-148. doi:10.1007/s00606-004-0200-9
19. Favre F, Jourda C, Grisoni M, et al. First *Vanilla planifolia* high-density genetic linkage map provides quantitative trait loci for resistance to *Fusarium oxysporum*. *Plant Dis.* 2023;107(10):2997-3006. doi:10.1094/PDIS-10-22-2386-RE
20. Favre, F, Jourda, C, Grisoni, et al. A genome-wide assessment of the genetic diversity, evolution and relationships with allied species of the clonally propagated crop *Vanilla planifolia* Jacks. ex Andrews. *Genetic Resources and Crop Evolution.* 2022;69(6), 2125-2139. doi:10.1007/s10722-022-01362-1

21. Bourge, M; Brown, S.C; Siljak-Yakovlev, S. Flow cytometry as tool in plant sciences, with emphasis on genome size and ploidy level assessment. *Genet. Appl.* 2018, 2, 1-12.
22. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094-3100. doi:10.1093/bioinformatics/bty191
23. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics.* 2018;34(5):867-868. doi:10.1093/bioinformatics/btx699
24. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods.* 2021;18(2):170-175. doi:10.1038/s41592-020-01056-5
25. Astashyn A, Tvedte ES, Sweeney D, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* 2024;25(1):60. Published 2024 Feb 26. doi:10.1186/s13059-024-03198-7
26. Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics.* 2023;24(1):288. Published 2023 Jul 18. doi:10.1186/s12859-023-05385-y
27. Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. MitoFinder: Efficient automated large-scale extraction of mitogenomic data

in target enrichment phylogenomics. *Mol Ecol Resour.* 2020;20(4):892-905.
doi:10.1111/1755-0998.13160

28. Durand NC, Shamim MS, Machol I, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 2016;3(1):95-98.
doi:10.1016/j.cels.2016.07.002

29. Dudchenko O, Batra SS, Omer AD, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* 2017;356(6333):92-95. doi:10.1126/science.aal3327

30. Durand NC, Robinson JT, Shamim MS, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* 2016;3(1):99-101. doi:10.1016/j.cels.2015.07.012

31. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ.* 2018;6:e4958. Published 2018 Jun 4.
doi:10.7717/peerj.4958

32. Vasimuddin M, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS); 2019:314-324.
doi:10.1109/IPDPS.2019.00041

33. Li H, seqtk: toolkit for processing sequences in FASTA/Q formats. 2012.
<https://github.com/lh3/seqtk>

34. Závodník M, Fajkus P, Franek M, et al. Telomerase RNA gene paralogs in plants - the usual pathway to unusual telomeres. *New Phytol.* 2023;239(6):2353-2366. doi:10.1111/nph.19110
35. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38(10):4647-4654. doi:10.1093/molbev/msab199
36. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics.* 2017;33(4):574-576. doi:10.1093/bioinformatics/btw663
37. Sallet E, Gouzy J, Schiex T. EuGene: an automated integrative gene finder for Eukaryotes and Prokaryotes. *Methods Mol Biol.* 2019;1962:97-120. doi:10.1007/978-1-4939-9173-0_6
38. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37(8):907-915. doi:10.1038/s41587-019-0201-4
39. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 2015;33(3):290-295. doi:10.1038/nbt.3122

40. AGAP. Funcannotate. Gitlab.
<https://gitlab.cirad.fr/agap/workflows/funcannotate>. Accessed April 2024
41. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236-1240. doi:10.1093/bioinformatics/btu031
42. Aramaki T, Blanc-Mathieu R, Endo H, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*. 2020;36(7):2251-2252. doi:10.1093/bioinformatics/btz859
43. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6:11. Published 2015 Jun 2. doi:10.1186/s13100-015-0041-9
44. Ou S, Su W, Liao Y, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019;20(1):275. Published 2019 Dec 16. doi:10.1186/s13059-019-1905-y
45. McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*. 2003;19(3):362-367. doi:10.1093/bioinformatics/btf878
46. Orozco-Arias S, Humberto Lopez-Murillo L, Candamil-Cortés MS, et al. Inpactor2: a software based on deep learning to identify and classify LTR-retrotransposons in plant genomes. *Brief Bioinform*. 2023;24(1):bbac511. doi:10.1093/bib/bbac511

47. Hoede C, Arnoux S, Moisset M, et al. PASTEC: an automatic transposable element classification tool. PLoS One. 2014;9(5):e91929. Published 2014 May 2. doi:10.1371/journal.pone.0091929
48. Quesneville H, Bergman CM, Andrieu O, et al. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 2005;1(2):166-175. doi:10.1371/journal.pcbi.0010022
49. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-1659. doi:10.1093/bioinformatics/btl158
50. Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett. 1999;174(2):247-250. doi:10.1111/j.1574-6968.1999.tb13575.x
51. Russo A, Alessandrini M, El Baidouri M, et al. Genome of the early spider-orchid *Ophrys sphegodes* provides insights into sexual deception and pollinator adaptation. Nat Commun. 2024;15(1):6308. Published 2024 Jul 26. doi:10.1038/s41467-024-50622-4
52. Flynn JM, Hubley R, Goubert C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117(17):9451-9457. doi:10.1073/pnas.1921046117

53. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics*. 2005;21 Suppl 1:i351-i358.
doi:10.1093/bioinformatics/bti1018
54. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*. 2007;23(8):1026-1028.
doi:10.1093/bioinformatics/btm039
55. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet*. 2003;106(3):411-422.
doi:10.1007/s00122-002-1031-0
56. Wang Y, Tang H, Debarry JD, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012;40(7):e49. doi:10.1093/nar/gkr1293
57. Henry IM, Dilkes BP, Miller ES, Burkart-Waco D, Comai L. Phenotypic consequences of aneuploidy in *Arabidopsis thaliana*. *Genetics*. 2010;186(4):1231-1245. doi:10.1534/genetics.110.121079
58. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
doi:10.1093/bioinformatics/btp352

59. Neph S, Kuehn MS, Reynolds AP, et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics*. 2012;28(14):1919-1920. doi:10.1093/bioinformatics/bts277
60. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*. 2021;2(3):100141. Published 2021 Jul 1. doi:10.1016/j.xinn.2021.100141
61. Howe K, Chow W, Collins J, et al. Significantly improving the quality of genome assemblies through curation. *Gigascience*. 2021;10(1):giaa153. doi:10.1093/gigascience/giaa153
62. Scherthan H. Detection of chromosome ends by telomere FISH. *Methods Mol Biol*. 2002;191:13-31. doi:10.1385/1-59259-189-2:13
63. Garg S. Computational methods for chromosome-scale haplotype reconstruction. *Genome Biol*. 2021;22(1):101. Published 2021 Apr 12. doi:10.1186/s13059-021-02328-9
64. Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol*. 2002;11(12):2453-2465. doi:10.1046/j.1365-294x.2002.01643.x
65. Shi L, Cui X, Shen Y. The roles of histone methylation in the regulation of abiotic stress responses in plants. *Plant Stress*. 2024;11(100303):100303. doi:10.1016/j.stress.2023.100303

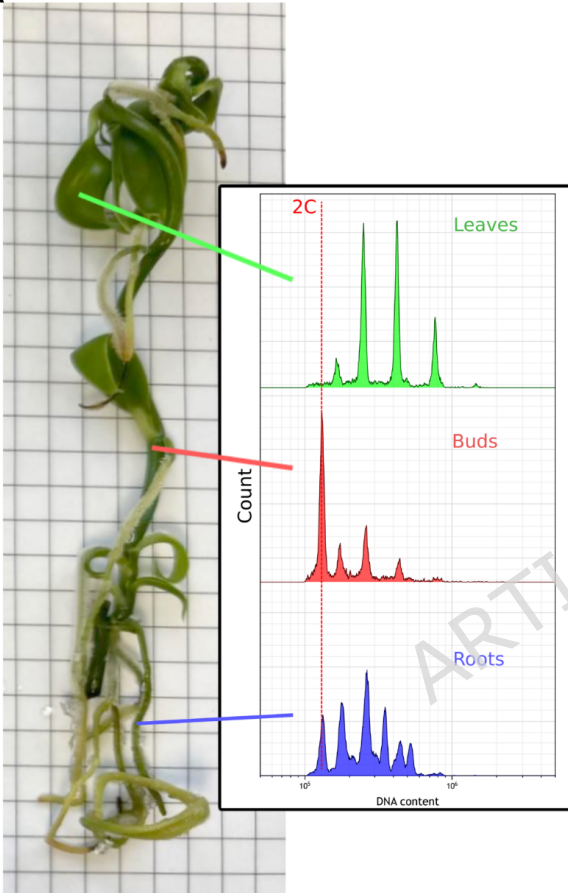
66. Quiroz S, Yustis JC, Chávez-Hernández EC, et al. Beyond the genetic Pathways, flowering regulation complexity in *Arabidopsis thaliana*. Int J Mol Sci. 2021;22(11):5716. Published 2021 May 27. doi:10.3390/ijms22115716
67. Takagi H, Hempton AK, Imaizumi T. Photoperiodic flowering in *Arabidopsis*: multilayered regulatory mechanisms of CONSTANS and the florigen FLOWERING LOCUS T. Plant Commun. 2023;4(3):100552. doi:10.1016/j.xplc.2023.100552
68. Wang R, Farrona S, Vincent C, et al. PEP1 regulates perennial flowering in *Arabis alpina*. Nature. 2009;459(7245):423-427. doi:10.1038/nature07988
69. Song J, Angel A, Howard M, Dean C. Vernalization - a cold-induced epigenetic switch. J Cell Sci. 2012;125(Pt 16):3723-3731. doi:10.1242/jcs.084764
70. Yang H, Berry S, Olsson TSG, Hartley M, Howard M, Dean C. Distinct phases of Polycomb silencing to hold epigenetic memory of cold in *Arabidopsis*. Science. 2017;357(6356):1142-1145. doi:10.1126/science.aan1121
71. Zhao F, Zhang H, Zhao T, Li Z, Jiang D. The histone variant H3.3 promotes the active chromatin state to repress flowering in *Arabidopsis*. Plant Physiol. 2021;186(4):2051-2063. doi:10.1093/plphys/kiab224
72. Sugimoto-Shirasu K, Roberts K. "Big it up": endoreduplication and cell-size control in plants. Curr Opin Plant Biol. 2003;6(6):544-553. doi:10.1016/j.pbi.2003.09.009

