Plant Communications, Volume *3*

Supplemental information

A chromosome-level, haplotype-phased *Vanilla planifolia* genome highlights the challenge of partial endoreplication for accurate whole-genome assembly

Quentin Piet, Gaetan Droc, William Marande, Gautier Sarah, Stéphanie Bocs, Christophe Klopp, Mickael Bourge, Sonja Siljak-Yakovlev, Olivier Bouchez, Céline Lopez-Roques, Sandra Lepers-Andrzejewski, Laurent Bourgois, Joseph Zucca, Michel Dron, Pascale Besse, Michel Grisoni, Cyril Jourda, and Carine Charron

SUPPLEMENTAL INFORMATION

Supplemental Note 1

Global strategy for the sequencing and assembly of *V. planifolia* CR0040 genome

The assembly of the CR0040 genome and the study of the partial endoreplication (PE) phenomenon required interdisciplinary work and a multitude of sequencing technologies, which has proven to be crucial for the detection to detect PE along the assembled chromosomes (Supplemental Figure 1).

Plant material

A traditional *Vanilla planifolia* cultivar from La Reunion was used for this study. It is conserved by CIRAD under the accession number CR-VA-00040 ("CR0040") in the Biological Resources Center Vatel (Saint Pierre, La Réunion) where it is maintained by cuttings under shade-house and by *in vitro* culture established from axillary buds. The vitroplants were propagated by micro-cuttings in basal Murashige and Skoog media (Duchefa Biochemie, The Netherlands) without any phytohormone. They were grown at 24-26°C, 12H light/day. Morphological data and genetic analyses confirm its taxonomic position within the *V. planifolia* species (Bory et al., 2008; Bouétard et al., 2010; Favre et al., 2022), and its membership of the group of vanilla cultivars usually grown in the Indian Ocean region. The profile of aromatic precursors identified in its fruits is typical of vanillas used for Bourbon vanilla production (Perez-Silva et al., 2006).

Flow cytometry

The total nuclear DNA amount was assessed by flow cytometry (FCM) as described in Bourge et al. (2018). Internal reference used for cytometry was *Petunia hybrida* Vilm. "PxPc6" (2C=2.85 pg). Leaves of the internal standard and nodal tissues from three to five month-old *in vitro V. planifolia* CR0040 were chopped using a razor blade in a plastic Petri dish with 1 ml of Gif nuclei-isolation buffer (45 mM $MgCl_2$, 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 and RNAse (2.5 U/mL). The suspension was filtered through 50-µm nylon mesh. The nuclei were stained with 70-100 µg/ml propidium iodide, a specific DNA fluorochrome intercalating dye, 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, Excitation 561 nm; emission through a 610/20 nm band-pass filter). Nuclei were identified by a gate on Side-Scatter and propidium iodide (PI)-Area and the cytogram of PI-Area versus PI-Height signals served to

select singlets, to eliminate doublets and to detect any degradation. Different cell ploidy levels were identified by their fluorescence intensity in PI, compared to the standard. The DNA histogram of nodal tissues from *Vanilla planifolia* CR0040 and nuclear classes in FCM samples are illustrated in the Supplemental Figure 2 and Supplemental Table 1, respectively. Note that there was a large variability between each endoploidy proportions for the 13 individuals.

The component F represents the Fixed part of the haploid genome which does not endoreplicate. The component P represents the part potentially participating in endoreplication. P and F are proportions (and not amounts) of the genome (%), whereas p and f are amounts (typically pg). Note that, as proportions, F and P have the same value whether referring to the haploid or to the diploid genome. As described in Brown et al. 2017, the replicate fraction P is assessed from the relative fluorescence intensity (I, arbitrary units) of peak#2 (4E, the first endocycle population) to peak#1 (2C nuclei): $P = [(Ipeak2 / Ipeak1) – 1] \times 100$

In quantitative terms, the haploid nucleus is (1f+1p) pg, and diploid nucleus is (2f+2p) pg. So 4E nuclei have four copies of the part of the genome which replicates, and two copies of the rest of the genome which does not replicate, in total 2f+4p (pg). The 8E nuclei have 2f+8p (pg), etc.

Cytogenetics

Root tips were obtained from *Vanilla planifolia* CR0040 plantlets cultivated in vitro in the VATEL biological resource center based in La Réunion. They were pretreated with 0.05% colchicine aqueous solution at room temperature during 1h and then fixed in absolute ethanol and glacial acetic acid (3:1) for at least two days at 4 °C. Root tips were hydrolyzed in 1 M HCl for 12 min at 60 °C, washed in distilled water at room temperature, and stained in 1% orcein in 45% acetic acid for about 30 min. Root tip meristems were squashed in a droplet of acetic carmine and observed under Zeiss Axiophot microscopes. Chromosome number was determined on five to seven well-spread metaphase plates from each of seven of those in vitro cultivated plantlets. The best metaphase plates were photographed using a CCD camera (RETIGA 2000R; Princeton Instruments, Evry, France).

For fluorochrome banding (Chromomycin A_3, Hoechst 33258 and DAPI- 4′,6-diamidino-2-phenylindole), the meristems were hydrolyzed for about 1 h at 37° C in an enzymatic mixture composed of 4% cellulase RS (Onozuka Yakult Honsha Co.), 1% pectolyase Y23 (Seishin Pharmaceutical Co, Tokyo, Japan), and 4% hemicellulase (Sigma Chemical Co) in 0.01M citrate buffer at pH 4.6. Thus, digested meristems were squashed into a drop of freshly prepared 50% acetic acid and the preparations were observed using an epifluorescence Zeiss Axiophot microscope with different combinations of excitation and emission filter sets (01, 07, 15, and triple 25). The best slides were frozen at -80ºC at least during 12 h. The coverslips were removed and the slides

were rinsed with absolute ethanol and air-dried. To detect GC- and AT-rich DNA regions, meristematic tissue was stained respectively with chromomycin A_3 (CMA) (Serva, France) according to Schweizer (1976), and with Hoechst 33258 (Ho; Sigma) according to Martin and Hesemann (1988) with minor modifications of Siljak-Yakovlev et al. (2002). The slides were mounted in citifluor AF2 (Agar Scientific Oxford Instruments, Stanstead, UK). The DAPI was used to observe an unspecific heterochromatin. The acquisition and treatment of images were performed using a highly sensitive CCD camera (RETIGA 2000R, Princeton Instruments, Evry, France) and an image analyzer (MetaVue, Evry, France).

High molecular weight DNA extraction and sequencing

High molecular weight (HMW) DNA was extracted from 1g of nodal tissues from *in vitro* cultured *Vanilla planifolia* CR0040 in order to minimize the endoreplicated phenomenon using QIAGEN genomic tip kit. DNA libraries and sequencing were performed at GeT-PlaGe core facility, INRAe Toulouse.

The Sequel2 HiFi library preparation and sequencing were performed according to the manufacturer's instructions "Procedure and Checklist Preparing HiFi SMRTbell Libraries using SMRTbell Express Template Prep Kit 2.0". 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 15µg of DNA was purified then sheared at 15kb using the Megaruptor3 system (Diagenode). A single strand overhang removal and a DNA and END damage repair step were performed on 10µg of sample using SMRTbell Express Template prep kit 2.0. Then blunt hairpin adapters were ligated to the library. The library was treated with an exonuclease cocktail to digest unligated DNA fragments. A size selection step using a 12kb cutoff was performed on the BluePippin Size Selection system (Sage Science) with "0.75% DF Marker S1 3-10 kb Improved Recovery" protocol. Using Sequel® II Binding kit 2.0 and sequencing kit 2.0, the primer V2 annealed and polymerase 2.0 bounded library was sequenced by diffusion loading onto 3 SMRTcells on Sequel2 instrument at 55pM with a 2 hours pre-extension and a 30 hours movie. HiFi reads are produced by calling consensus from subreads generated by multiple passes of the enzyme around a circularized template from a single zero mode waveguide. HiFi reads were generated with the "Circular Consensus Sequencing (CCS)" pipeline in SMRT Link v8.0 (ccs, version 4.0.0) with default parameters (minimum predicted accuracy of 0.99, minimum number of passes = 3)".

The Chromium libraries were prepared according to 10X Genomics' protocols using the Genome Reagent Kits v2. The sample quantity and quality controls were validated on Qubit, Nanodrop and

Femto. Optimal performance has been characterized on input gDNA with a mean length greater than 50 kb. The libraries were prepared from 3 µg of HMW gDNA (cut off at 50Kb using the BluePippin system). Briefly, in the microfluidic Genome Chip, a library of Genome Gel Beads is combined with HMW template gDNA in Master Mix and partitioning oil to create Gel Bead-In-EMulsions (GEMs) in the Chromium. Each Gel Bead is functionalized with millions of copies of a 10x™ Barcoded primer. Upon dissolution of the Genome Gel Bead in the GEM, primers containing (i) an Illumina R1 sequence (Read 1 sequencing primer), (ii) a 16 bp 10x Barcode, and (iii) a 6 bp random primer sequence are released. Read 1 sequence and the 10x™ Barcode are added to the molecules during the GEM incubation. P5 and P7 primers, read 2, and sample index are added during library construction. 10 cycles of PCR were applied to amplify libraries. Library quality was assessed using a Fragment Analyser and libraries were quantified by QPCR using the Kapa Library Quantification Kit. The libraries have been performed on an Illumina HiSeq3000 using a paired-end read length of 2x150 pb with the Illumina HiSeq3000 sequencing kits.

Oxford Nanopore Technologies (ONT) library preparation and sequencing were performed according to the manufacturer's instructions "1D gDNA selecting for long reads (SQK-LSK109)". 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 Fragment analyzer (AATI) High Sensitivity DNA Fragment Analysis Kit. Purification steps were performed using AMPure XP beads (Beckman Coulter). 8 libraries were prepared, for each library, 5µg of DNA was purified then sheared at 20kb to 40kb using the megaruptor 1 system (diagenode). A one step DNA damage repair + END-repair + dA tail of double stranded DNA fragments was performed on 2µg of sample. Then adapters were ligated to the library. Libraries were loaded onto 3 R9.4.1 revD GridION flowcells and sequenced on GridION instrument at 30fmol within 48H and onto 5 R9.4.1 revD PromethION flowcells and sequenced on PromethION instrument at 25 fmol within 72H.

DNA-seq data production per technology and instrument are indicated in the Supplemental Table 2.

Ultra-HMW DNA extraction and Optical mapping

Ultra-HMW DNA were purified from one gram of nodal tissues at the base of the young leaves according to the Bionano Prep Plant tissue DNA Isolation Liquid Nitrogen Grinding Protocol (30177 - Bionano Genomics) with the following specifications and modifications. Briefly, the tissues were broken in liquid nitrogen and then ground with rotor stator in the homogenization buffer. Nuclei were washed and then embedded in agarose plugs. After overnight proteinase K digestion in the presence of Lysis Buffer (Bionano Genomics) and one-hour treatment with RNAse A (Qiagen), plugs were washed four times in 1x Wash Buffer (Bionano Genomics) and five times

in 1x TE Buffer (ThermoFisher Scientific). Then, plugs were melted two minutes at 70°C and solubilized with 2 µL of 0.5 U/µL AGARase enzyme (ThermoFisher Scientific) for 45 minutes at 43°C. A dialysis step was performed in 1x TE Buffer (ThermoFisher Scientific) for 45 minutes to purify DNA from any residues. The DNA samples were quantified by using the Qubit dsDNA BR Assay (Invitrogen). The presence of mega base size DNA was visualized by pulsed field gel electrophoresis. Labeling and staining of the uHMW DNA were performed according to the Bionano Prep Direct Label and Stain (DLS) protocol (30206 - Bionano Genomics). Briefly, labeling was performed by incubating 750 ng genomic DNA with 1× DLE-1 Enzyme (Bionano Genomics) for 2 hours in the presence of 1× DL-Green (Bionano Genomics) and 1× DLE-1 Buffer (Bionano Genomics). Following proteinase K digestion and DL-Green cleanup, the DNA backbone was stained by mixing the labeled DNA with DNA Stain solution (Bionano Genomics) in presence of 1× Flow Buffer (Bionano Genomics) and 1× DTT (Bionano Genomics), and incubating overnight at room temperature. The DLS DNA concentration was measured with the Qubit dsDNA HS Assay (Invitrogen). Labelled and stained DNA was loaded on the Saphyr chip. Loading of the chip and running of the Bionano Genomics Saphyr System were all performed according to the Saphyr System User Guide (30247 - Bionano Genomics). Data processing was performed using the Bionano Genomics Access software (https://bionanogenomics.com/support-page/bionano-access-software/). A total of 590 Gb data with molecules larger than 150kb was produced and then *de novo* assembled according to the Access software. It produced 950 genome maps with a N50 of 4.6 Mbp for a total genome map length of 2,115 Gbp.

Supplemental Note 2

Genome assembly and pseudomolecules construction

Illumina 10X, ONT and PacBio HiFi reads were assembled using respectively Supernova (v.2.1.1, Weisenfeld et al., 2017), Flye (v.2.4.2-release, https://github.com/fenderglass/Flye, Kolmogorov et al., 2019) followed by three iterative polishing with Illumina data using Pilon (v.1.22, https://github.com/broadinstitute/pilon, Walker et al., 2014) and Hifiasm v.0.13 (Cheng et al., 2021), with default parameters. A meta-assembly of corrected ONT reads from Canu and PacBio HiFi reads was tentatively generated using Hifiasm. The assembly metrics were computed with QUAST 5.1.0 (Gurevich et al., 2013) and summarized in Supplemental Table 3.

Hybrid scaffolding was performed between the best assembly (PacBio HiFi assembly) and optical genome maps with hybridScaffold pipeline and default parameters (https://bionanogenomics.com/wp-content/uploads/2018/04/30073-Bionano-Solve-Theory-of-

Operation-Hybrid-Scaffold.pdf). We obtained 874 hybrid scaffolds with the maximum size of 32 Mb, reaching 1.9 Gb and N50 of 6.2Mb. The not anchored contigs represent 1,529 Mb and 23,037 contigs from the PacBio assembly. This hybrid scaffold file is composed of the 2 allelic versions of the genome. In order to separate the alleles, we used an in-house script. Briefly, the fasta file is transformed into an optical maps file (.cmap). This cmap file is aligned against itself using Bionano tools to create a xmap file that has the information of all the scaffold id that align against each scaffold id. With this information, a parsing algorithm, coded in java, produces 2 files with 1 allele in each file. Contigs not belonging to scaffolds were split among haplotype using purge dups (https://github.com/dfguan/purge_dups). We then only kept contigs that were unassigned to scaffolds and added those from the "hap" file to haplotype A and those from the "purged" file to haplotype B. To organize the scaffolds into pseudomolecules, haplotype A fasta file was aligned against Daphna haplotype A chromosomes using the same Bionano tools. In that way, 36 scaffolds were organized in 14 chromosomes for haplotype A and 49 scaffolds for haplotype B.

Evaluation of assembly quality

The metrics for the best assembly were: assembly for haplotype A has 1.5 Gb, N50 = 3 Mb and 3,874 scaffolds or contigs and assembly for haplotype B has 1.9 Gb, N50 =0.4Mb and 17,655 scaffolds or contigs (Supplemental Table 4). We then looked for universal single copy orthologs genes with BUSCO 5.0.0. This analysis was done for the whole assembly and for both haplotypes apart on three levels of taxonomy using the following databases: viridiplantae_odb10, embryophyta_odb10 and liliopsida_odb10 (Supplemental Table 5).

Supplemental Note 3

RNA extraction and sequencing

Total RNA was extracted from nine organs (fruit, leaf, flower, stem, soil root, aerial root, axillary stem bud, flower bud and ovary) using the RNeasy Plant Mini kit, according to the manufacturer's protocol (Qiagen, Hilden, Germany). RNA purity, quantification, and integrity were evaluated using a NanoDrop 2000 spectrophotometer (Thermo Scientific, USA) and Agilent 4200 bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Only samples with an RNA Integrity Number (RIN) ≥ 8 were subjected to subsequent analysis and indicated in Supplemental Table 6.

RNA-seq libraries were prepared according to Illumina's protocols using the Illumina TruSeq Stranded mRNA sample prep kit to analyze mRNA at the GeT-PlaGe core facility, INRAe Toulouse. Briefly, mRNAs were selected using poly-T beads and then fragmented to generate

double stranded cDNA. Adaptors were ligated to cDNA and 11 cycles of PCR were applied to amplify libraries. Library quality was assessed using a Fragment Analyzer and then quantified by qPCR using the Kapa Library Quantification Kit. RNA-seq experiments were performed on an Illumina NovaSeq using a paired-end read length of 2x150 pb with the Illumina NovaSeq sequencing kits.

RNA-seq assembly and gene expression

RNA-seq reads of ten samples (the nine sequenced organs and a mix of them) were mapped onto the *Vanilla planifolia* CR0040 contig assembly using hisat2 (v.2.1.0) with default parameters (Kim et al., 2019), which resulted in one alignment file per sample. Each of these files were then coordinate-sorted and converted to BAM format in order for them to be compatible with the StringTie (v.2.0.3, Pertea et al., 2015), the RNA-seq alignment assembler used with default parameters. The resulting output files (gtf format) were used for a last StringTie run with the "--merge" option, thus producing a non-redundant set of transcripts as a single output file, corresponding to the final set of assembled transcripts. Main statistics of the RNA-seq experiment are detailed in Supplemental Table 7.

To estimate gene expression in the nine different tissues and to identify putative novel isoforms, we followed the StringTie's recommended protocol by realigning the corresponding RNA-Seq reads to the reference with hisat2 (v.2.2.1) using perl scripts provided by hisat2 (*extract_exon.py* and *extract_splice_sites.py*) to extract exon and splice site information from the annotation of the CR0040 final assembly. Alignment files were sorted by coordinates and converted to BAM format with the samtools utility (v.1.10). Then, RNA-seq reads were assembled using StringTie (v.2.0.3) with a GTF file containing transcripts coordinates from the annotation of the CR0040 assembly (-G option to specify the file), that is used to guide the assembly. The output files were then merged to remove redundancy between samples (StringTie v.2.0.3 with --merge and -G options) and to generate a single gtf file containing their coordinates. Finally, a final StringTie run was performed in order to re-estimate transcript abundance and generate read coverage tables (-B and -e options were used) for each sample, in regards to the merged set (-G option). These tables were then used by the prepDe.py script to compute genes and transcripts count matrices (see RNA-seq tracks in https://vanilla-genome-hub.cirad.fr/content/v-planifolia-cr0040).

A BUSCO analysis was performed using the Viridiplantae database to estimate the genic completeness for each transcriptome and for merged transcriptomes (Supplemental Table 8).

Structural and functional genome annotation

Automatic gene prediction was performed using EuGene Eukaryotic Pipeline (EGNEP v.1.5 with EuGene v.4.2a), an integrative gene finder software that is able to combine several sources of information in order to predict genes (Sallet et al., 2019). This pipeline can infer a structural gene annotation considering homology with known sequences, structural information in the form of GFF3 files and statistical information.

EuGene's internal prediction model can be built and trained using proteomic and transcriptomic data. Thus, a set of annotated *V. planifolia Daphna* proteins (NCBI Genomes: haplotype A, BioProject Accession: PRJNA633886, GenBank assembly accession: GCA_016413895.1 and haplotype B, BioProject Accession: PRJNA668740, GenBank assembly accession: GCA_016413885.1) as well as a combination of the non-redundant set of transcripts of *V. planifolia* CR0040 assembled with Stringtie and the *V. planifolia* Daphna annotated transcripts (*.rna_from_genomic.fna.gz) were submitted to the pipeline for this task. The rest of the proteomic data used as evidence are a combination of *Phalaenopsis equestris* proteins retrieved from RefSeq (NCBI RefSeq Genome: BioProject Accession: PRJNA382149, Annotation Release 100 accession: GCF_001263595.1) and the Uniprot/Swissprot *Liliopsida* database (2020_06 version). One Trinity assembly of a *Vanilla planifolia Jacks* (Solano-De la Cruz et al., 2019) retrieved from the NCBI database (NCBI Transcriptome: BioProject Accession: PRJNA554104, GEO Series: GSE134155) was also used as transcriptomic evidence. The pipeline comes with repeat detection software such as Red (Girgis et al., 2015) and LTRharvest (Ellinghaus et al., 2008), and similarity searches against RepBase (REPET, version 20.05) are also performed to clean the proteomic datasets at the start of the annotation process.

As EuGene also uses statistical models for splice sites detection, we built an orchids specific model from the genomic and transcriptomic data of *V. planifolia Daphna*, *Dendrobium catenatum* (NCBI RefSeq Genome: BioProject Accession: PRJNA453230, Annotation Release 101 acc: GCF_001605985.2), *Phalaenopsis equestris* (NCBI RefSeq), and our *V. planifolia* CR0040 clone using the egn_build_wam.pl companion script.

In order to assign functions to the predicted protein coding genes. InterProScan domain searches as well as similarity searches (BlastP) against Uniprot/Swissprot and Uniprot/TrEMBL databases were performed. Putative gene models were sorted into six confidence classes (from ISS_1 to ISS_6) based on their functional annotation. In order to filter protein coding genes, ISS_6 genes (96,224 sequences), ISS_5 genes with protein sequence shorter than 150 amino acids (aa) and with untranslated region (UTR) shorter than 35 base pairs (bp), and ISS_4 genes with protein sequences shorter than 80 amino acids and without UTR sequence were removed (98,718 sequences). Among remaining putative genes, ISS_5 genes, ISS_4 genes and ISS_3 fragments genes coding for protein sequences with matches in RepBase were also removed (3,527 sequences). In addition, putative

gene models with protein sequences predicted to have MULE transposase interpro domains (IPR018289) without match on FAR protein (FAR-RED impaired response 1, transcription activator that derived from ancient transposases) domain (IPR031052), or reverse transcriptase domain (IPR000477) without NMAT (nuclear intron maturase 1 mitochondrial, IPR024937) domain and TERT family domains (telomerase reverse transcriptase, IPR003545) were removed (589 sequences). Putative genes encoding for gagpol polyprotein and not associated to Benchmarking Universal Single-Copy Orthologs (BUSCO) and putative genes located on organellar contigs were removed (23,237 sequences). After filtering steps, 59,128 high-confidence protein coding genes were kept. The number of these genes per chromosome and per haplotype as well as main statistics of these genes are indicated in the Supplemental Tables 9 and 12.

Functions were assigned through InterProScan domain searches as well as similarity searches against Uniprot/Swissprot and Uniprot/TrEMBL databases (BlastP). Gene Ontology (GO) terms were assigned through InterProScan (Jones et al., 2014) results while Enzyme Classification (EC) numbers were predicted combining both tools PRIAM (Claudel-Renard et al., 2003) and BlastKOALA (Kanehisa et al., 2016).

Enzymes and metabolic pathways were predicted from the protein-coding genes using Pathway Tools (Karp et al., 2002). A file in the PathoLogic format was created, which included gene functional descriptions, GO terms and enzyme commission numbers. This file will be used to set up Pathways tools (see Supplemental Note 6). In order to assess the completeness of the resulting set of protein coding gene models, a BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis (v4.0.5) was carried out using three different databases, namely Viridiplantae, Embryophyta and Liliopsida (Supplemental Table 10).

From repeats detected by EGNEP, 1,472 interspersed repeats were selected based on the consensi found by RepeatModeler (v2.0.1) (Flynn et al., 2020), enriched with consensi of RepeatScout (v1.0.5) (Price et al, 2005) and transposable element genes (TEG) predicted from EGNEP (REGN TEG), according to the following procedure of classification, filtering and clustering.

Consensi families were generated by RepeatScout (RS) and RepeatModeler (RM) from the CR0040 genome assembly (21,529 sequences): 1,247 TEG, 1,958 RS consensi and 1,236 RM consensi were then classified with REPET (v3.0) and PASTEC (v2.0) (Hoede et al., 2014) according to the Wicker's TE classification (Wicker et al., 2007) and using RepBase (REPET, version 23.12), cleaned by removing SSR and rDNA sequences, requalified with REPET postProcessClassif.py, then potential host genes were removed and finally clustered with CD-HIT (v4.8.1) (Fu et al., 2012). A first clustering (cd-hit -c 1 -sc 1) was made between the three filtered sets (1,072 RM, 1,748 RS and 284 TEG) to remove identical sequences. 46 repetitive sequences were removed by manual curation of the 42 clusters with more than 2 sequences (out of the 3,057 CD-HIT clusters

including 3,015 singletons) leading to 3,058 sequences. A second clustering (cd-hit -c 0.85 -sc 1 -aS 0.75) was made on the 3,058 sequences to help us to select 767 transposable elements TE, as follow: 338 RM TEs, 103 RS TE CD-HIT cluster representatives, 35 REGN TEG cluster representatives, 7 RS SINEs (RSX) that were not cluster representatives, 259 RS TE singletons not classified as unknown retrotransposon (RXX) and 25 REGN TEG singletons classified as DNA transposon (DTX). A third clustering (cd-hit -c 0.85 -sc 1 -aS 0.75) was made on the 693 RM unknown interspersed repeats (neither classified by RM nor by PASTEC) leading to a reduction of the set to 639 sequences. This sequence set was concatenated with the 66 RS TE singletons of the previous clustering and classified as unknown retrotransposon (RXX) and considered as doubtful TE, leading to 705 unknown interspersed repeats. Then, the genome assembly was annotated with RepeatMasker (v4.1.1, Tarailo-Graovac et al, 2009) using the two CR0040 interspersed repeat banks separately (sequences of 767 transposable elements and 705 unknown interspersed repeats available at https://vanilla-genome-hub.cirad.fr/filebrowser/download/188). Bedtools intersect (v2.29.2, Quinlan et al., 2010) allowed the TE annotation to be given priority over the unknown interspersed repeat annotation. Summary report of the repeat annotation is detailed in the Supplemental Table 11.

Supplemental Note 4

Gene families clustering

Polypeptide sequences of five plant species and CR0040 (14 chromosomes and CR0040_A0 and CR0040_B0 random mosaic chromosomes of unanchored scaffolds) were retrieved in order to carry out comparative genomic analyses between them: *Arabidopsis thaliana (*TAIR10, Phytozome) (Lamesch et al., 2012), *Oryza sativa* (IRGSP-1.0, RAP-DB), *Phalaenopsis equestris* (NCBI RefSeq), *Vanilla planifolia* Daphna (NCBI Genomes) and *Phalaenopsis aphrodite* (Orchidstra 2.0, Chao et al., 2017). For genes with alternative splicing variants, only the polypeptide corresponding to the longest transcript was kept and the two haplotypes of each *V. planifolia* cultivar were separated for the analyses. A total of 237,645 proteins were clustered with Orthofinder2 (v.2.4.0) using default parameters (Emms et al., 2019). A total of 212,852 proteins (89.5%) were clustered to 27,815 orthogroups (https://vanilla-genome-hub.cirad.fr/filebrowser/download/187) with 10.5% of the whole set were part of species-specific orthogroups, 8,317 were conserved in all the species/haplotypes and 1,426 corresponded to single-copy orthogroups. With this Orthofinder analysis, made from proteomes derived from the 14 chromosomes and unanchored scaffolds, we began to glimpse into the vanilla pangenome. The vanilla core genome is composed of 14,403

families and 86,688 genes, common to both CR40 and Daphna genomes. The dispensable genome of CR0040 contains 3,637 families and 17,258 genes specific to CR0040. The dispensable genome of Daphna contains 3,804 Daphna specific families and 13,529 genes.

The inferred orthogroups were visualized with the UpsetR (Conway et al., 2017) R package to generate the UpSet plot (Supplemental Figure 5). Among the compared datasets, 596 orthogroups were only shared by orchids and 1,029 were only shared by the 4 haplotypes of *V. planifolia* (Daphna and CR0040). Number of conserved and specific gene groups are detailed in the Supplemental Table 14.

Expansion and reduction of gene families with CAFE

Forty-seven protein coding genes involved in cell cycle regulation in *A. thaliana* were focused on (Supplemental Table 17), and orthogroups containing them were analyzed in terms of family evolution using CAFE (v5, Mendes et al., 2020). Orthogroups showing expansions and/or contractions in copy numbers, among the six compared plant lineages, were identified using a P value threshold of 0.001. For this, orthogroups (made from proteomes derived from the 14 chromosomes only) and the species tree produced by orthofinder were used as input. Prior to this analysis, orthogroups that contained genes in only one species or with more than 100 gene copies for at least one species were removed. The species tree was made into an ultrametric one with the make_ultrametric_tree.py script bundled with the orthofinder package using an original divergence of 152 Mya for the analyzed species, as referenced in the TreeTime database (Kumar et al., 2017). These filtered orthogroups, and the ultrametric species tree were utilized to estimate the expansion/contraction of orthogroups using one lambda (λ) model, where λ is a random birth-death parameter.

Construction of syntenic blocks

The SynMap workflow (Haug-Baltzell et al., 2017) of the CoGe comparative genomics platform (Lyons and Freeling, 2008) was used to generate syntenic dotplots between two organism genomes (haplotypes A of CR0040 and Daphna, as well as *P. aphrodite*). SynMap relies on pairwise comparison of Coding Sequences (CDS) between the two genomes and allows to compute the synonymous mutation rate of syntenic CDS pairs.

SynMap requires the genome sequence and a gene annotation file of each compared organism. Prior to the analyses, annotated genes in each species were filtered to keep only the longest isoform of each gene. For each genome in the comparison, the chromosome fasta file and the gene annotation gff3 file were loaded using the iCommands to the CyVerse's cloud-based Data Store. Be careful that all CDS of the same mRNA (*e.g.* ID=VANPL_A_00001g000010.mRNA1) should have the

same identifier (e.g. ID=VANPL_A_00001c000010; Parent=VANPL_A_00001g000010.mRNA1) otherwise SynMap will overestimate the number of CDSs and this will cause problems during the tandem gene filtering step. First, tandem duplications are filtered out with the blast_to_raw program (SynMap Analysis Log file of CR0040_A, https://genomevolution.org/coge/data/diags/62209/62209/1j9ym.log, source code https://github.com/LyonsLab/coge). Syntenic pairs of CDSs are then identified by finding collinear series of putative homologous sequences using DAGChainer (Haas et al., 2004). A global alignment is performed on these syntenic CDS pairs using the Needleman-Wunsch algorithm implemented in nwalign (https://pypi.org/project/nwalign/) using the BLOSUM62 scoring matrix, after which a back translation of this alignment into a codon alignment is processed. Then, codeml-coge, a modified implementation of CodeML from the PAML package (Yang et al., 2007) is used to compute the rate of synonymous substitutions per synonymous site for each pair of CDSs (dS called Ks on the SynMap interface). In the particular case where a genome was aligned against itself, the identical CDS pairs were removed from the analysis and therefore do not appear in either the SynMap dotplot (no dots on the central diagonal) or the Ks distribution. However, the dotplot is symmetrical around the central diagonal so only half the dotplot can be considered.

Visual quality control of the assemblies and whole genome duplication exploration using pairwise orchid genome synteny

In order to check the quality of the 14 chromosomes of the CR0040 A haplotype (CR0040_A) *V. planifolia* genome, to compare the 14 chromosomes of haplotype A of both *Vanilla* cultivars and to study the pan-orchid alpha Whole Genome Duplication (α^o WGD), a series of analyses were performed with the CoGe Synmap pipeline, as described above. Synmap pairwise genome synteny analyses between CR0040_A and Daphna A haplotype (D_A) and between themselves were illustrated in the Supplemental Figure 7. The dotplot of the CDS pairs of the 14 chromosomes of CR0040_A (Supplemental Figure 7E) and the histogram of the CDS pair Ks values (Supplemental Figure 7F) gives a genomic representation suggesting that the 14 chromosomes of CR0040_A and the gene annotation is complete and of high quality. The comparison of the CR0040_A dotplot with the D_A dotplot (Supplemental Figure 7A) shows that CR0040 A looks cleaner, with more continuous paralog diagonals. Also, the dotplot of D_A CDS against itself displays an unexpected dotted central diagonal indicating physically close duplicated genes. The cyan color of the diagonal indicates that the similar CDS pairs belong to the peak with a very low number of synonymous substitutions per synonymous site ($\log10(Ks) \approx -1.5$, light blue bars in the Supplemental Figure 7B). This artificial peak would correspond to allelic forms due to the heterozygosity and not to WGD paralogs. This peak could come from a problem of dual haplotype conservation during contig

phasing (difficulties to separate the two haplotypes when too close). The D_A karyotype illustrating the α° WGD (Hasing et al., 2020) was aligned onto the 14 chromosomes of the D_A to identify by projection the doplot diagonals corresponding to these 13 blocks of ohnologs (α° WGD paralogs). The ancestral karyotype of CR0040_A was drawn, using an in-house Perl script, from the D_A gene pair blocks (Table S10 in Hasing et al., 2020), transferred to CR0040_A genes via reciprocal best hits computed in Supplemental Note 6. On the CR0040_A dotplot, the diagonals corresponding to the 13 ohnologous blocks were also found, but with slight differences. For example, the difference found for block number 13, between chromosome 10 and 11, could be due to the fact that CR0040_A chromosome 11 (47 Mb with Ns; Supplemental Table 12) is longer than the D_A chromosome 11 (38 Mb; Supplemental Table 13). Similarly, on the dotplot of D_A versus CR0040_A (Supplemental Figure 7D), the karyotypes were aligned along the 14 chromosomes, allowing to find the diagonals corresponding to the 13 ohnolog blocks present on D_A and CR0040_A. The histogram of Ks values of CDS pairs between D_A and CR0040_A (Supplemental Figure 7C) shows two distinct peaks. The first one corresponds to allelic gene pairs between D_A and CR0040_A and the second one to onholog pairs. These analyses validate the already known pan-orchid WGD. Additional short diagonals on the CR0040_A dotplot and the second peak on the CR0040_A Ks histogram suggest an older WGD, probably the tau (τ^m) of Monocots (Hasing et al., 2020). The blue-green diagonals would correspond to the α° WGD (*e.g.* log10(Ks) < 0.2) while the shorter green-red ones would correspond to the monocot τ^m WGD (*e.g.* log10(Ks) > 0.2). The paralogous diagonals between chromosomes 3-5-6-14 could be taken as an example on the CR0040_A dotplot (Supplemental Figure 7E). Knowing that the α° WGD is relatively old (90 Mya; Hasing et al., 2020) and the sequence of the 14 pseudomolecules of CR0040_A is of high fidelity, it is possible that τ^m traces are seen. The biology of the species is also a consideration. Vanilla is perennial with predominant vegetative multiplication that could decrease the mutation rate (Sandler et al., 2020). The classical formula T = Ks / (2r) was applied, on a rough example from Supplemental Figure 7F, with a mutation rate 'r' of 6.5E-9 synonymous substitutions per synonymous site per generation (Tang et al., 2010). For instance, with a log10(Ks) peak median of 0.09, then the α° WGD would be dated at ≈ 95 Mya and with a median at 0.35, the τ^m WGD would be dated at ≈ 172 Mya, in agreement with Jiao et al., 2014.

Missing pairs of chromosomes 15 and 16 in the *V. planifolia* genome were tentatively searched by comparison with the *P. aphrodite* genome. Due to problems with *P. aphrodite* CDS positions at the chromosome level, gene sequences of *P. aphrodite* contigs were aligned to the *P. aphrodite* chromosomes using Liftoff (Shumate and Salzberg, 2021) as described in Supplemental Note 6 and these results were used for comparative genomics with CR0040 haplotype A sequence. Synmap pairwise genome synteny analyses of the 14 pseudomolecules of CR0040_A *V. planifolia*, the

pseudomolecules and longest scaffolds of *P. aphrodite* and between themselves were illustrated in the Supplemental Figure 8.

A first correspondence between the chromosomes of the two species was obtained using the SynMap Syntenic path assembly (SPA) option (Supplemental Figure 8D, see correspondences at https://genomevolution.org/coge/data/diags/62209/62346/html/master_62209_62346.CDS-CDS.last.tdd10.cs0.filtered.dag.all.go_D20_g10_A5.aligncoords.gcoords_ct0.w1000.spa-1.mcs1000000.ks.sr.cs1.csoN.log.nsd.spa_info.txt). This initial analysis does not show any *P. aphrodite*'s chromosomes that would not have a match in CR0040. Given the information available: genome size, chromosome size, complete BUSCO score for the annotated protein coding genes, and comparative genomics results, there is no strong evidence that essential protein-coding genes are missing from the current CR0040 genome assembly (chromosomes 15 and 16 would be in the unplaced scaffolds).

Comparison between CR0040 and *P. aphrodite* for the analysis of the pan-orchid WGD is difficult, in part because the 19 chromosomes of the latter are not well enough assembled. However, the chromosome 1 of CR0040 contains internal duplications as if the same two ancestral chromosomes (pre-α° WGD) had been merged (Supplemental Figure 8C). This is not observed in *P. aphrodite* that seems to have four chromosomes (1, 2, 10b, 17) corresponding to the Vanilla chromosome 1. The comparison of the two Ks histograms of a genome against itself (Supplemental Figure 8B and 7F) indicated the same kind of peak profile but narrower for *P. aphrodite*. Comparison of the vanilla genome with the genome of other orchids to search for the missing vanilla chromosomes 15 and 16 does not seem to be an appropriate approach because these chromosomes are probably small, specific and contain few genes. Their absence would therefore not be an obstacle to the study of the evolution of monocot genomes.

Supplemental Note 5

Detection of non-endoreplicated genomic regions

Reads from each sequencing technology used in this study (HiFi long reads, ONT long reads and Illumina short reads from CR0040), as well as ONT long reads and Illumina short reads from Daphna were mapped onto the CR0040 assembly. Minimap2 (Li, 2018) was used for long reads mapping with the following parameters: -a -x asm20 -t8, and Illumina short reads were mapped using bwa-mem2 (https://github.com/bwa-mem2/bwa-mem2) with default parameters. Sequencing depths were extracted with the samtools depth script for each position. These values were averaged for genomic windows of 20Kb which were used to compute global statistics on chromosomes and to plot depth distributions along the 28 chromosomes and the two random mosaic chromosomes

(chr0). Mean sequencing depth of the genome was 19.57, 29.82 and 52.50 for CR0040 cultivar with HiFi, ONT and Illumina reads, and 27.29 and 46.03 for Daphna cultivar with ONT and Illumina reads. The chromosome 11A is less covered by sequencing due to the presence of large gaps in its assembly. The sequencing depth of CR0040 genome per technology is summarized in the Supplemental Table 15.

The sequencing depth of CR0040 genome using Daphna data sequences (Hasing et al., 2020) is summarized in the Supplemental Table 16. The mean sequencing depth of the two random mosaic chromosomes (chr0) was lower than the one of the 14 pseudochromosomes for long reads sequencing technologies (Supplemental Tables 15 and 16). To detect sequencing depth bias and to limit the risk to detect false positives, the mean of sequencing depth for every 20 successive windows of 20Kb was computed using Illumina reads for Daphna and using long reads (HiFi + ONT) for CR0040. All regions with a sequencing depth inferior to 20% of mean sequencing depth of chromosomes were selected for both Daphna and CR0040. Regions with a decrease in sequencing depth for both cultivars were identified using the 'bedtools intersect' (version 2.29.2) and correspond to the endoreplication pattern 1. Low coverage regions identified from the HiFi CR0040 reads mapping that did not overlap with those identified from the Daphna Illumina reads mapping corresponded to the endoreplication pattern 2. All automatically identified regions were manually validated and refined by visualization of sequencing depth drops for each CR0040 chromosome and for all available sequencing datasets (Supplemental Figure 9).

Supplemental Figure 10 shows low depth k-mers information in addition to the coverage depth information presented in Supplemental Figure 9. The ratios of k-mers of depth between 5x and 15x were computed for genomic windows of 20Kb and plotted along each pair of chromosomes.

Supplemental Note 6

Database implementation for Vanilla Genome Hub (VGH)

The VGH was constructed using the Tripal v3 framework, a specific toolkit for the construction of online community genomic databases, by integrating the GMOD Chado database schema and the Drupal open source platform (https://www.drupal.org/). The genome sequences, predicted gene models, mRNA and protein sequences were loaded into the database using the 'Data Loaders' function of Tripal. The Drupal theme of the website is derived from the hardwood genomics website (https://github.com/statonlab/hardwood).

Gene Search

Gene search can be done in two ways, (i) by querying genes using the Tripal MegaSearch module (Jung, et al, 2021) and various filters such as name, chromosome position and functional annotation (putative function, InterPro domain and gene ontology, Figure 6B), (ii) by similarity using BLAST. The BLAST search interface provides access to datasets from the current genome release, allowing users to conduct sequence searches against the coding sequences, transcripts, proteins and the whole-genome assembly (Figure 6C).

For both, results are displayed as a dynamic table that summarizes the relevant information on the corresponding search with a link to the gene report and onto the genome browser.

Gene report page provides the user with a view of the metadata associated with the gene models. These include protein domain composition identified by InterProScan, Gene Ontology (GO) annotations and KEGG BlastKOALA (Figure 6D).

Genome Browser

Genome visualization was implemented using the Tripal JBrowse module (https://github.com/tripal/tripal_jbrowse) (Figure 6E). The module embeds JBrowse (Buels et al., 2016), an interactive, client-side genome browser, into a Drupal webpage. JBrowse allows users to visualize features of the reference genome. Each chromosome can be selected from a drop-down menu, and the browser displays information about the sequence and corresponding gene models. The gene expression profiles are also displayed in JBrowse. To display RNA-Seq expression profiles in JBrowse, the read alignment file in BAM format for each accession was converted to coverage tracks in bigwig format using DeepTools2, and these coverage tracks were then loaded onto JBrowse.

Metabolic pathway

The Pathologic file previously produced was processed by Pathway Tools, which predicted a total of 678 metabolic pathways in vanilla. A pathway database, VanillaCyc (https://vanilla-genome-hub.cirad.fr/content/pathways-tools), was built based on these predicted pathways using the web server of Pathway Tools (Figure 6F). Users can search and browse the predicted pathways, as well as perform comparative and omics data analysis through the VanillaCyc database.

If we take the example of the methyltransferase gene family (Enzyme Commission Number 2.1.1.-), 415 methyltransferases were predicted on both haplotypes CR0040_A and CR0040_B. The O-methyltransferase are involved, among other things, in four phenylpropanoid biosynthetic pathways predicted by Pathway Tools:

The phenylpropanoid biosynthesis (PWY-361)

The free phenylpropanoid acid biosynthesis (PWY-2181)

The vanillin biosynthesis I (PWY-5665)

The phenylpropanoids methylation (ice plant; PWY-7498)

If we focus on the example of two CR0040 methyltransferases:

VANPL_B_00007t006130 (OG0000841), the only Caffeic acid O-methyltransferase predicted in the vanillin biosynthesis I pathway and also putatively involved in the phenylpropanoid biosynthesis and

VANPL_B_00011t00710 (OG0000449), OMT4 (Vpl_s126Bg26946.1) homolog possibly involved in the phenylpropanoid biosynthesis and the phenylpropanoids methylation (ice plant).