73. Cookson SJ, Radziejwoski A, Granier C. Cell and leaf size plasticity in *Arabidopsis*: what is the role of endoreduplication?. *Plant Cell Environ.* 2006;29(7):1273-1283. doi:10.1111/j.1365-3040.2006.01506.x
74. De Veylder L, Beeckman T, Inzé D. The ins and outs of the plant cell cycle. *Nat Rev Mol Cell Biol.* 2007;8(8):655-665. doi:10.1038/nrm2227
75. Stiehler F, Steinborn M, Scholz S, Dey D, Weber APM, Denton AK. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics.* 2021;36(22-23):5291-5298. doi:10.1093/bioinformatics/btaa1044
76. Zhang P, Xu T, Wang S, et al. Highly accurate ab initio gene annotation with ANNEVO. *Nat Methods.* Published online March 12, 2026. doi:10.1038/s41592-026-03036-7
77. Kausch AP, Horner HT. Increased nuclear DNA content in raphide crystal idioblasts during development in *Vanilla planifolia* L. (Orchidaceae). *Eur J Cell Biol.* 1984;33(1):7-12.
78. Srivastava S, Avvaru AK, Sowpati DT, Mishra RK. Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics.* 2019;20(1):153. Published 2019 Feb 22. doi:10.1186/s12864-019-5516-5
79. Zhu L, Wu H, Li H, et al. Short tandem repeats in plants: genomic distribution and function prediction. *Electron J Biotechnol.* 2021;50:37-44. doi:10.1016/j.ejbt.2020.12.003

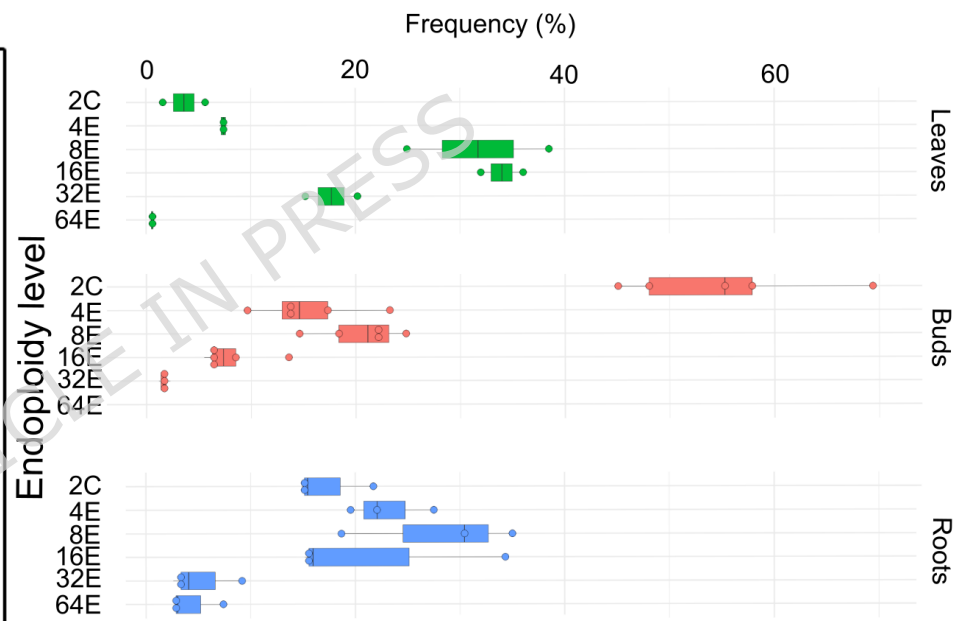
80. Mirkin SM. Expandable DNA repeats and human disease. *Nature*. 2007;447(7147):932-940. doi:10.1038/nature05977
81. Aguilera A, García-Muse T. Causes of genome instability. *Annu Rev Genet*. 2013;47:1-32. doi:10.1146/annurev-genet-111212-133232
82. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet*. 2002;30(2):194-200. doi:10.1038/ng822
83. Fujimori S, Washio T, Higo K, et al. A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *FEBS Lett*. 2003;554(1-2):17-22. doi:10.1016/s0014-5793(03)01041-x
84. Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. *Genome Res*. 2014;24(11):1894-1904. doi:10.1101/gr.177774.114
85. Bailey KA, Reeve JN. DNA repeats and archaeal nucleosome positioning. *Res Microbiol*. 1999;150(9-10):701-709. doi:10.1016/s0923-2508(99)00122-9
86. Mitas M. Trinucleotide repeats associated with human disease. *Nucleic Acids Res*. 1997;25(12):2245-2254. doi:10.1093/nar/25.12.2245
87. Bacolla A, Gellibolian R, Shimizu M, et al. Flexible DNA: genetically unstable CTG.CAG and CGG.CCG from human hereditary neuromuscular disease genes. *J Biol Chem*. 1997;272(27):16783-16792. doi:10.1074/jbc.272.27.16783

88. Zhang J, Fakharzadeh A, Pan F, Roland C, Sagui C. Atypical structures of GAA/TTC trinucleotide repeats underlying Friedreich's ataxia: DNA triplexes and RNA/DNA hybrids. *Nucleic Acids Res.* 2020;48(17):9899-9917. doi:10.1093/nar/gkaa665
89. Ohshima K, Kang S, Larson JE, Wells RD. Cloning, characterization, and properties of seven triplet repeat DNA sequences. *J Biol Chem.* 1996;271(28):16773-16783. doi:10.1074/jbc.271.28.16773
90. Campuzano V, Montermini L, Moltò MD, et al. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science.* 1996;271(5254):1423-1427. doi:10.1126/science.271.5254.1423
91. Leach TJ, Chotkowski HL, Wotring MG, Dilwith RL, Glaser RL. Replication of heterochromatin and structure of polytene chromosomes. *Mol Cell Biol.* 2000;20(17):6308-6316. doi:10.1128/MCB.20.17.6308-6316.2000
92. Beliveau BJ, Joyce EF, Apostolopoulos N, et al. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc Natl Acad Sci U S A.* 2012;109(52):21301-21306. doi:10.1073/pnas.1213818110
93. Jiang J. Fluorescence in situ hybridization in plants: recent developments and future applications. *Chromosome Res.* 2019;27(3):153-165. doi:10.1007/s10577-019-09607-z