Phylogenetic trees of the orthogroups of these polypeptides showed clusters of genes encoding for Caffeic acid O-methyltransferase (COMT, OG0000841; Supplemental Figure 11A, B) and for O-methyltransferase (OMT, OG0000449; Supplemental Figure 11C, D). The two OrthoFinder gene families, OG0000449.fa and OG0000841.fa, are available in the VGH (see download section below). They were reanalyzed with https://ngphylogeny.fr/ using the PhyML/OneClick workflow and advanced setting. Default parameters were used except for PhyML. Tree topology search and branch support option were set to “best of nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR)” and approximate likelihood-ratio test, respectively. Gene names are indicated in blue, green, red, black, magenta and orange for CR0040 *V. planifolia*, Daphna *V. planifolia*, *Arabidopsis thaliana*, *Oryza sativa* subsp. *japonica* and *Phalaenopsis equestris*, respectively.

Regarding these two OrthoFinder orthogroups, we observed that there is a higher number of genes predicted in the Daphna genome than in the CR0040 one and for CR0040, methyltransferases are preferentially found on haplotype B (Supplemental Figure 11). This could suggest phenomena that may be concomitant such as variations of the dispensable genome, assembly problems in the two *Vanilla planifolia* genomes or annotations. Indeed, in both gene families, the Daphna haplotype A appears to contain tandem methyltransferase clusters (green boxes on DA07 and DA11 genomic regions in JBrowse; Supplemental Figure 11B, D). However, it can be seen in the genomic region containing a tandem methyltransferase cluster on Daphna_A chromosome 7, that the cluster is interrupted by genes that are repeated in 5' (grey boxes in the Supplemental Figure 11B): Protein disulfide-isomerase LQY1, Type I inositol-1,4,5-trisphosphate 5-phosphatase, Pentatricopeptide repeat-containing protein, Activating signal cointegrator 1. This may reflect the assembly of two allelic long reads instead of them being separated in each of the two Daphna haplotypes. This is consistent with other remarks already made along these lines in this article (e.g. optical mapping, syntenic dotplot). As illustrated with theses isozyme examples, it is therefore possible that this is the same reason why the chromosome scaffolded part is larger in Daphna than in CR0040 and why the

number of paralogs is overestimated in Daphna compared to CR0040. On the contrary, the CR0040_A chromosomes would possibly underestimate tandem gene clusters.

GO enrichment analysis

From a list of genes of interest, it can be interesting to look for enriched biological processes, molecular functions, or cellular components. This functionality is brought by DIANE (Dashboard for the Inference and Analysis of Networks from Expression data), as described in Cassan, et al (2021). Results can be obtained as a data table, a dotplot of enriched GO terms with associated gene counts and p-values, or an enrichment map linking co-occurring GO terms (Figure 6G).

Synteny analysis and visualization

Syntenic blocks and homologous gene pairs within syntenic blocks were identified in the two vanilla genome sequences, including comparisons both within each genome and between any two genomes (haplotypes). The protein sequences were first aligned against themselves (within each genome) as well as between each other (pairwise comparisons) using a reciprocal best hit approach. Based on the BLASTP results and gene positions, syntenic blocks were determined using MCScanX with default parameters (Wang et al., 2012). To further facilitate comparative analyses of genomes, the Synvisio extension module (https://www.drupal.org/project/synvisio) was implemented. The module embeds Synvisio (Bandi, 2020), an interactive multiscale synteny visualization tool for MCScanX, into a Drupal webpage and provides a simple interface for managing and creating Synvisio instances (Figure 6H). Liftoff (Shumate and Salzberg, 2021) was used to align the gene sequences of Daphna to the CR0040 genome. For each gene, Liftoff finds the alignments of the exons that maximize sequence identity while preserving the transcript and gene structure. This track can be loaded onto JBrowse to visualize the collinearity between the 2 sets of annotations.

Download section

A download section grants users direct access to the data used by the different tools that compose the hub. Assembly of pseudomolecules as well as their structural and functional annotations are available in FASTA and in GFF3 formats respectively at https://vanilla-genome-hub.cirad.fr/content/download. This section was created using the Filebrowser module (https://www.drupal.org/project/filebrowser).

Supplemental References

Bandi, V. K. (2020). SynVisio: a multiscale tool to explore genomic conservation, Master of Science (M.Sc.) dissertation, University of Saskatchewan. 82p.

Bory, S., Da Silva, D., Risterucci, A.M., Grisoni, M., Besse, P., and Duval, M.F. (2008). Development of microsatellite markers in cultivated vanilla: Polymorphism and transferability to other vanilla species. *Scientia Horticulturae***,** 115, 420-425. 10.1016/j.scienta.2007.10.020.

Bouétard, A., Lefeuvre, P., Gigant, R., Bory, S., Pignal, M., Besse, P., and Grisoni, M. (2010). Evidence of transoceanic dispersion of the genus *Vanilla* based on plastid DNA phylogenetic analysis. *Molecular Phylogenetics and Evolution*, 55(2), 621-630.

Bourge, M., Brown, S. C., and Siljak-Yakovlev, S. (2018). Flow cytometry as tool in plant sciences, with emphasis on genome size and ploidy level assessment. *Genetics and Applications*, 2(2), 1-12.

Brown, S.C., Bourge, M., Maunoury, N., Wong, M., Bianchi, M.W., Lepers-Andrzejewski, S., Besse, P., Siljak-Yakovlev, S., Dron, M., and Satiat-Jeunematre, B. (2017). DNA remodeling by strict partial endoreplication in Orchids, an original process in the plant kingdom. *Genome Biology and Evolution*, 9, 1051-1071. 10.1093/gbe/evx063.

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L., and Holmes, I. H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology*, 17, 66. https://doi.org/10.1186/s13059-016-0924-1.

Cassan, O., Lèbre, S. and Martin, A. (2021). Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. *BMC Genomics* 22, 387. https://doi.org/10.1186/s12864-021-07659-2.

Chao, Y. T., Yen, S. H., Yeh, J. H., Chen, W. C., and Shih, M. C. (2017). Orchidstra 2.0-A Transcriptomics Rresource for the orchid family. *Plant and cell physiology*, 58(1), e9. https://doi.org/10.1093/pcp/pcw220.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods*, 18(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5.

Claudel-Renard, C., Chevalet, C., Faraut, T., and Kahn, D. (2003). Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic acids research*, 31(22), 6633–6639. https://doi.org/10.1093/nar/gkg847.

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364.

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics*, 9, 18. https://doi.org/10.1186/1471-2105-9-18.

Favre, F., Jourda, C., Grisoni, M., Rivallan, R., Dijoux, J. B., Hascoat, J., Lepers, S., Besse, P., and Charron, C. (2022) A genome-wide assessment of the genetic diversity, evolution and relationships with allied species of the clonally propagated crop *Vanilla planifolia*. *Genetic Resources and Crop Evolution*, in press. https://doi.org/10.1007/s10722-022-01362-1.

Flynn JM, Hubley R, Goubert C, Rosen, J., Clark, A. G., Feschotte, C., and Smit, A. F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17):9451-9457. doi:10.1073/pnas.1921046117.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

Girgis H. Z. (2015). Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC bioinformatics*, 16, 227. https://doi.org/10.1186/s12859-015-0654-5.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Haas BJ, Delcher AL, Wortman JR and Salzberg SL. (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*. 20(18):3643-3646. doi:10.1093/bioinformatics/bth397.

Hasing, T., Tang, H.B., Brym, M., Khazi, F., Huang, T.F., and Chambers, A.H. (2020). A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. *Nature Food* 1:811-819. 10.1038/s43016-020-00197-2.

Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E., and Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics*, 33(14), 2197–2198. https://doi.org/10.1093/bioinformatics/btx144.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). PASTEC: an automatic transposable element classification tool. *PloS one*, 9(5), e91929. https://doi.org/10.1371/journal.pone.0091929.

Jiao, Y.N., Li, J.P., Tang, H.B., and Paterson, A.H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* **26**:2792-2802. 10.1105/tpc.114.127597.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics,* 30(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Jung, S., Cheng, C. H., Buble, K., Lee, T., Humann, J., Yu, J., Crabb, J., Hough, H. and Main, D. (2021) Tripal MegaSearch: a tool for interactive and customizable query and download of big data. *Database*. doi: 10.1093/database/baab023.

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of molecular biology*, *428*(4), 726–731. https://doi.org/10.1016/j.jmb.2015.11.006.

Karp, P. D., Paley, S., and Romero, P. (2002). The Pathway Tools software. *Bioinformatics*, 18, S225–S232. https://doi.org/10.1093/bioinformatics/18.suppl_1.s225.

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, 37(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4.

Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37:540.

Kumar, S., Stecher, G., Suleski, M. and Hedges, SB. (2017) TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* doi:10.1093/molbev/msx116.

Lamesch, P., Berardini, TZ., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, DL., Garcia-Hernandez, M., Karthikeyan, AS., Lee, CH., Nelson, WD., Ploetz, L., Singh, S., Wensel, A. and Huala, E. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research*. 40 Database issue D1202-10.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094-3100.

Lyons, E. and Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal* 53 (4), 661-673.

Martin, J., and Hesemann, C. U. (1988). Cytogenetic investigations in wheat, rye and triticale. I. Evaluation of improved Giemsa C-and fluorochrome banding techniques in rye chromosomes. *Heredity*, 61(3), 459-467.

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics,* btaa1022. Advance online publication. https://doi.org/10.1093/bioinformatics/btaa1022.

Perez-Silva, A., Odoux, E., Brat, P., Ribeyre, F., Rodriguez-Jimenes, G., Robles-Olvera, V., Garcia-Alvarado, M.A., and Gunata, Z. (2006). GC-MS and GC-olfactometry analysis of aroma compounds in a representative organic aroma extract from cured vanilla (*Vanilla planifolia* G. Jackson) beans. *Food chemistry,* 99:728-735. 10.1016/j.foodchem.2005.08.050.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, 33(3), 290–295. https://doi.org/10.1038/nbt.3122.

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics*, 21, i351–i358. https://doi.org/10.1093/bioinformatics/bti1018.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Sallet, E., Gouzy, J., and Schiex, T. (2019). EuGene: An automated integrative gene finder for Eukaryotes and Prokaryotes. *Methods in molecular biology* (Clifton, N.J.), 1962, 97–120. https://doi.org/10.1007/978-1-4939-9173-0_6.

Sandler, G., Bartkowska, M., Agrawal, A.F., and Wright, S.I. (2020). Estimation of the SNP mutation rate in two vegetatively propagating species of duckweed. *G3-Genes Genomes Genetics* 10:4191-4200. 10.1534/g3.120.401704.

Schweizer, D. (1976). Reverse fluorescent chromosome banding with chromomycin and DAPI. *Chromosoma*, 58(4), 307-324.

Shumate, A., and Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations, *Bioinformatics*, 37, 1639–1643, https://doi.org/10.1093/bioinformatics/btaa1016.

Siljak-Yakovlev, S., Cerbah, M., Coulaud, J., Stoian, V., Brown, S. C., Zoldos, V., Jelenic, S., and Papes, D. (2002). Nuclear DNA content, base composition, heterochromatin and rDNA in *Picea omorika* and *Picea abies*. *Theoretical and Applied Genetics*, 104(2), 505-512.

Solano-De la Cruz, M.T., Adame-García, J., Gregorio-Jorge, J., Jiménez-Jacinto, V., Vega-Alvarado, L., Iglesias-Andreu, L. G., Escobar-Hernández, E. E., and Luna-Rodríguez, M. (2019) Functional categorization of *de novo* transcriptome assembly of *Vanilla planifolia* Jacks. potentially points to a translational regulation during early stages of infection by *Fusarium oxysporum f. sp. vanillae*. *BMC Genomics*, 20(1):826. PMID: 31703622.

Tang, H., Bowers, J. E., Wang, X., and Paterson, A. H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(1), 472–477. https://doi.org/10.1073/pnas.0908007107.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, https://doi.org/10.1002/0471250953.bi0410s25.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., 1 Young, S.K. and Earl, A.M. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9:e112963. doi: 10.1371/journal.pone.0112963.

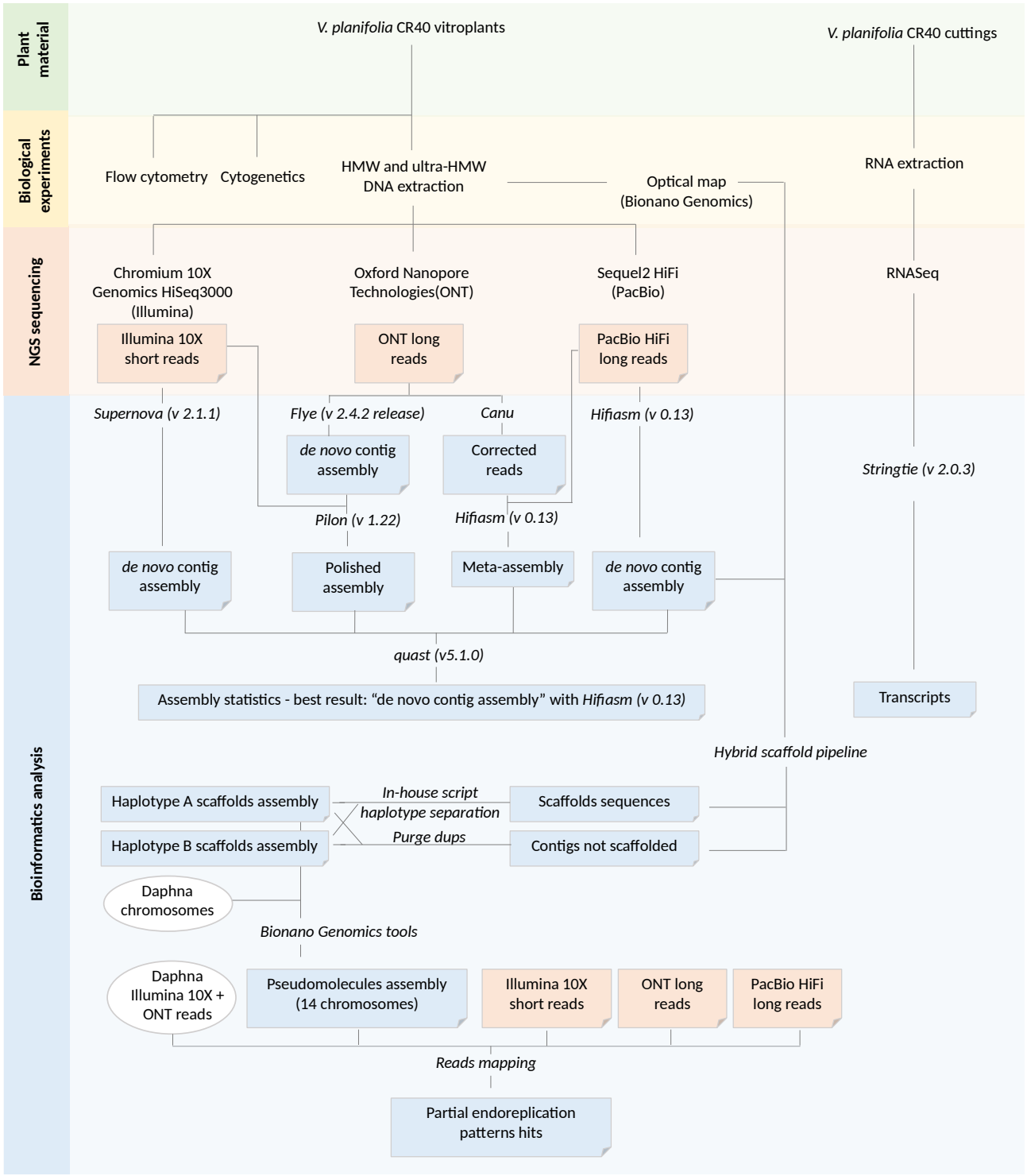
Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., Kissinger, J.C. and Paterson, A.H. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40(7):e49. doi: 10.1093/nar/gkr1293.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome research*, *27*(5), 757-767.

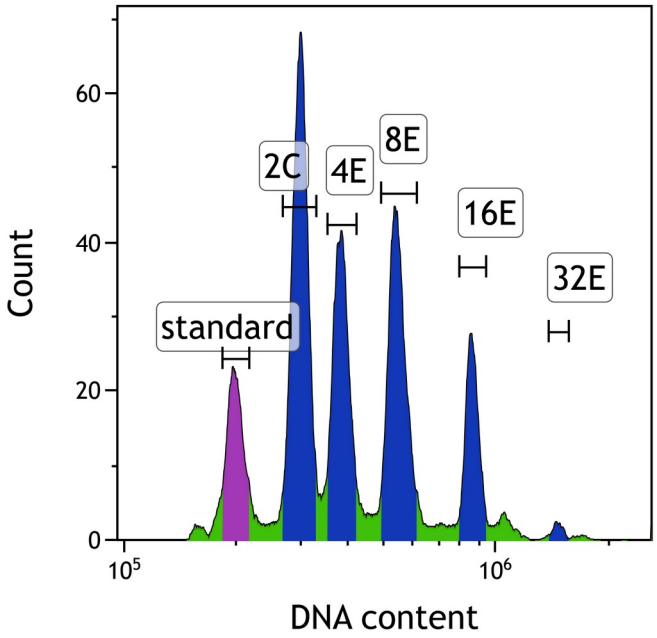
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics*, 8(12), 973–982. https://doi.org/10.1038/nrg2165.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586-1591. doi:10.1093/molbev/msm088.

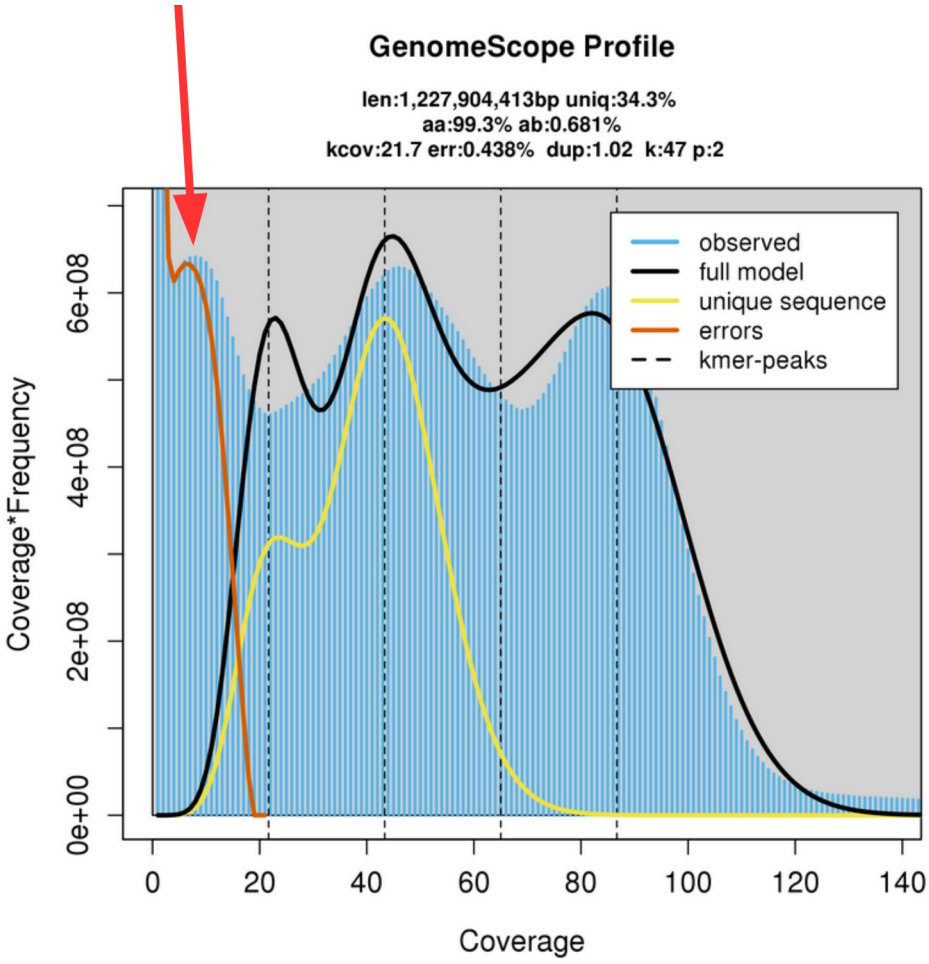
Supplemental figures



Supplemental Figure 1 Schematic representation of CR0040 genome sequencing strategy. Data represented with white circles were retrieved from *V. planifolia* Daphna genome (Hasing et al, 2020).



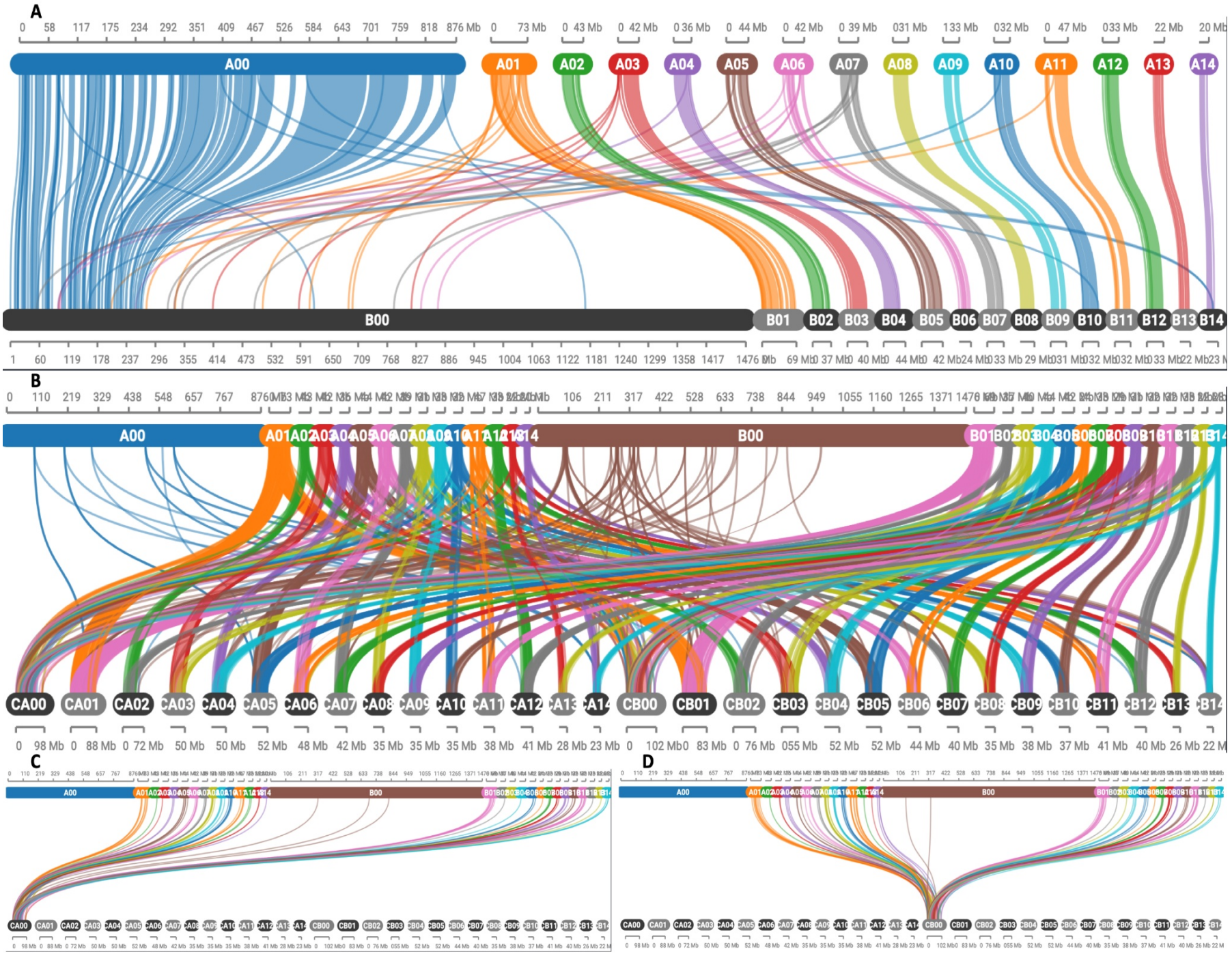
Supplemental Figure 2 Example of DNA content histogram of nodal tissue from *Vanilla planifolia* CR0040.



Supplemental Figure 3 GenomeScope2 k-mer profile of the CR0040 genome assembly

The red arrow shows k-mers from the non-endoreplicated part of the CR0040 genome whose distribution is centered at 10X. The two k-mers distributions centered at 42X and 84X, represent the classical diploid distribution with the heterozygous and homozygous k-mer content.

3

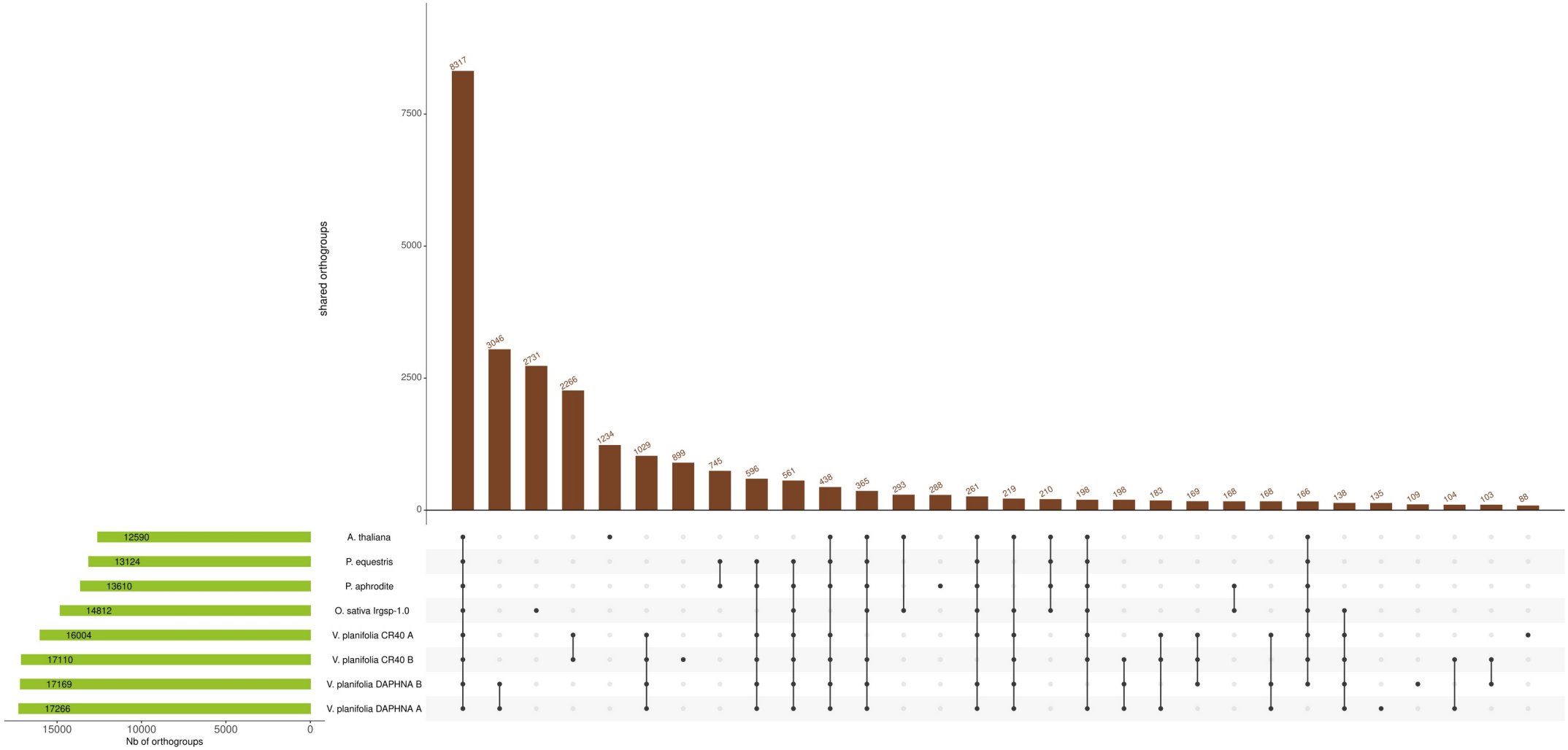

A
A00
A01
A02
A03
A04
A05
A06
A07
A08
A09
A10
A11
A12
A13
A14
B00
B01
B02
B03
B04
B05
B06
B07
B08
B09
B10
B11
B12
B13
B14
B
A00
B00
CA00
CA01
CA02
CA03
CA04
CA05
CA06
CA07
CA08
CA09
CA10
CA11
CA12
CA13
CA14
CB00
CB01
CB02
CB03
CB04
CB05
CB06
CB07
CB08
CB09
CB10
CB11
CB12
CB13
CB14
C
D

4

Supplemental Figure 4 Synteny view of the two *V. planifolia* genomes.

A. Comparison of MCScanX allelic blocks of both CR0040 haplotypes A and B. B) Comparison of MCScanX allelic blocks of both CR0040 haplotypes and of both Daphna haplotypes. C) Same as B) but it only shows syntenic relationships between the unknown random pseudomolecule of Daphna haplotype A (A0) and the CR0040 pseudomolecules in order to pinpoint improvement of CR0040 scaffolding. D) same as C) but with Daphna B0. Screenshots were done from https://vanilla-genome-hub.cirad.fr/synvisio.

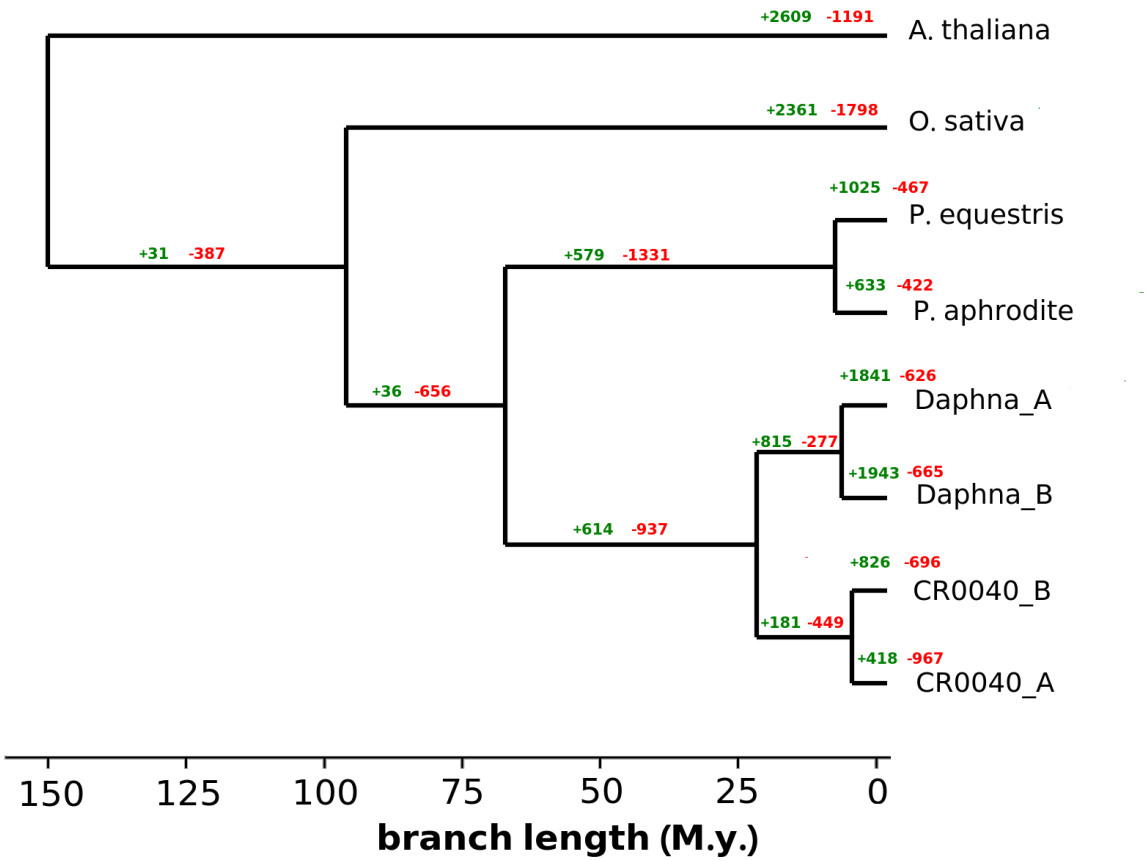
5



Supplemental Figure 5 UpSet plot of the inferred orthogroups.

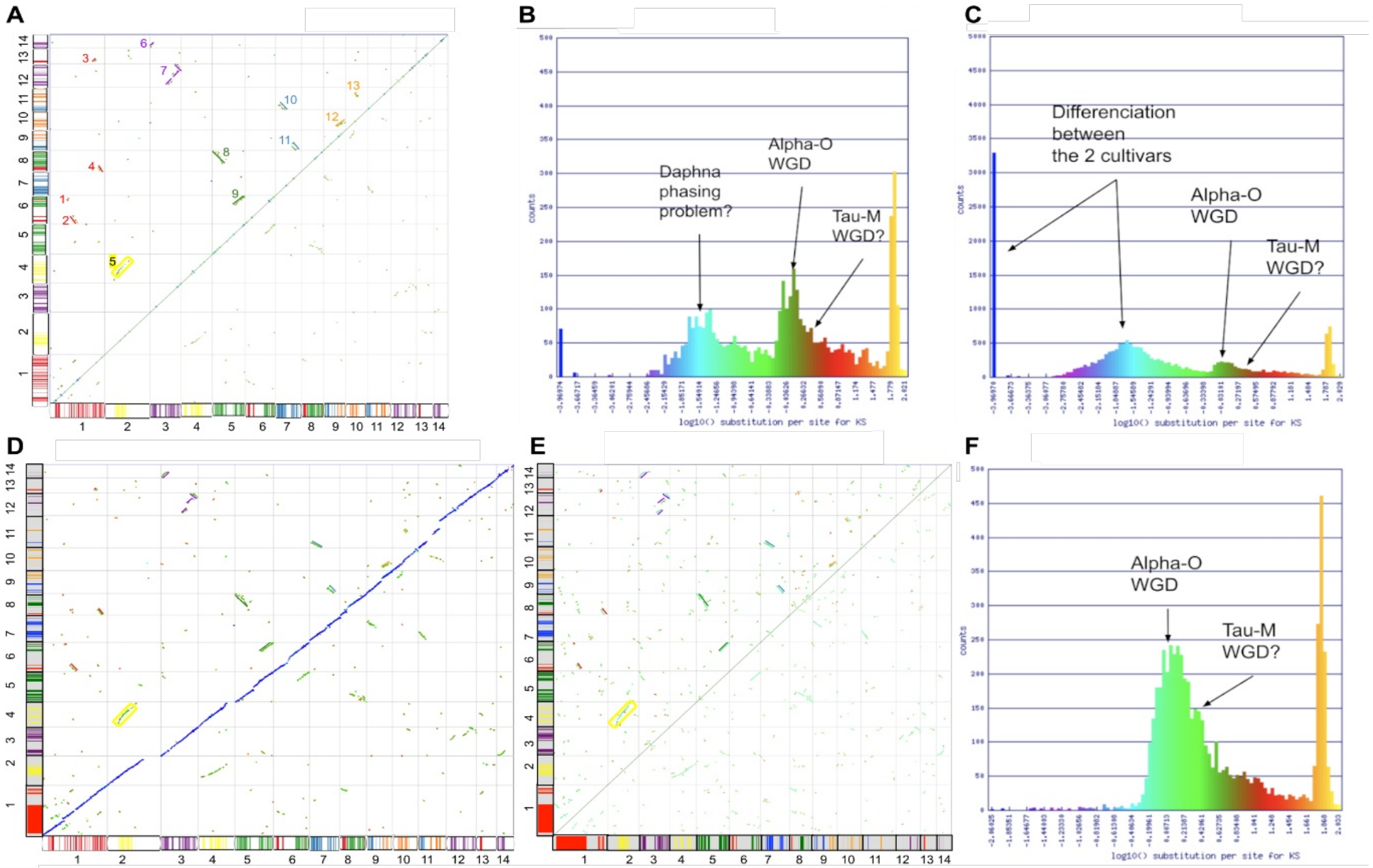
The number of orthogroups identified with OrthoFinder for each species and *V. planifolia* haplotype is displayed inside horizontal green bars. Numbers above vertical brown bars represent numbers of orthogroups shared by a given intersection of species/haplotypes illustrated downside each vertical bar. For clarity reasons, only the 30 most frequent intersections were displayed on the plot (Nb, numbers).

6



Supplemental Figure 6 Expansion and reduction of gene families in six proteomes. Orthogroups from OrthoFinder on six proteomes (CR0040_A, CR0040_A, Daphna_A, *Phalaenopsis equestris, P. aphrodite, Arabidopsis thaliana, Oryza sativa*) were analyzed with CAFE. The values represent the number of genes gained/lost.

7

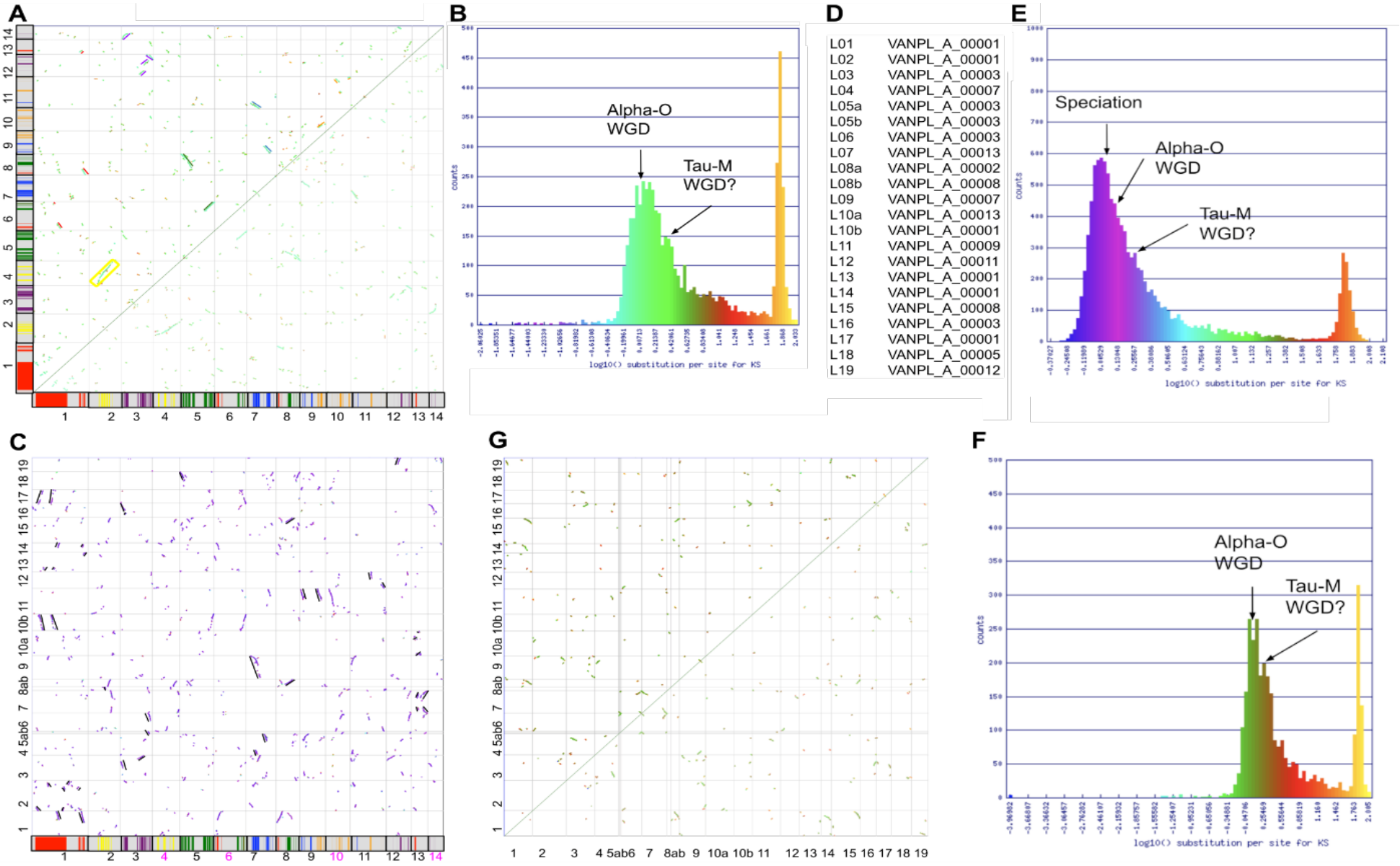

A
B
Daphna
phasing
problem?
Alpha-O
WGD
Tau-M
WGD?
counts
log10() substitution per site for KS
C
Differenciation
between
the 2 cultivars
Alpha-O
WGD
Tau-M
WGD?
counts
log10() substitution per site for KS
D
E
F
Alpha-O
WGD
Tau-M
WGD?
counts
log10() substitution per site for KS

8

Supplemental Figure 7 Synmap pairwise genome synteny analyses between CR0040 and Daphna cultivars and between themselves.

Haplotype A syntenic comparisons of Daphna CDS against themselves (A, B, see details in https://genomevolution.org/r/1jei9), between CR0040 and Daphna cultivars (C, D, see details in https://genomevolution.org/r/1jeid) and of CR0040 against itself (E, F, see details in https://genomevolution.org/r/1j9ym) were visualized with dotplots (A, D, E) and histograms of Ks values (B, C, F). Daphna karyotype painted according six pairs of duplicated regions illustrating the pan-orchid α^{o} WGD of Figure 4 (Hasing et al., 2020) has been cut out and positioned below the corresponding chromosomes of the dotplots (A, D). 13 paralog blocks were then projected onto the diagonals. Furthermore, using reciprocal best hits, the Daphna karyotype could be transferred to CR0040 (D, E).