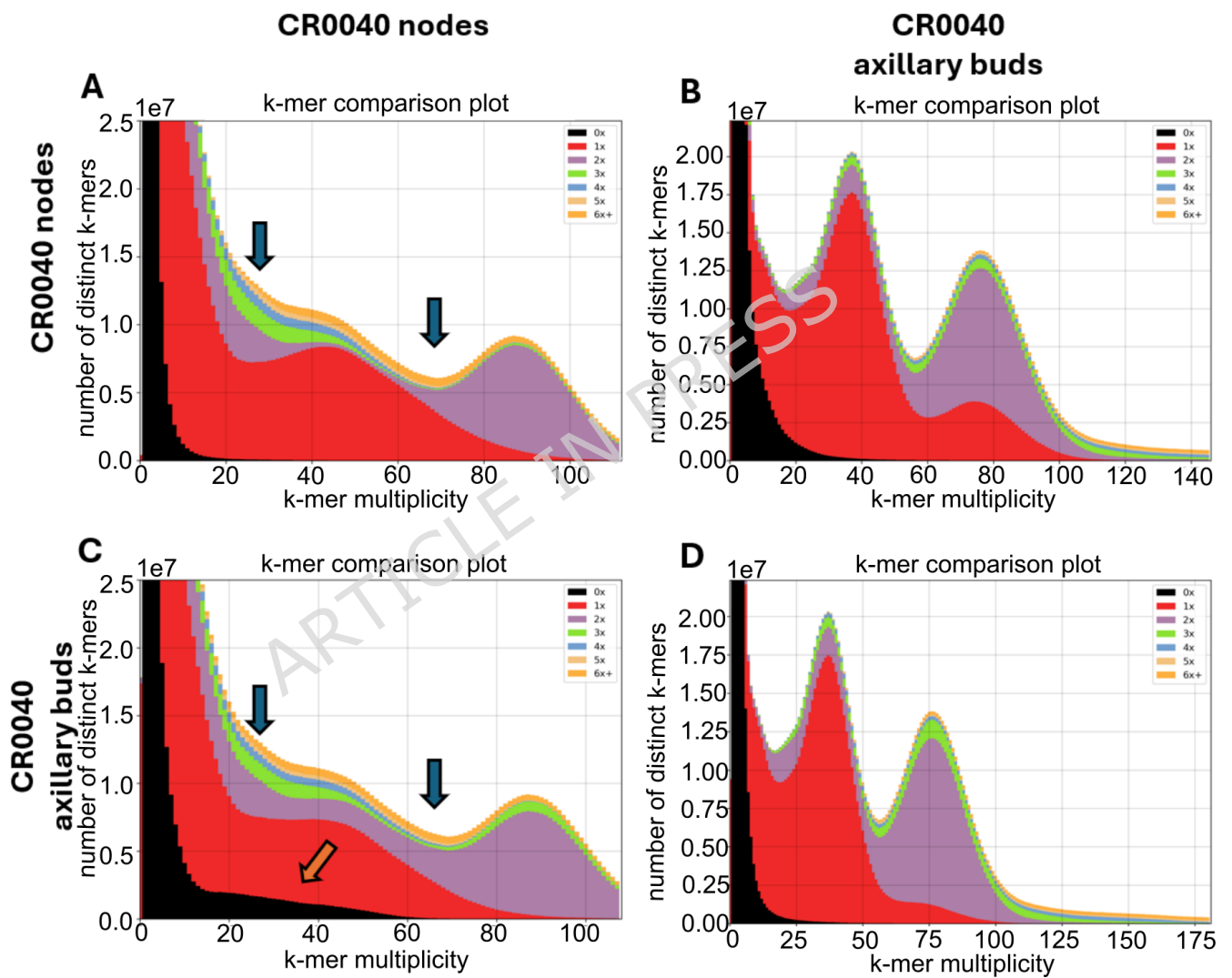
A



B

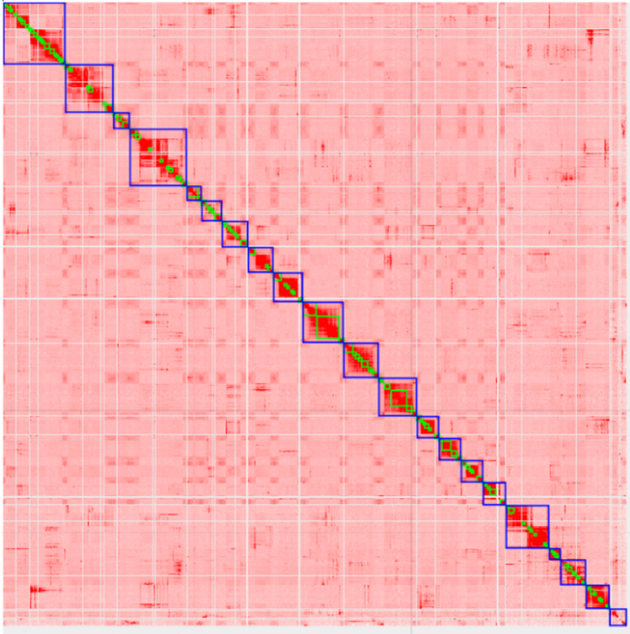


Reads



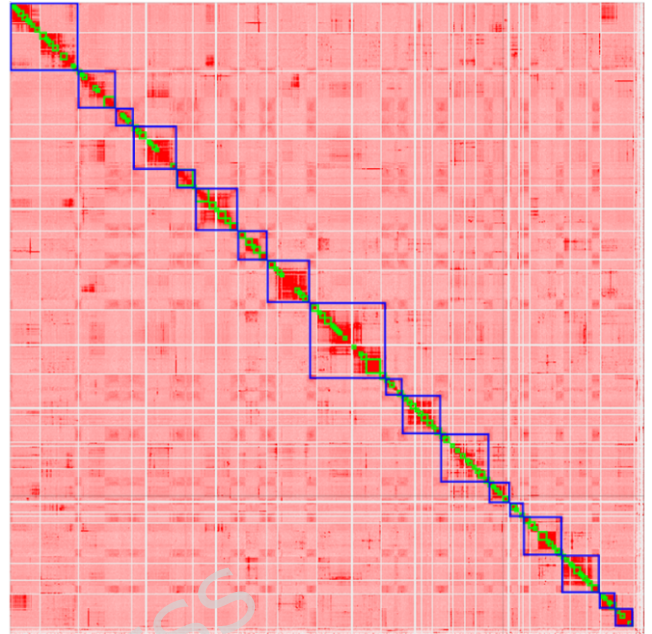
A

Haplotype A

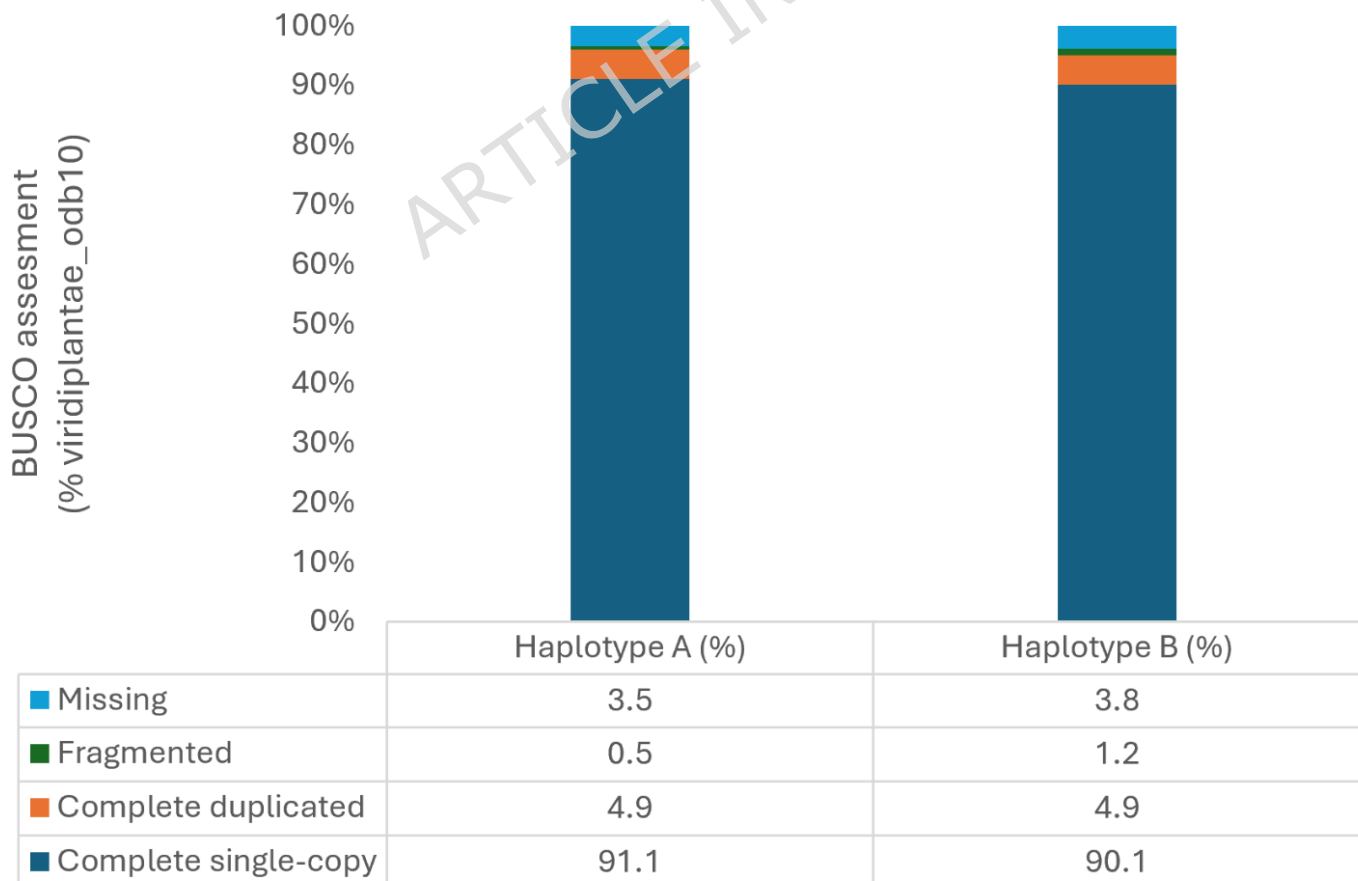