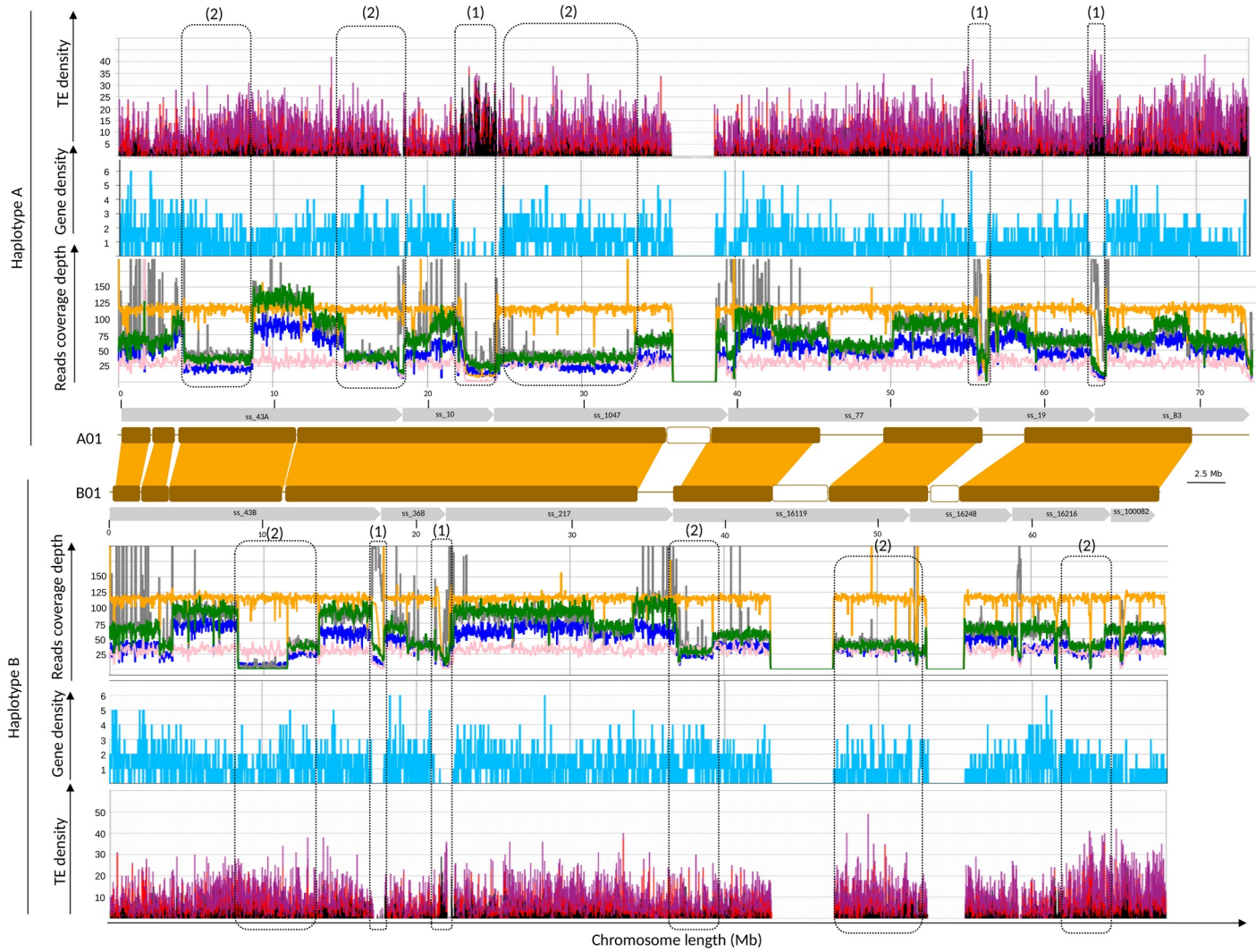
9


A
B
Alpha-O
WGD
Tau-M
WGD?
log10() substitution per site for KS
counts
D
L01 VANPL_A_00001
L02 VANPL_A_00001
L03 VANPL_A_00003
L04 VANPL_A_00007
L05a VANPL_A_00003
L05b VANPL_A_00003
L06 VANPL_A_00003
L07 VANPL_A_00013
L08a VANPL_A_00002
L08b VANPL_A_00008
L09 VANPL_A_00007
L10a VANPL_A_00013
L10b VANPL_A_00001
L11 VANPL_A_00009
L12 VANPL_A_00011
L13 VANPL_A_00001
L14 VANPL_A_00001
L15 VANPL_A_00008
L16 VANPL_A_00003
L17 VANPL_A_00001
L18 VANPL_A_00005
L19 VANPL_A_00012
E
Speciation
Alpha-O
WGD
Tau-M
WGD?
log10() substitution per site for KS
counts
C
G
F
Alpha-O
WGD
Tau-M
WGD?
log10() substitution per site for KS
counts

10

Supplemental Figure 8 Synmap pairwise genome synteny analyses between CR0040 *V. planifolia* and *P. aphrodite* and between themselves. Haplotype A syntenic comparisons of CR0040 CDS against themselves (A, B, see details in https://genomevolution.org/r/1j9ym), between CR0040 and *P. aphrodite* genome (C, D, E, see details in https://genomevolution.org/r/1jekr) and *P. aphrodite* genome itself (G, F, see details in https://genomevolution.org/r/1jekt) were visualized with dotplots (A, C, G), histograms of Ks values (B, E, F) and correspondence between the chromosomes of the two species was indicated according to Syntenic Path Assembly (D; SPA option; https://genomevolution.org/r/1jjmm). The CR0040 karyotype comes from the transfer of that of Daphna (Hasing et al., 2020) using reciprocal best hits. It has been cut out and positioned below the corresponding chromosomes of dotplots (A, C). Thus, six pairs of duplicated regions illustrating the pan-orchid α^o WGD allowed to project paralog blocks onto the diagonals (A). The orthologous blocks between CR0040 and *P. aphrodite* were highlighted with black lines (C) according to SPA results (D).

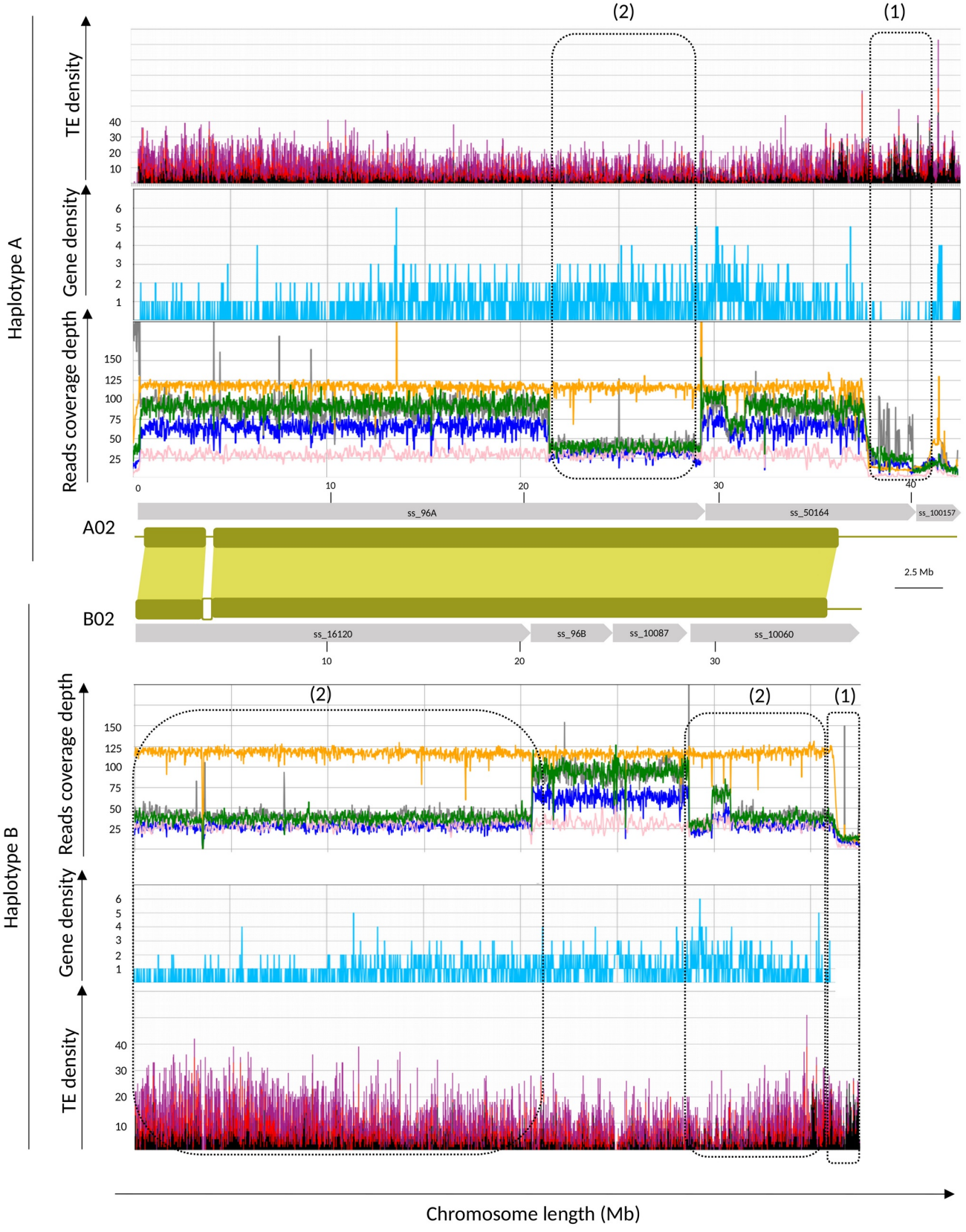
11



Supplemental Figure 9 Sequencing depth along the CR0040 chromosomes obtained by mapping Daphna Illumina (yellow) and ONT (pink) reads, CR0040 PacBio Hifi (blue), Nanopore (green) and Illumina (grey) reads on the CR0040 assembly. Synteny between homologous chromosomes are represented by central boxes. Gaps (N stretches) explaining sudden drops in sequencing depth are shown with white blocks. (1) Pattern 1 corresponds to low level of sequencing depth for all data. (2) Pattern 2 corresponds to inverted level of sequencing depth for CR0040 between haplotypes A and B, and constant level of sequencing depth for both Daphna haplotypes. Gene and retrotransposons distributions along the chromosomes are represented by a blue line chart and a stacked histogram (copia: red, gypsy: purple, other retrotransposons: black) respectively.

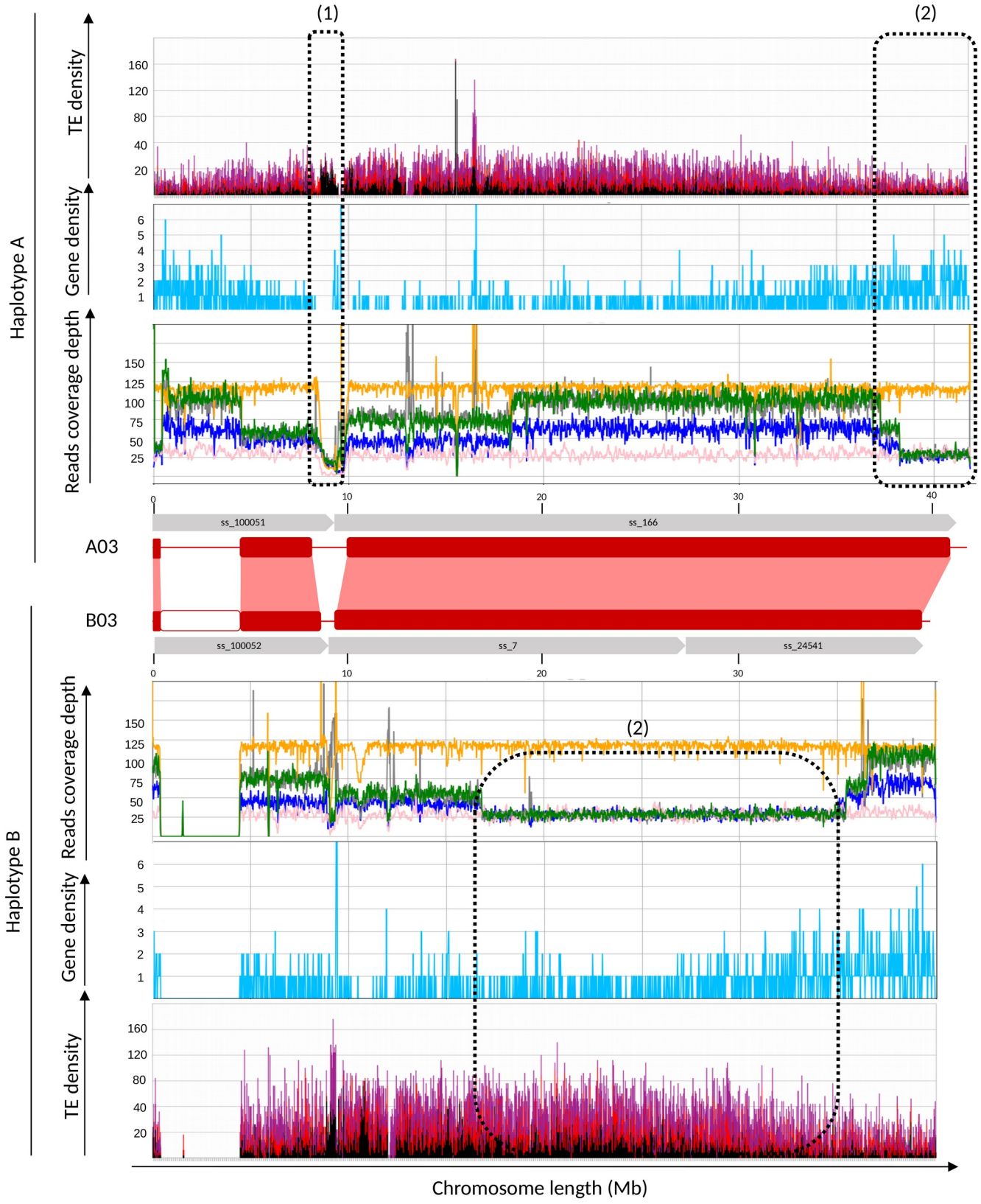
12

Supplemental Figure 9 continued



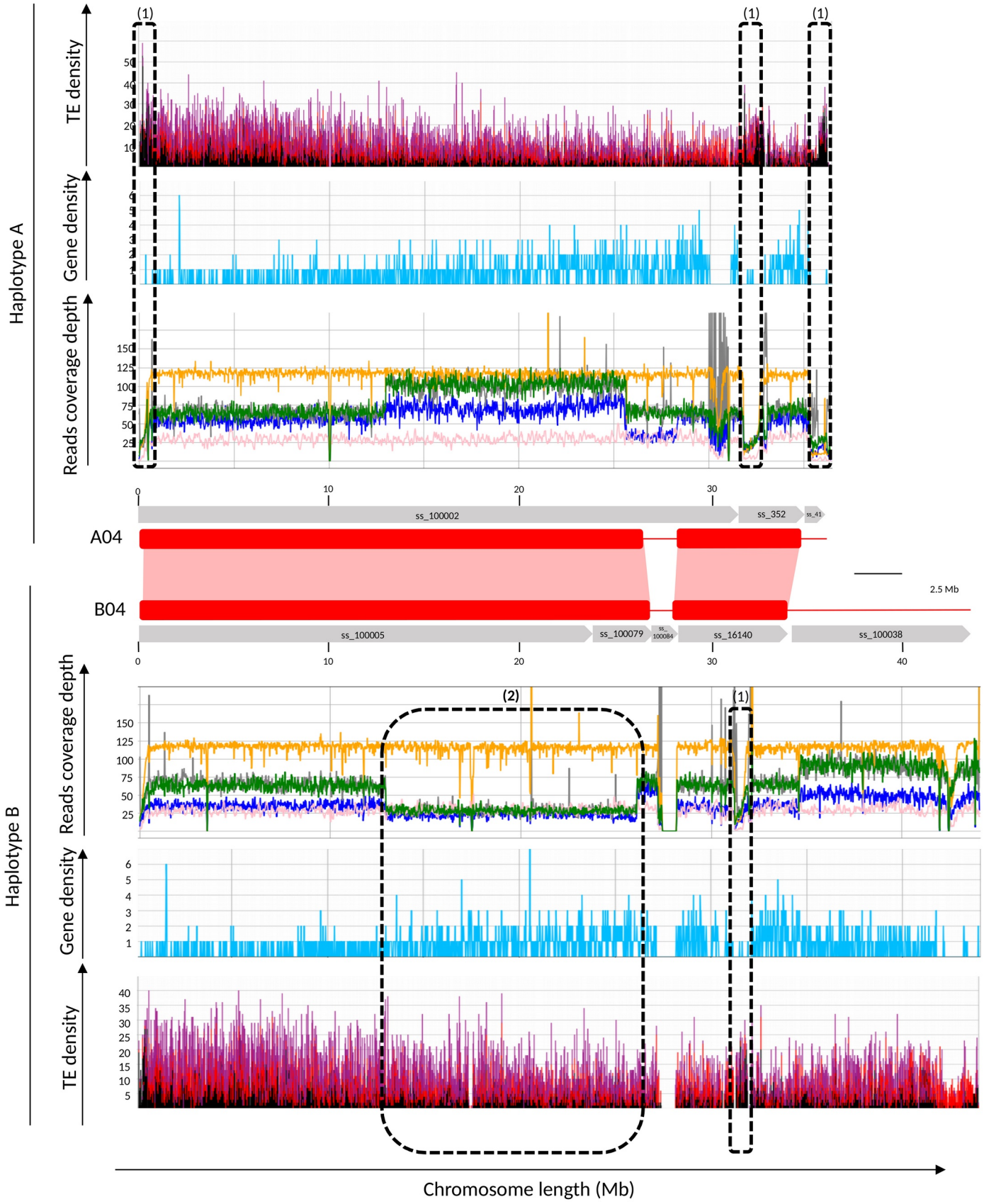
13

Supplemental Figure 9 continued



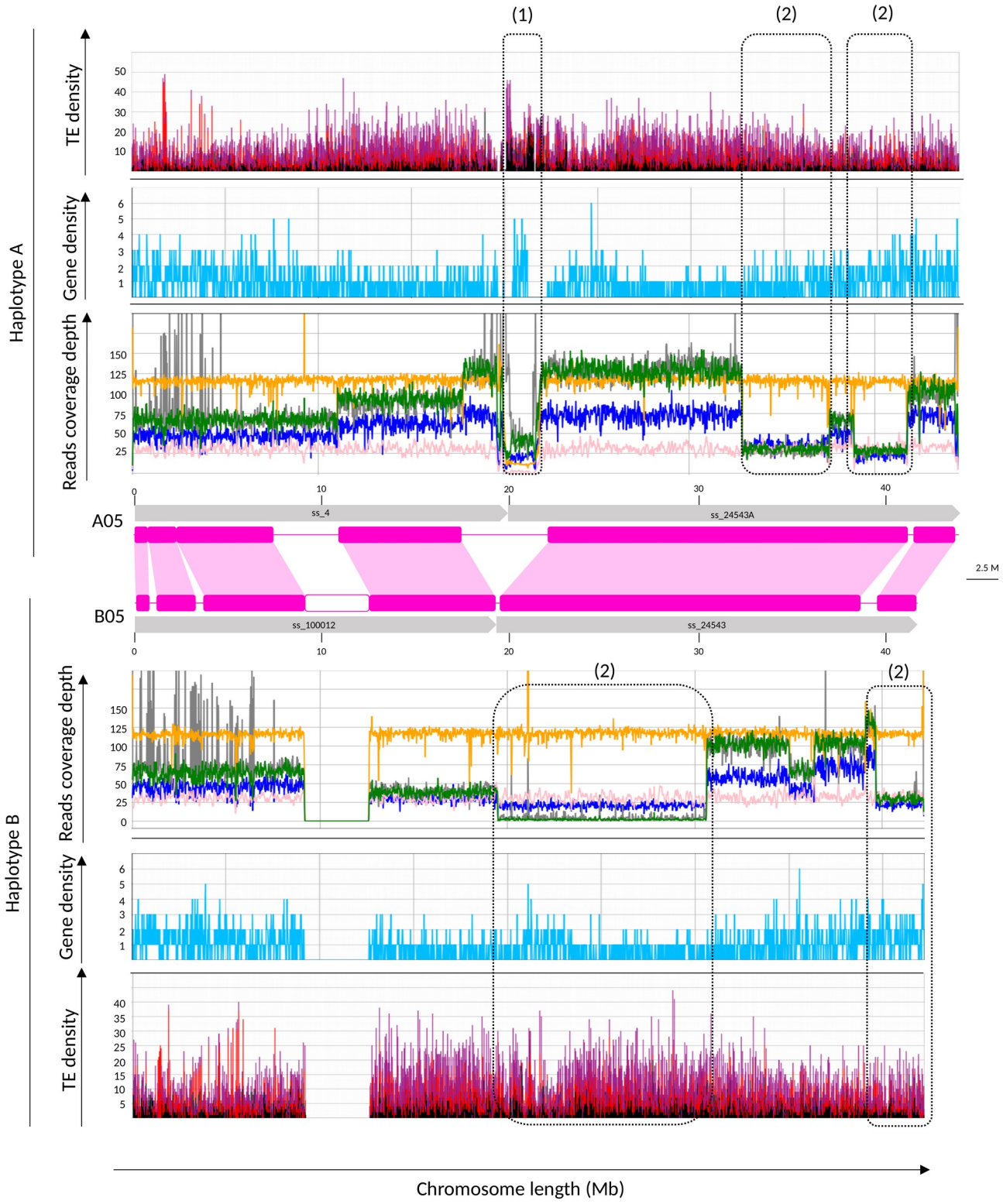
14

57 Supplemental Figure 9 continued



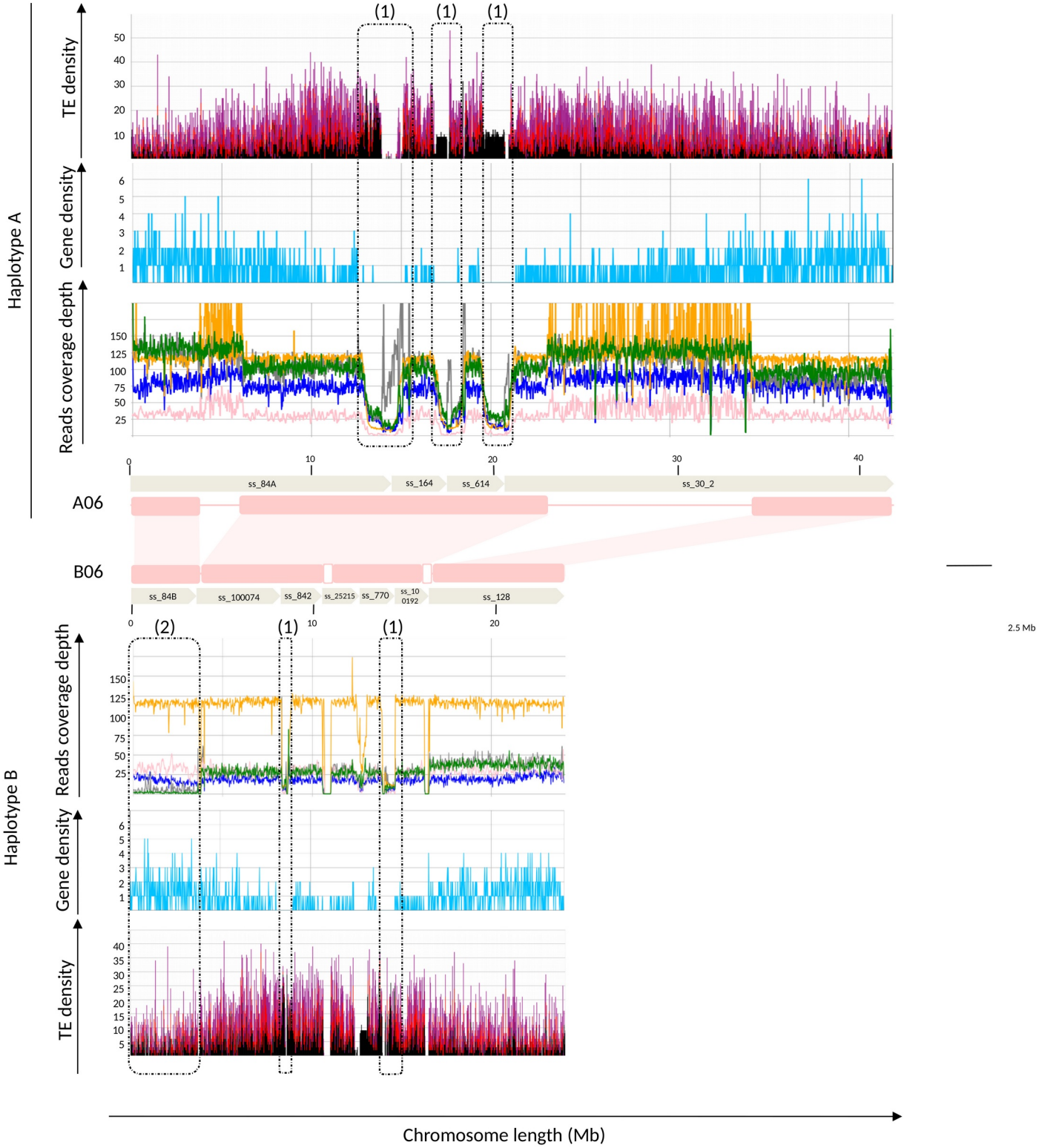
15

Supplemental Figure 9 continued



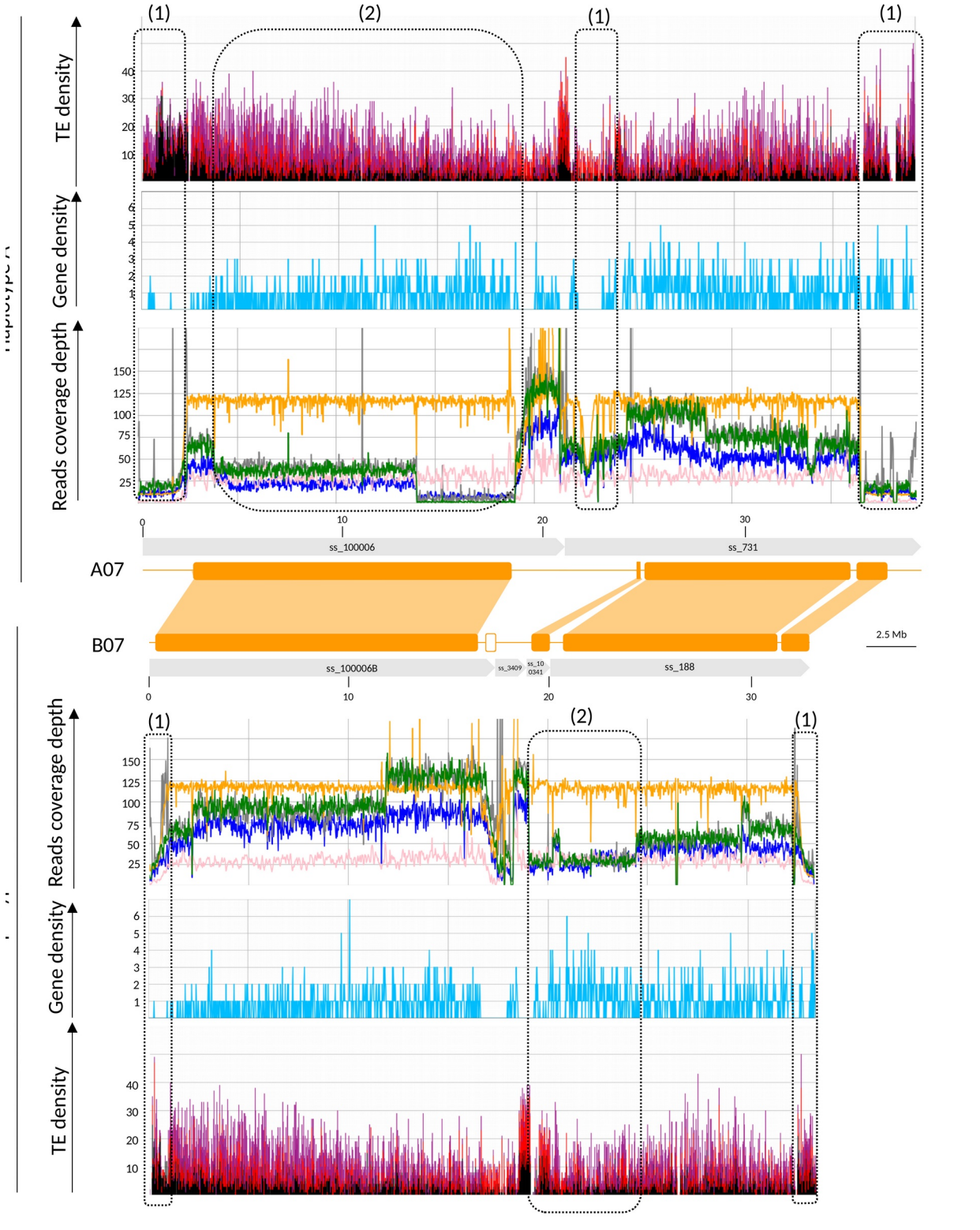
16

59 Supplemental Figure 9 continued



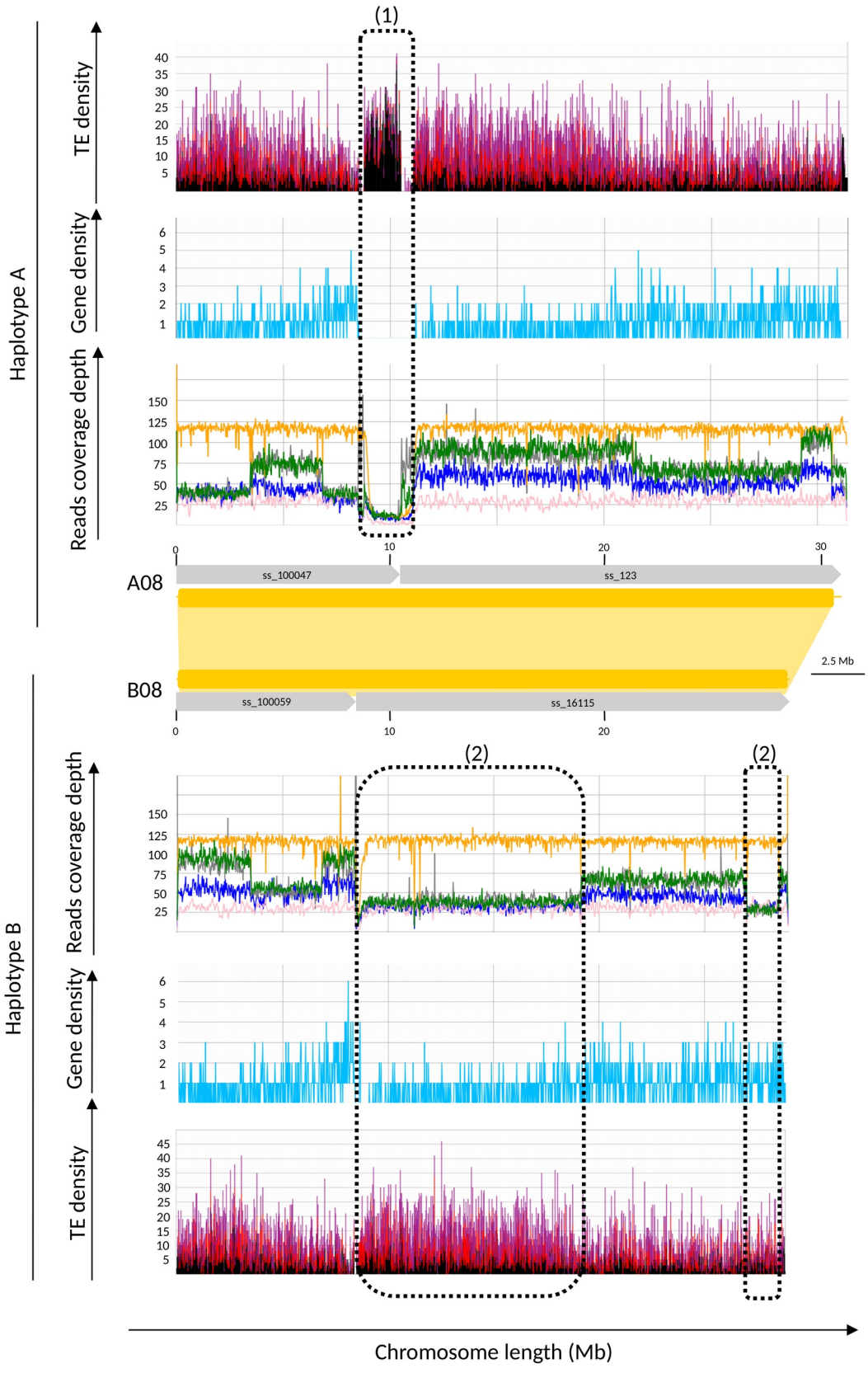
17

60 Supplemental Figure 9 continued

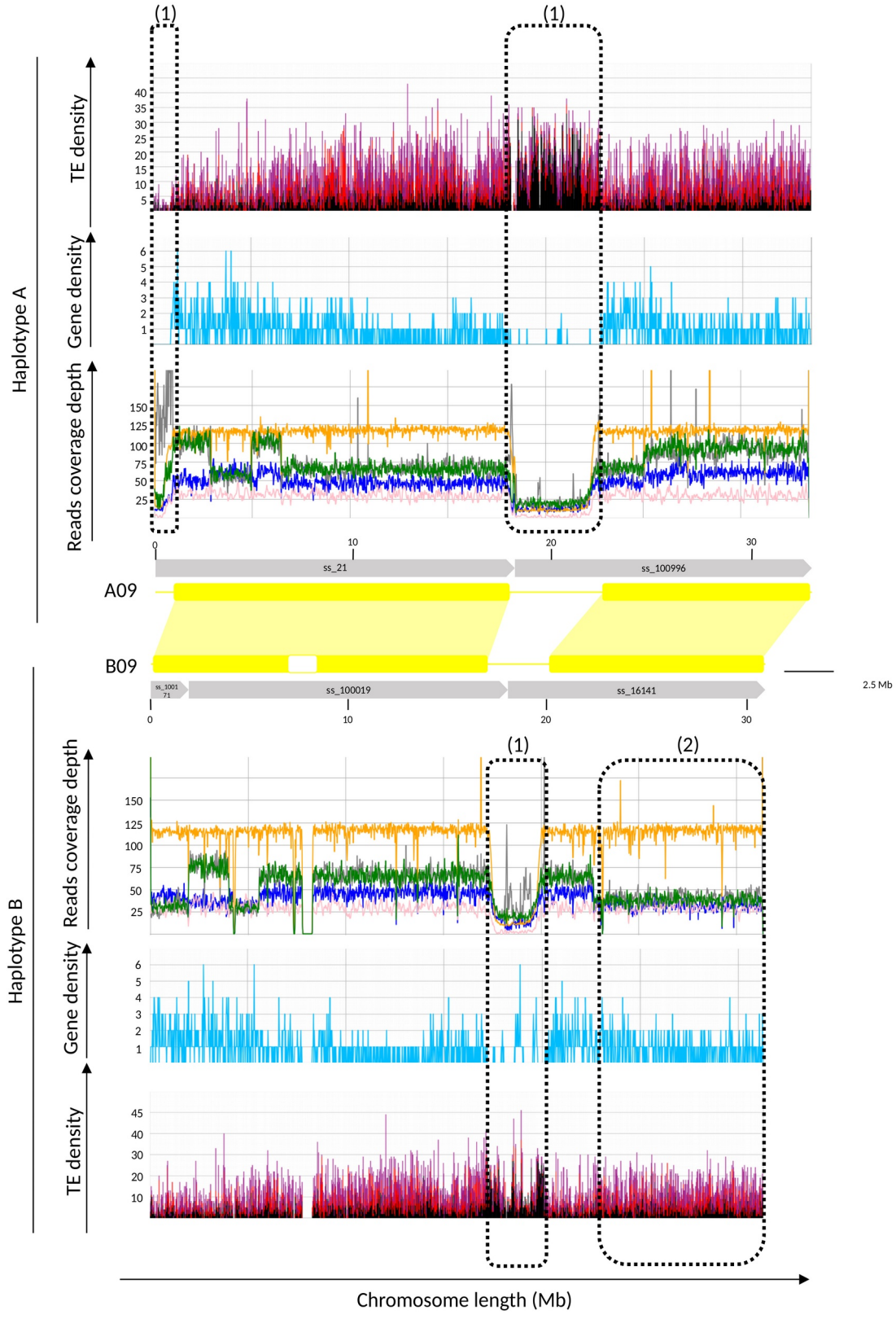


18

61 Supplemental Figure 9 continued

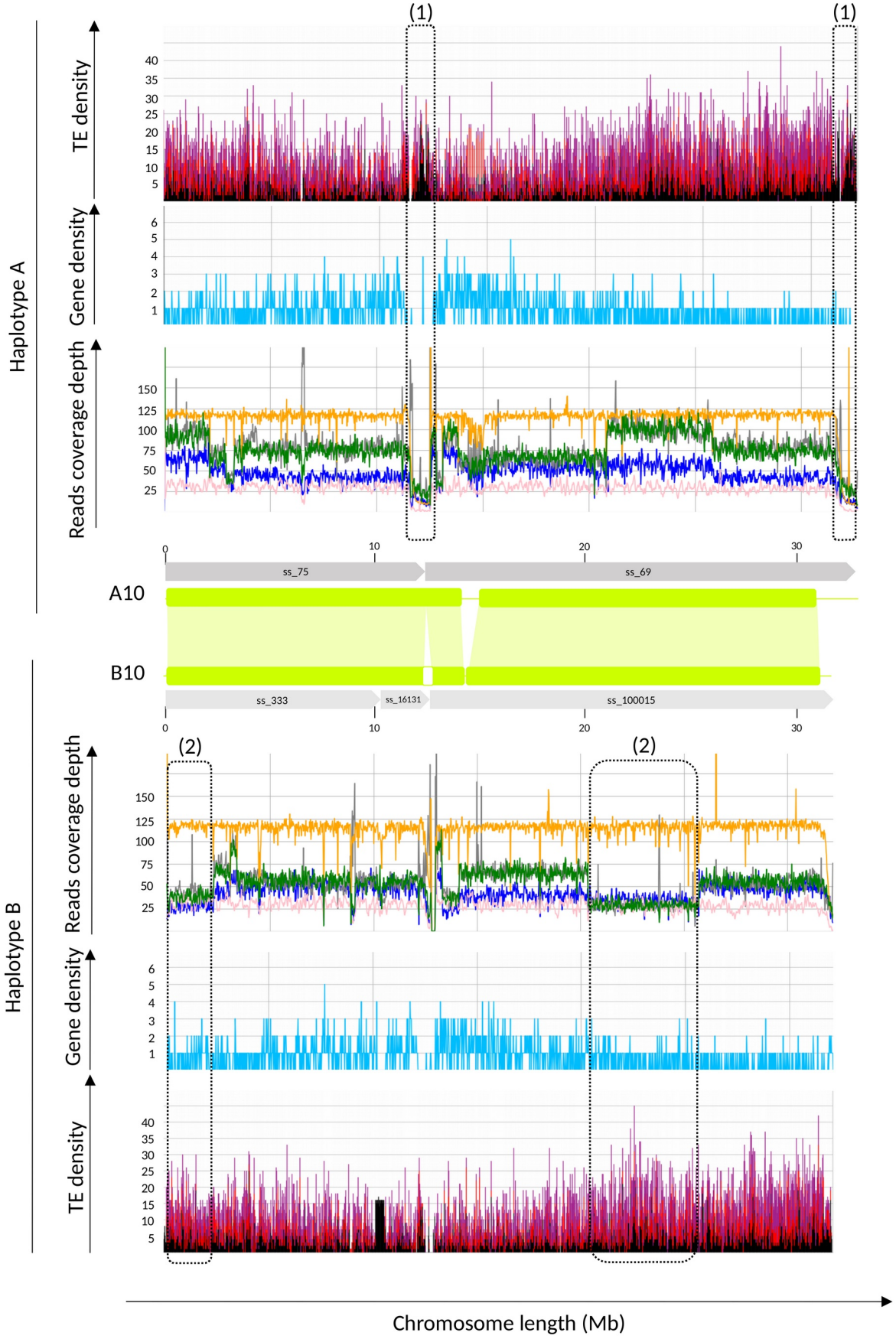


62 Supplemental Figure 9 continued



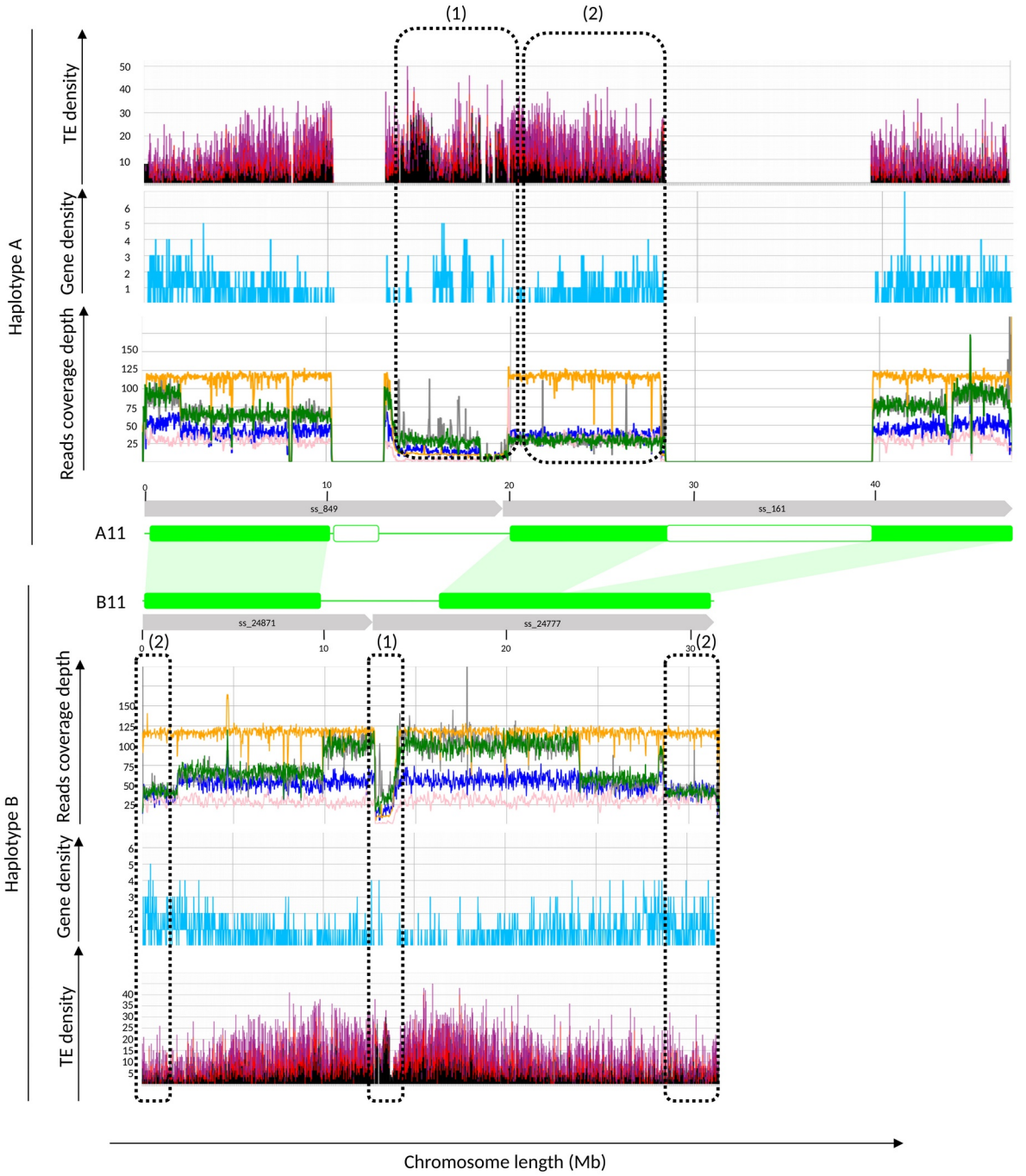
20

63 Supplemental Figure 9 continued



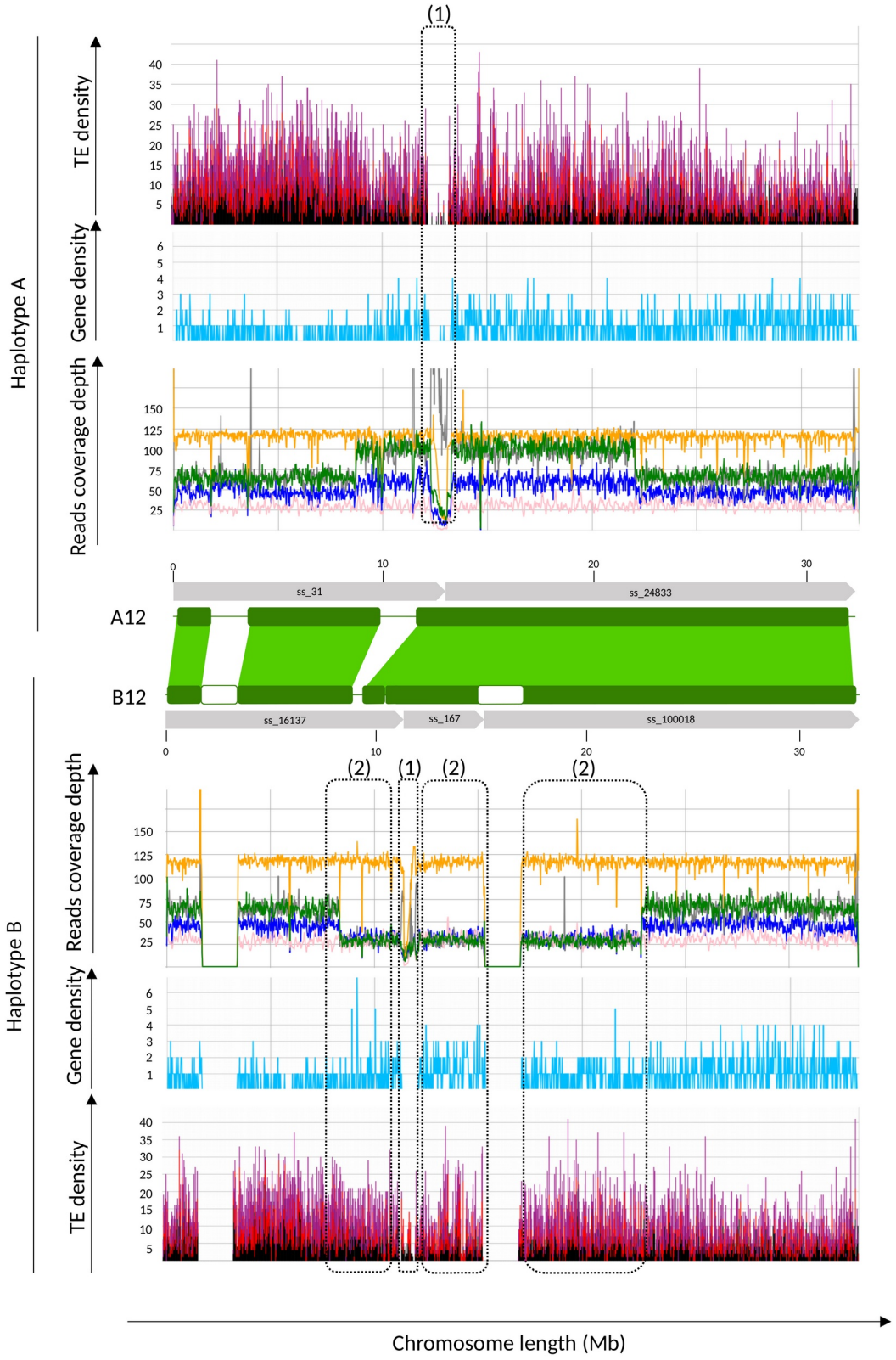
21

64 Supplemental Figure 9 continued

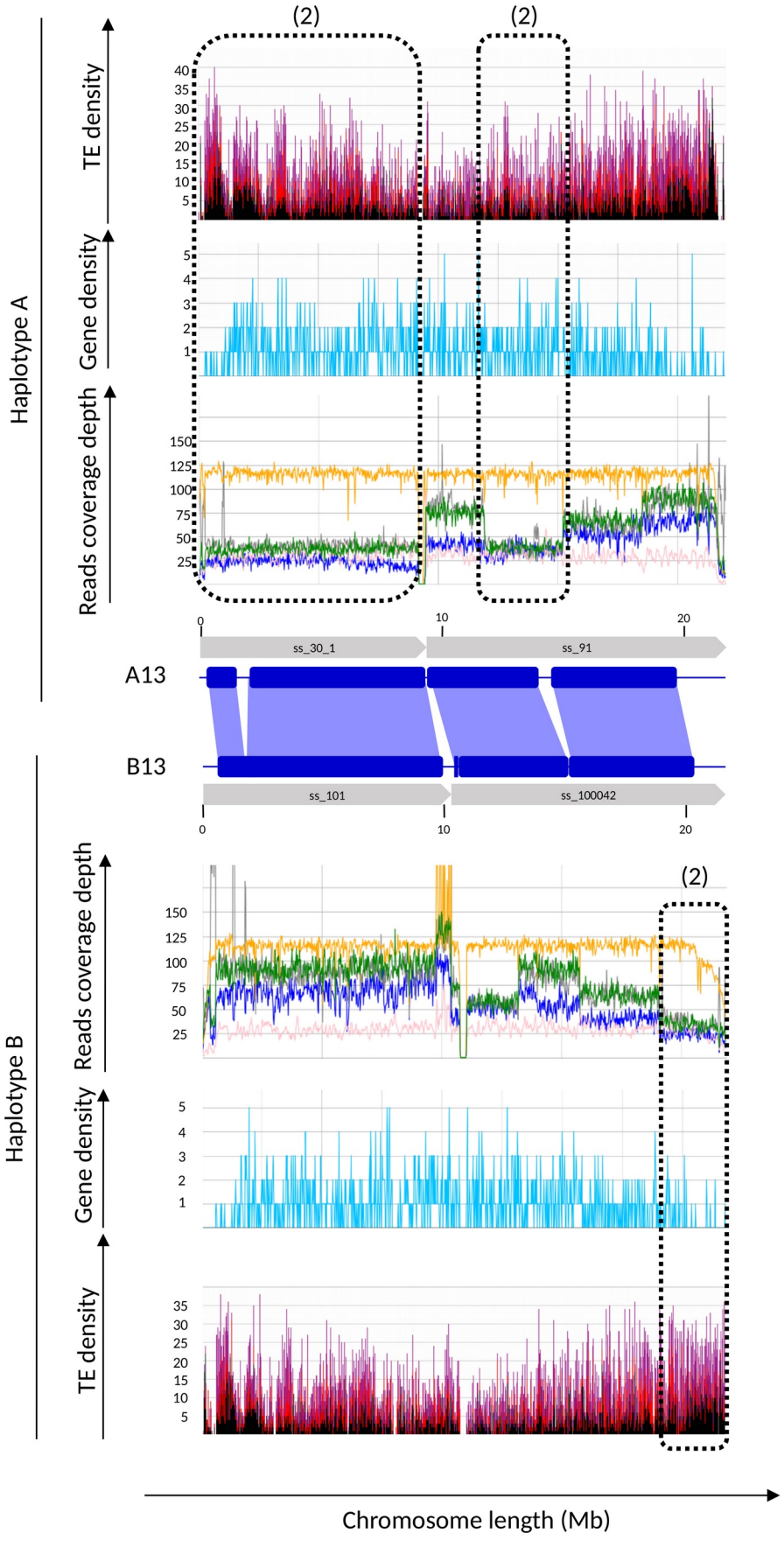


22

65 Supplemental Figure 9 continued