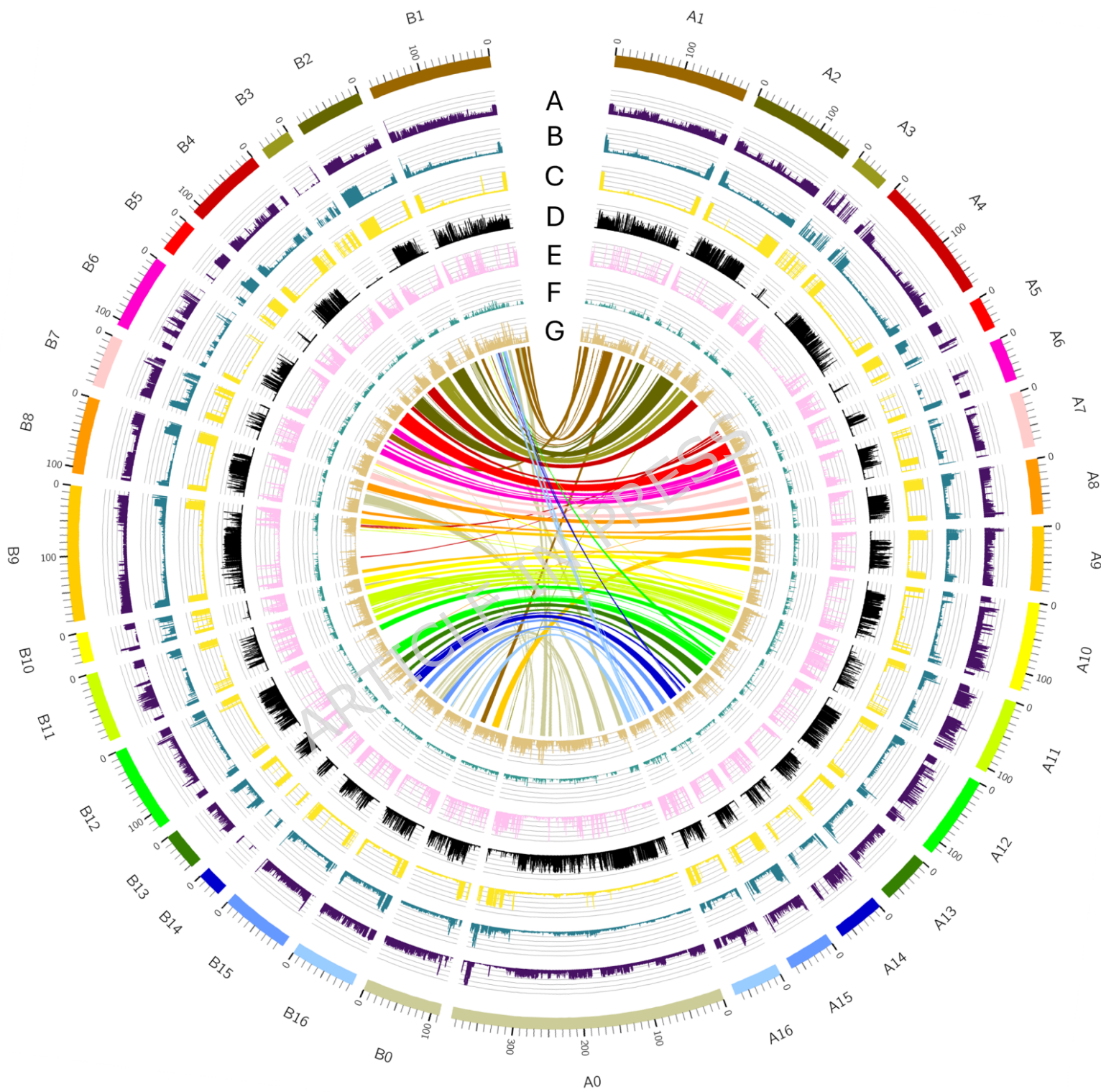
B

Haplotype B

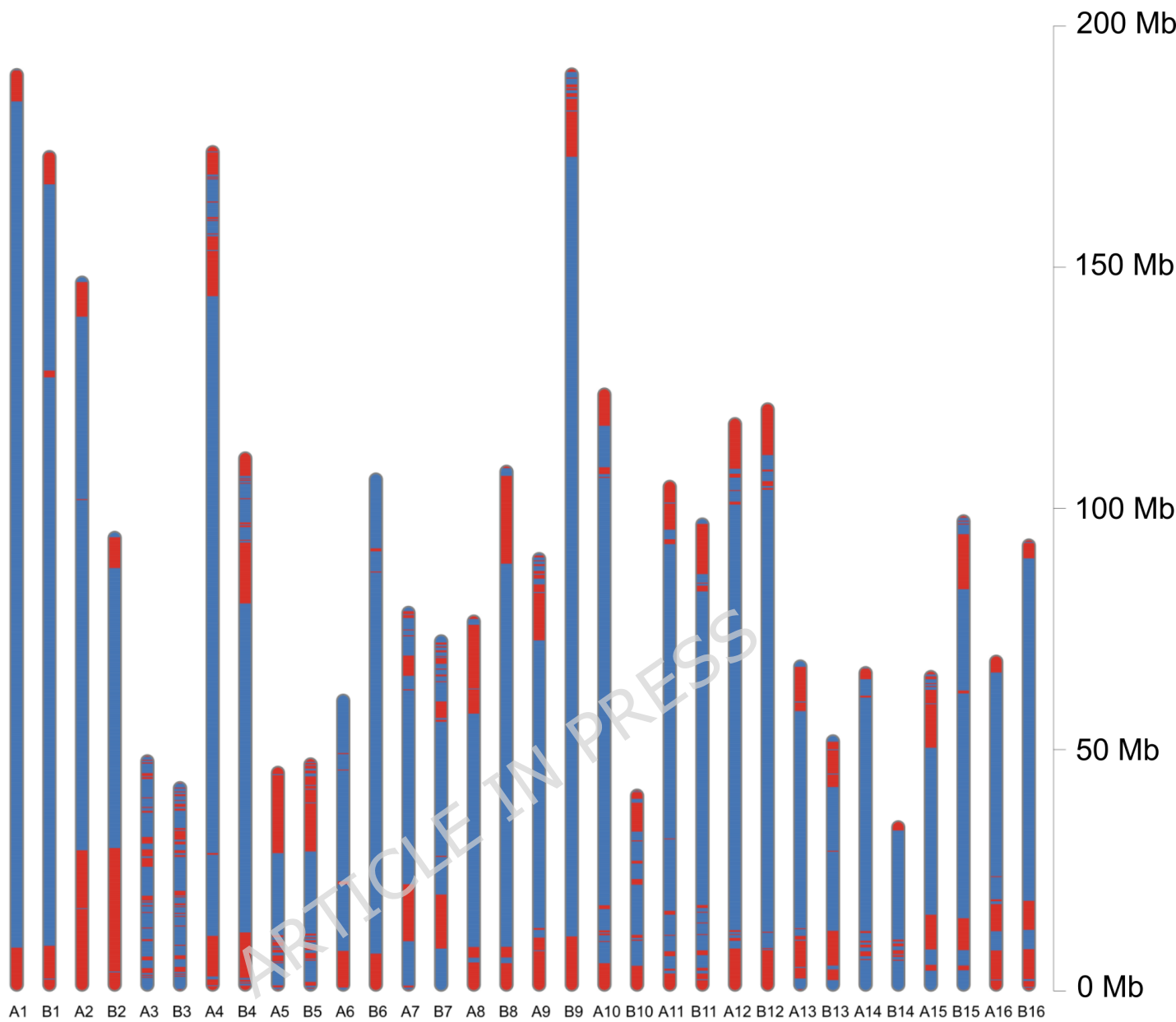


C

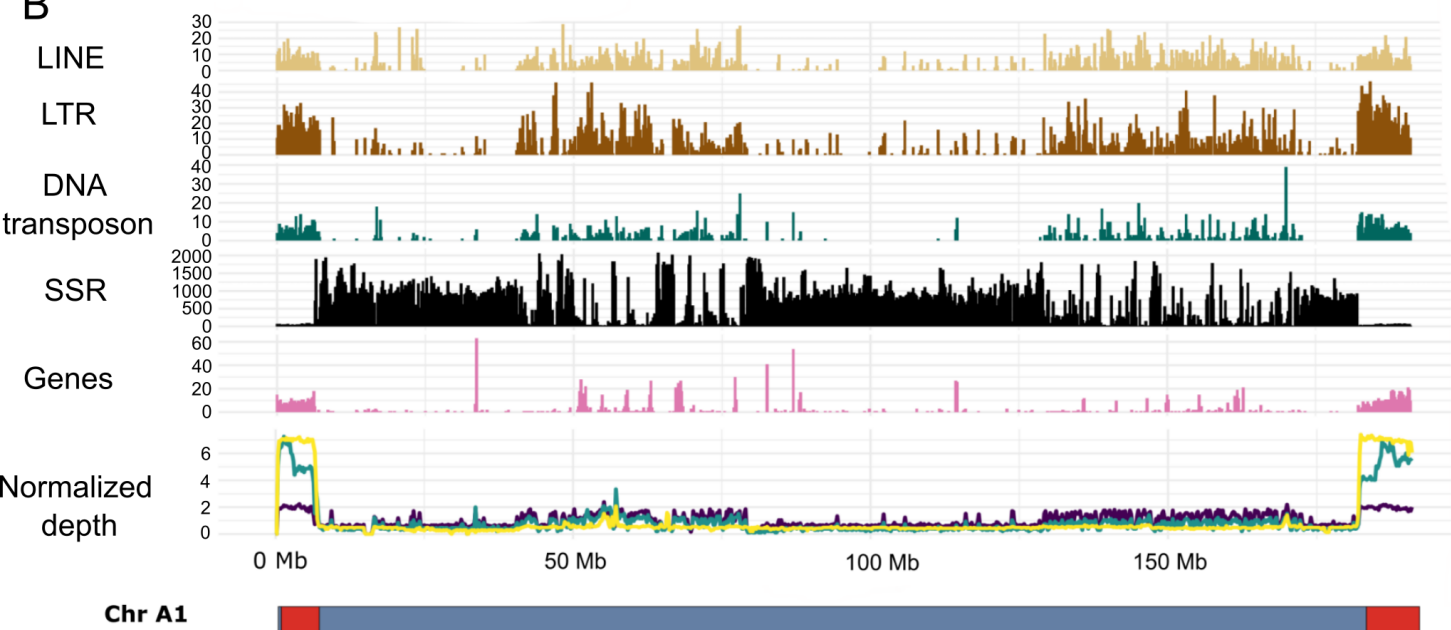




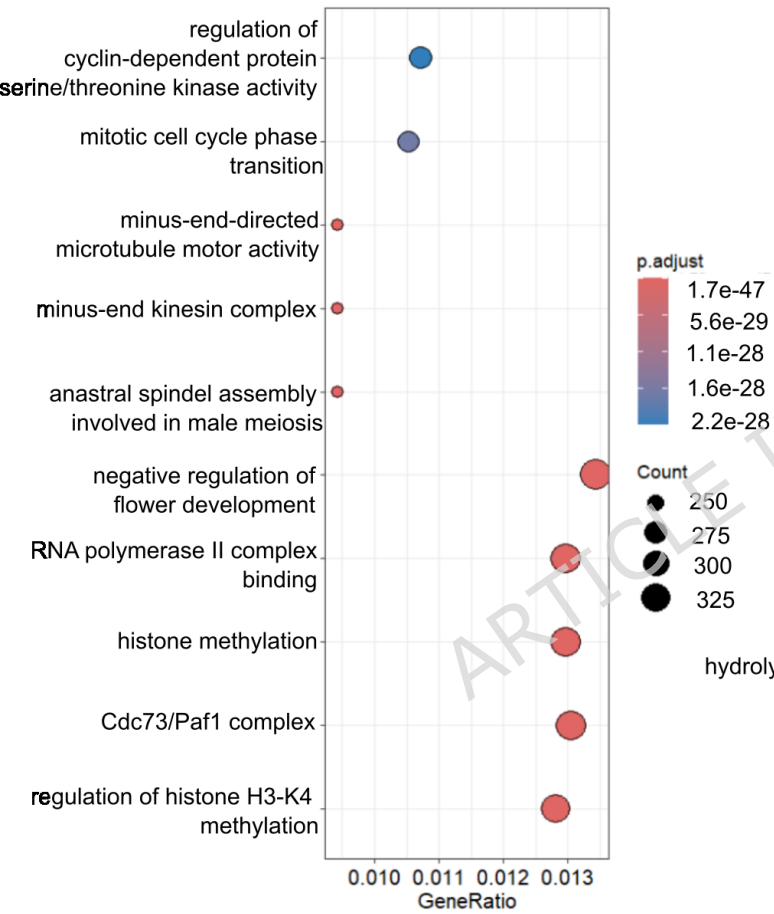