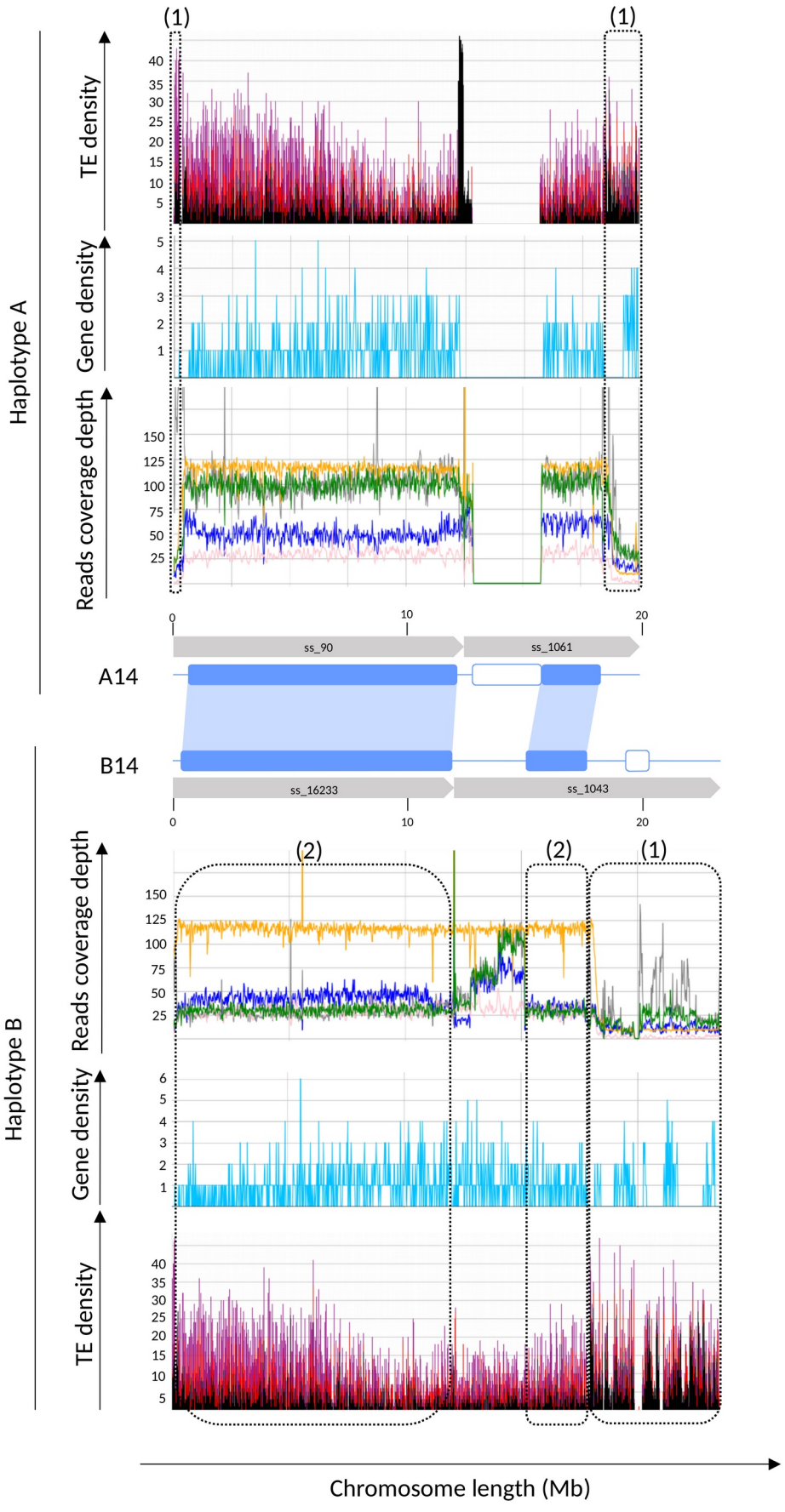


23

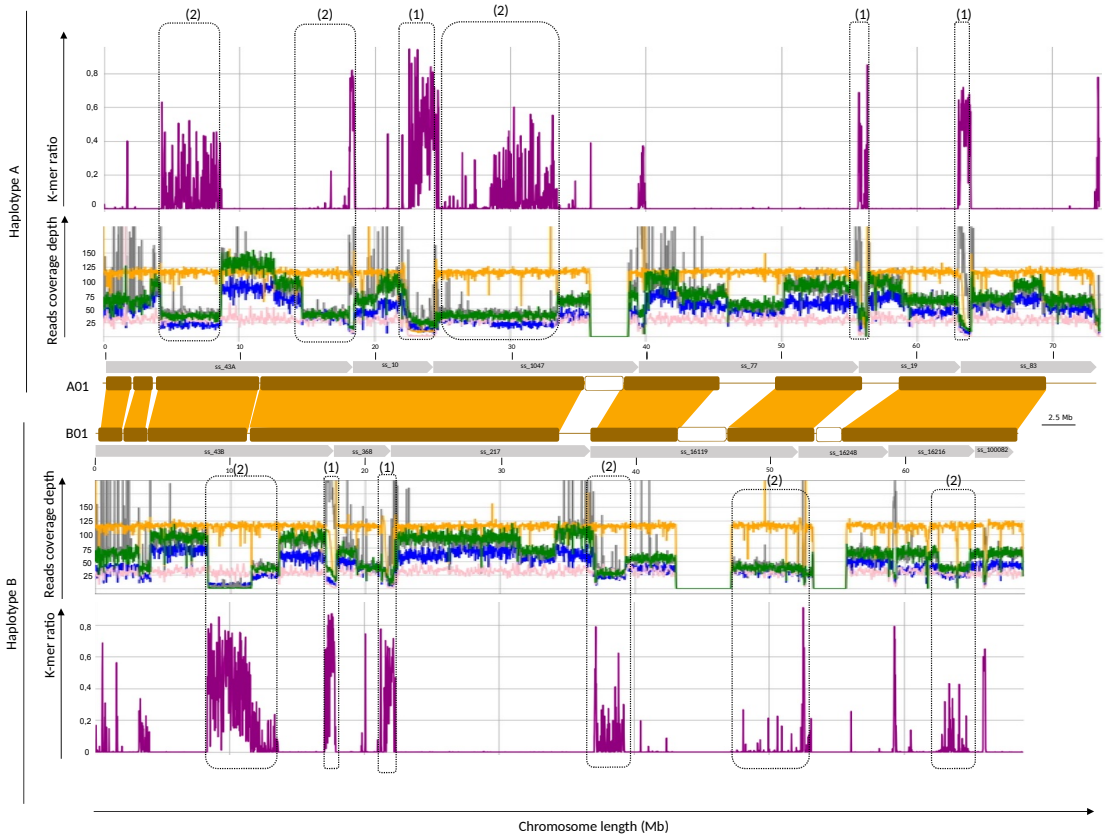

Haplotype A
TE density
Gene density
Reads coverage depth
(2)
(2)
ss_30_1
ss_91
A13
B13
ss_101
ss_100042
Haplotype B
Reads coverage depth
Gene density
TE density
(2)
Chromosome length (Mb)

24

67 Supplemental Figure 9 continued



25

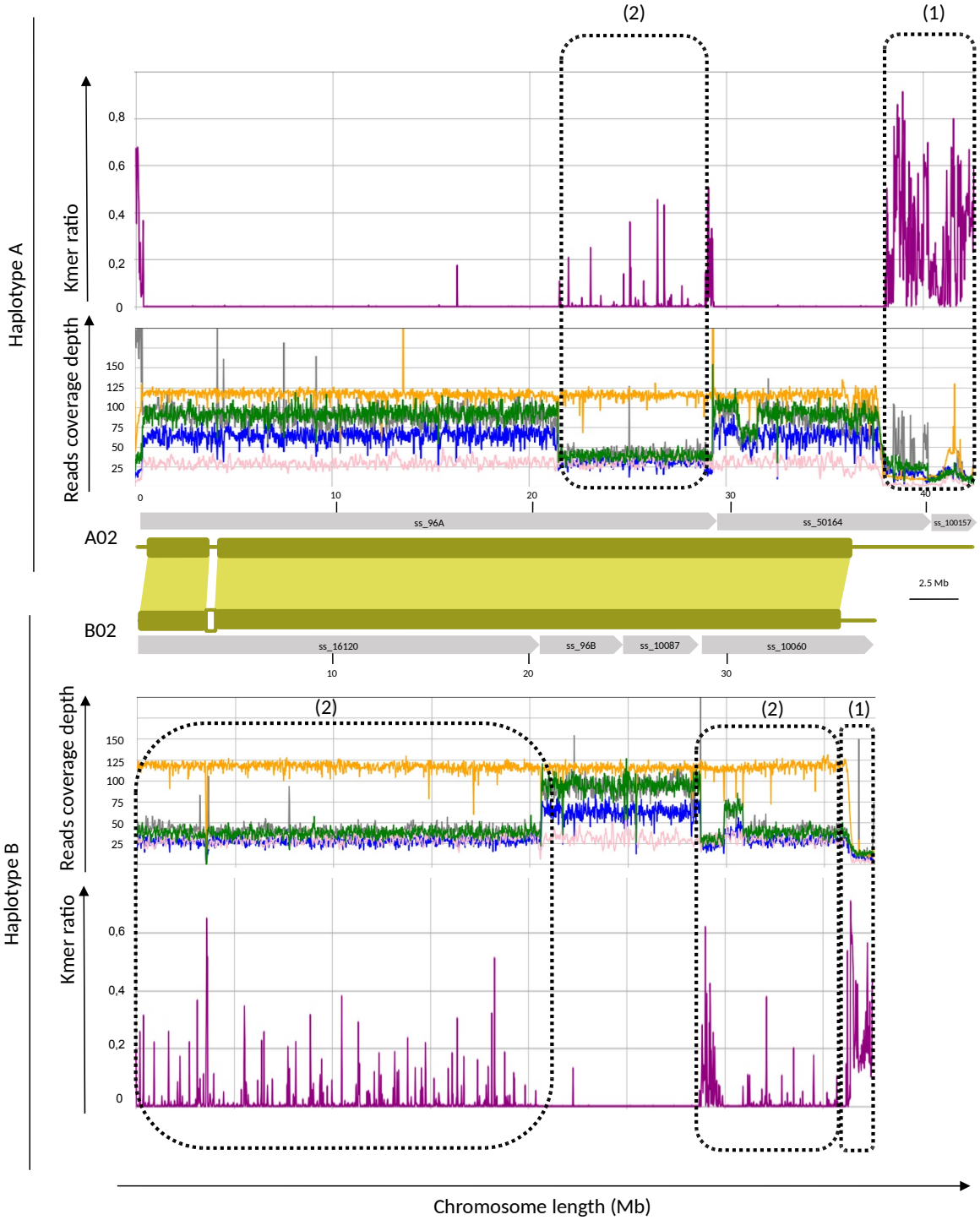


68 Supplemental Figure 10 Sequencing reads coverage depth and 5X-15X depth k-mers density
69 distribution along the 14 pairs of chromosomes

70 Obtained by mapping Daphna Illumina reads (yellow), CR0040 PacBio Hifi (blue), Nanopore
71 reads (green) and optical mapping data (red) on the CR0040 assembly. The ratio of k-mers
72 (47-mers) of depth between 5X and 15X are represented along each pair of chromosomes
73 (genomic windows of 20Kb). For other details see the legend of Supplemental Figure 9.