A



B



A



B

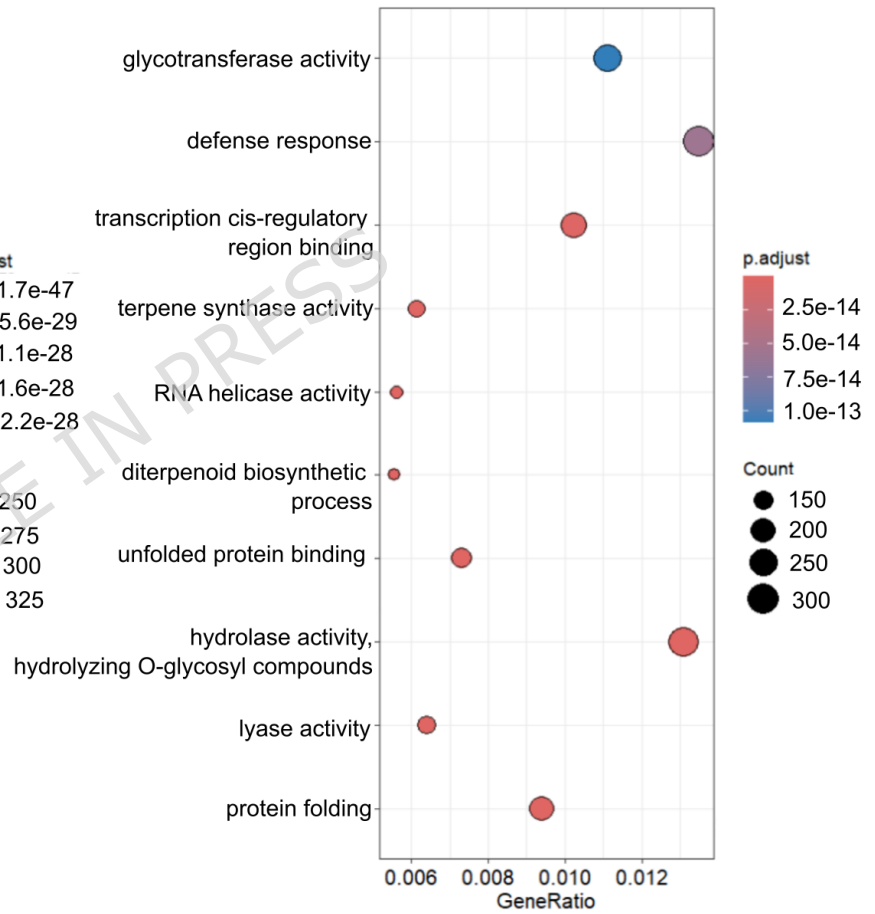


Table 1. Final assembly and annotation statistics of the diploid *V. planifolia* CR0040 genome

Total assembly size (Gb)	3.53
Total sequences number	476
N50 (Mb)	105.76
Maximum sequence length (Mb)	191.23
Sequence in chromosomes (% sequence)	86.32
GC content (%)	31.87
Number of protein-coding genes	63,872
Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness	96.7
Total interspersed repeats content (% sequence)	32.98
Total SSR content (% sequence)	6.87

Gb, Giga bases; Mb, Mega bases