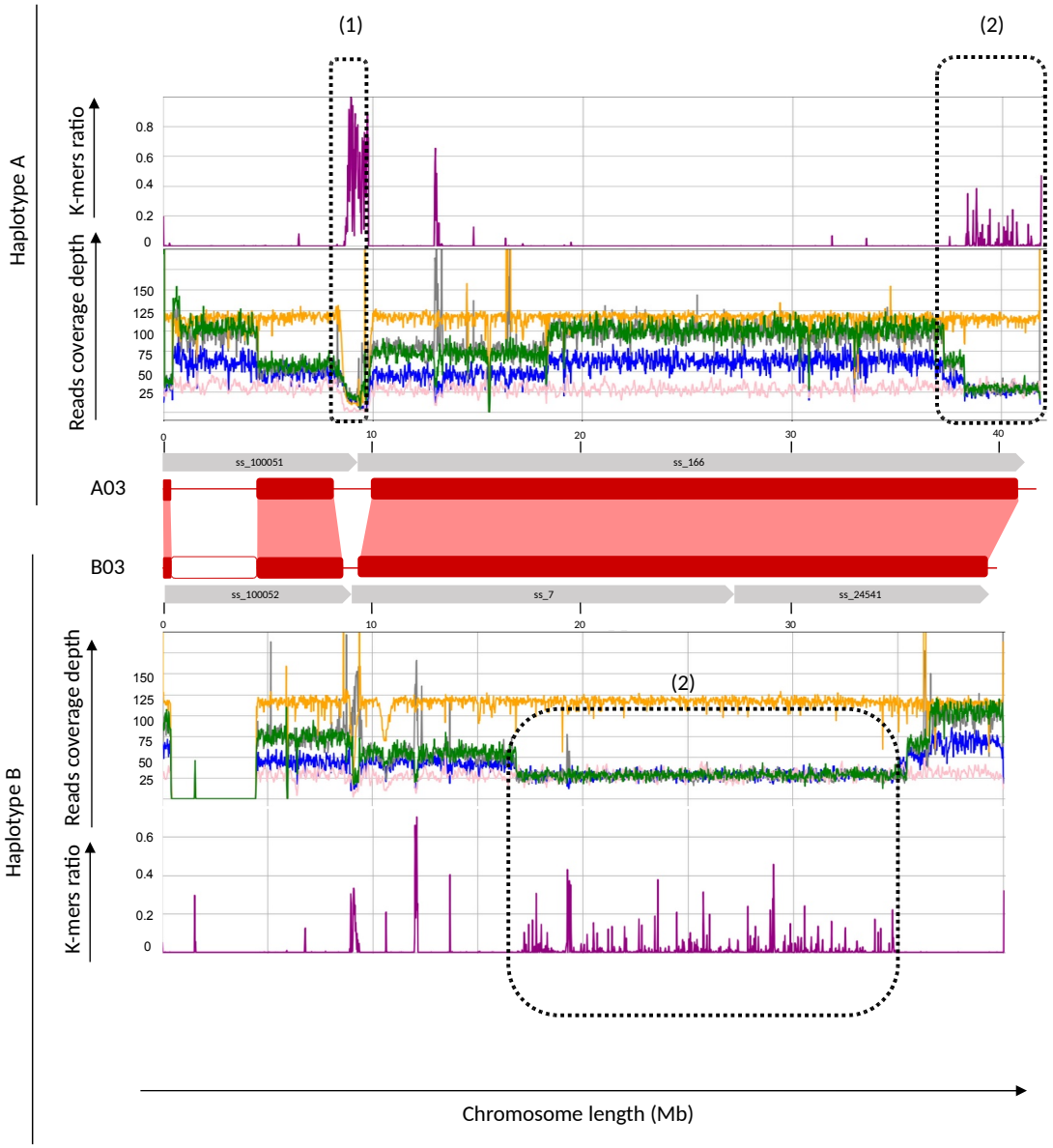
26

74 Supplemental Figure 10 continued



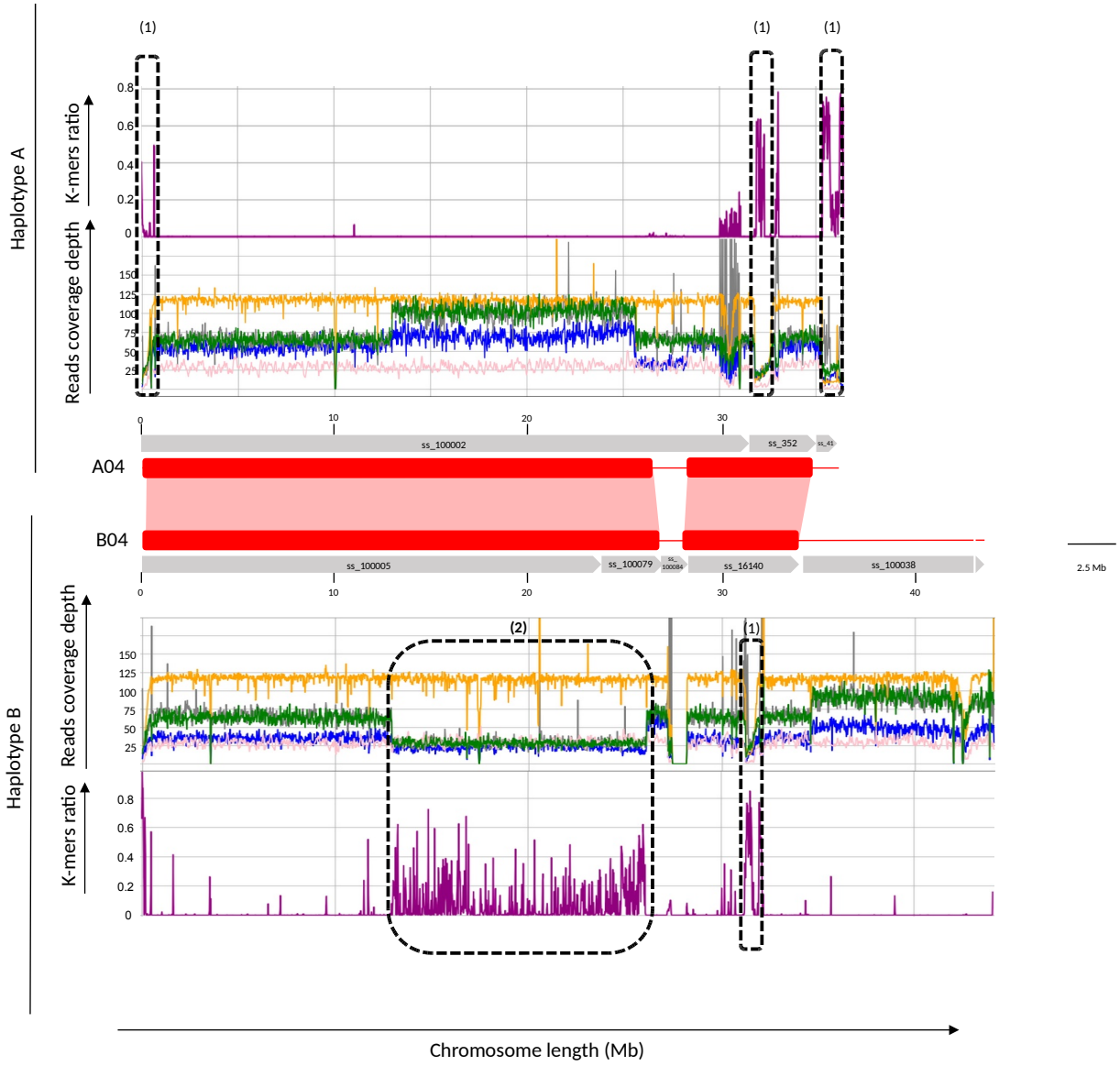
27

75 Supplemental Figure 10 continued



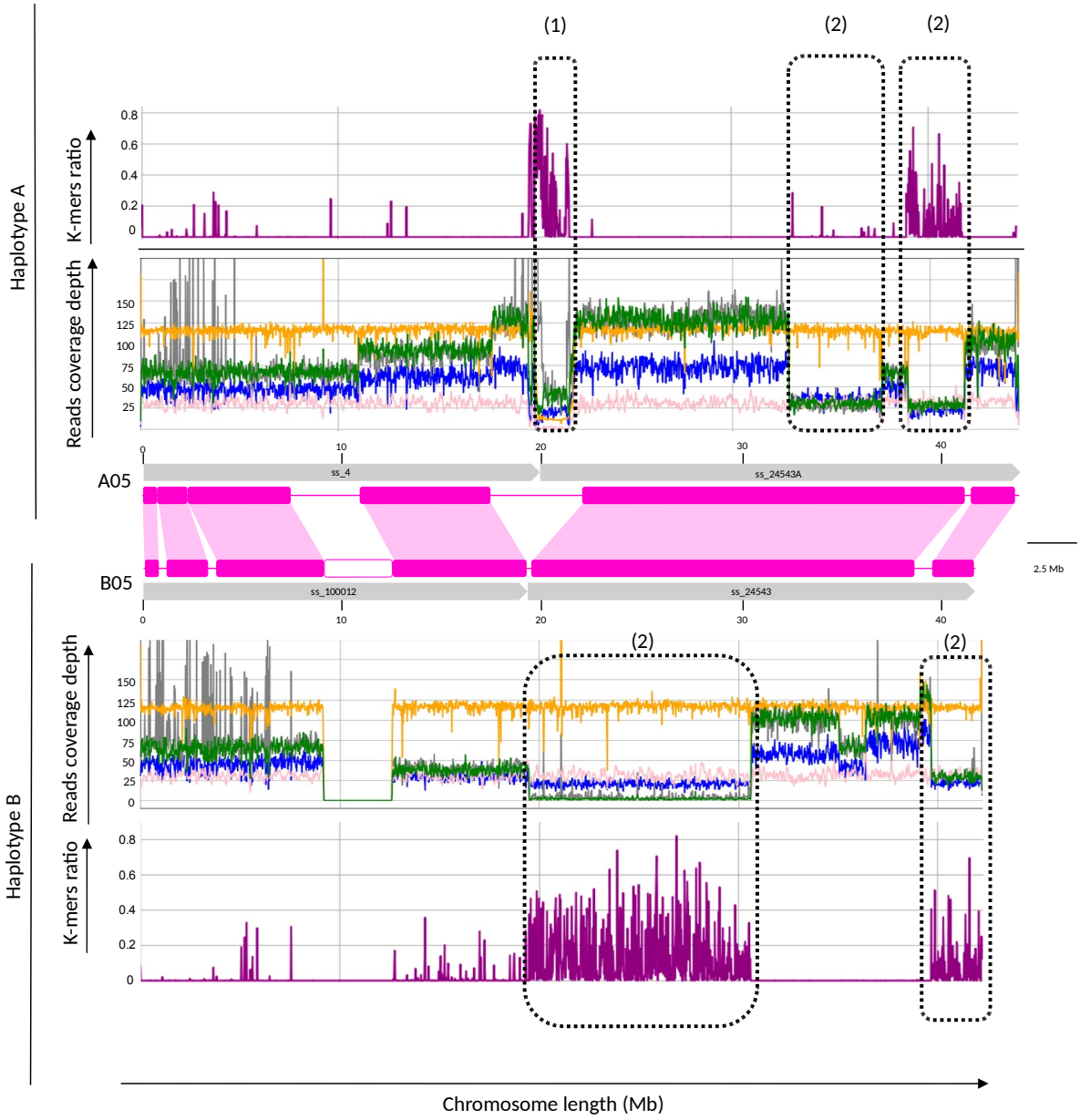
28

76 Supplemental Figure 10 continued



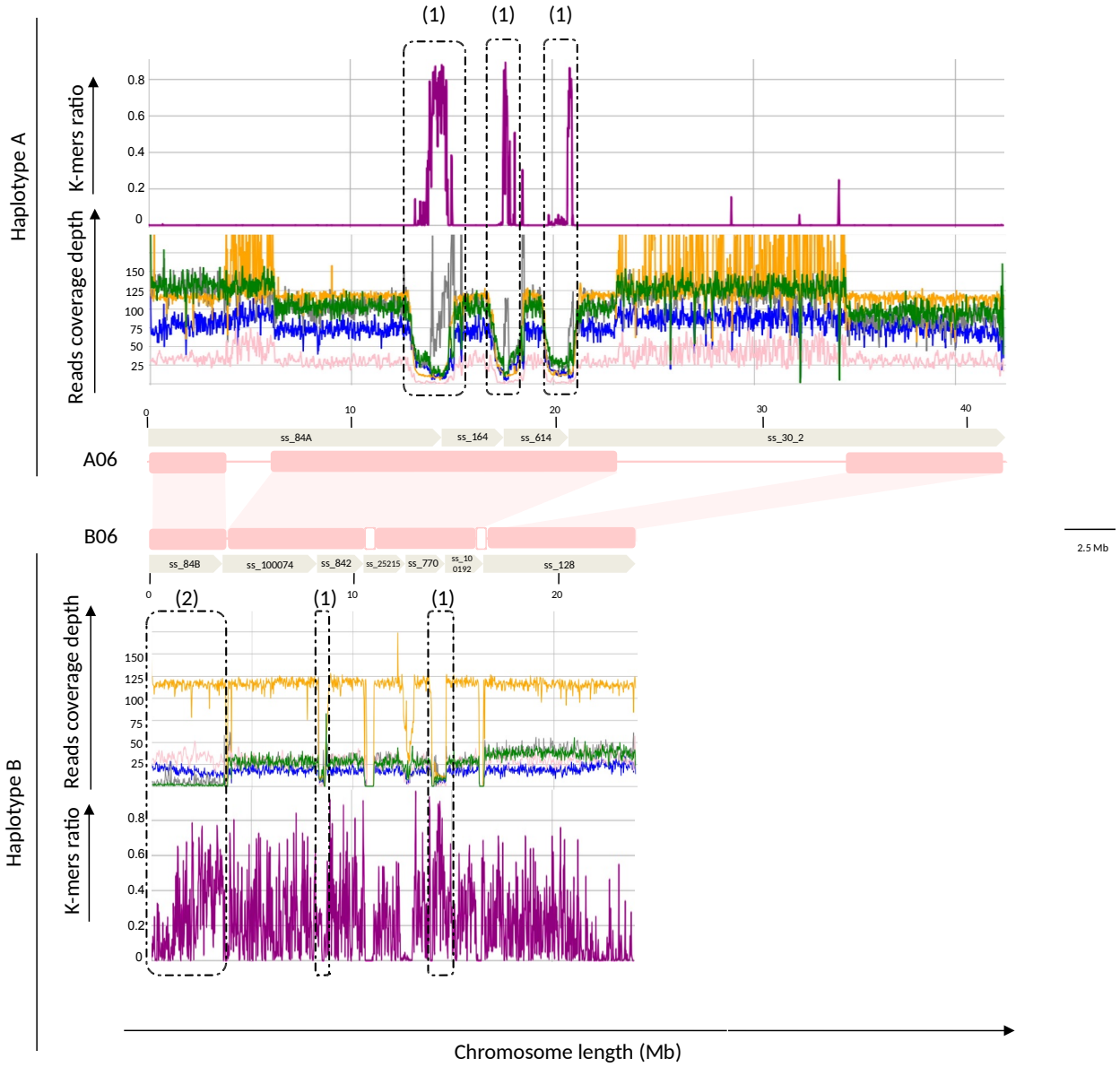
29

77 Supplemental Figure 10 continued



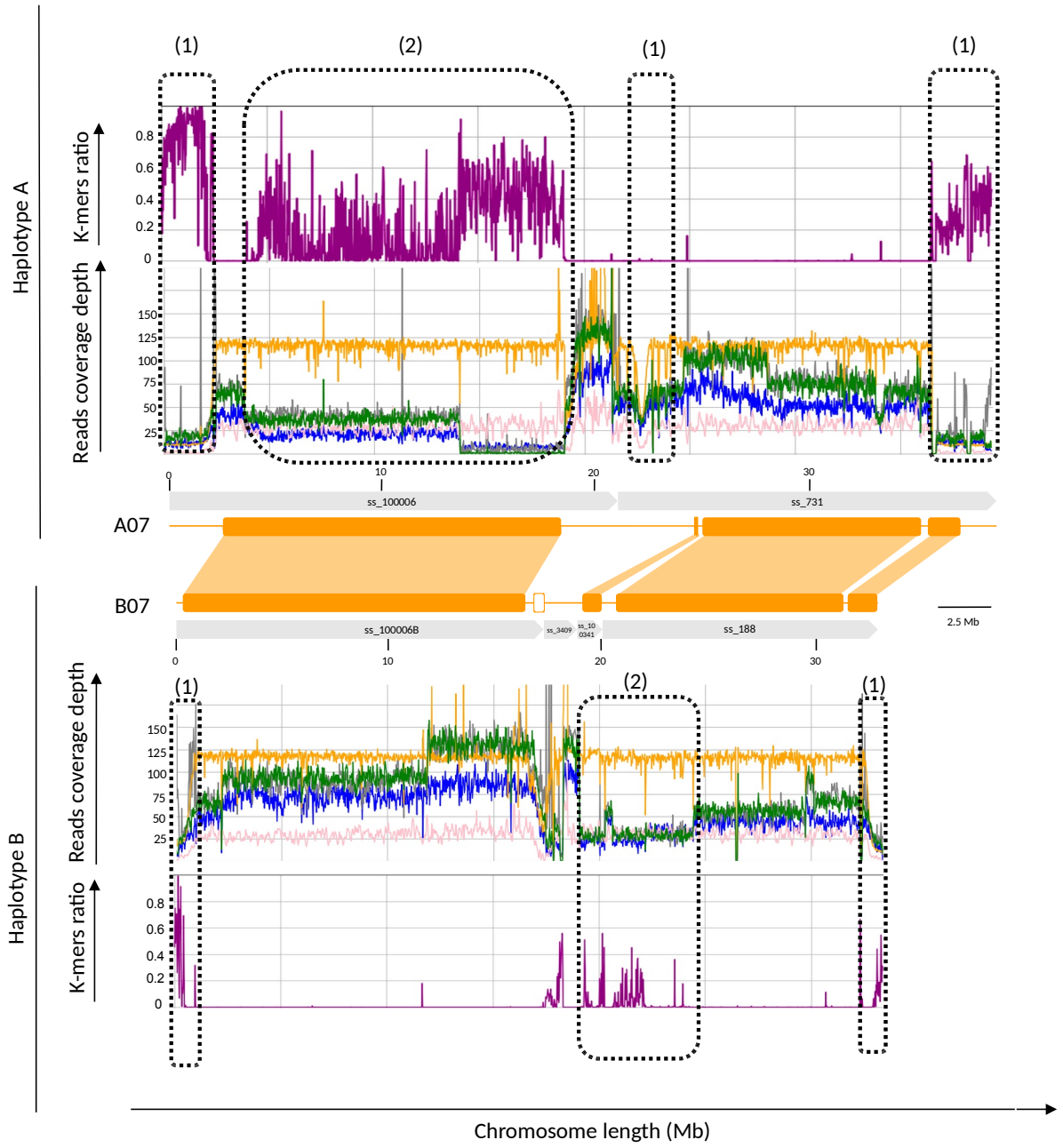
30

78 Supplemental Figure 10 continued



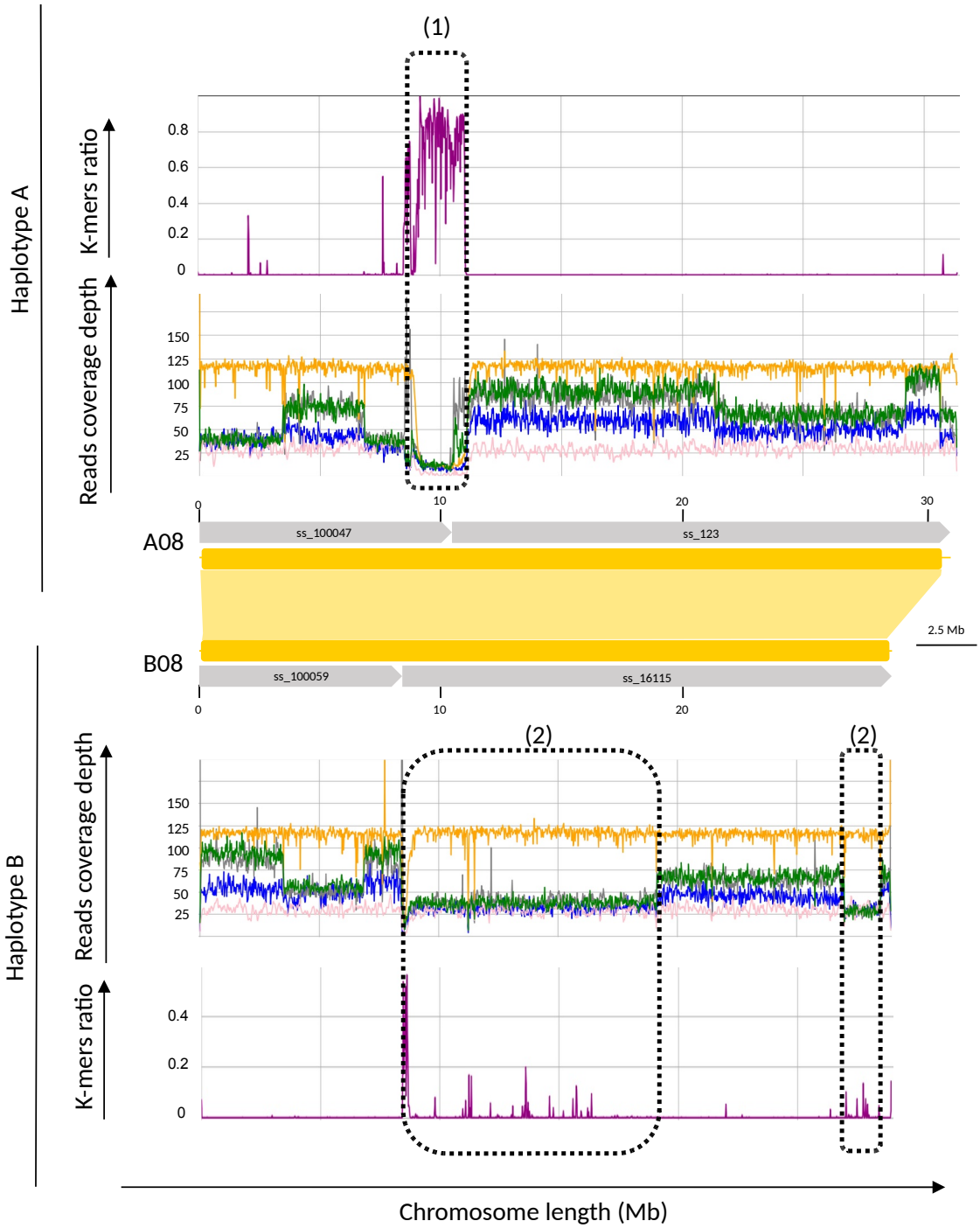
31

79 Supplemental Figure 10 continued



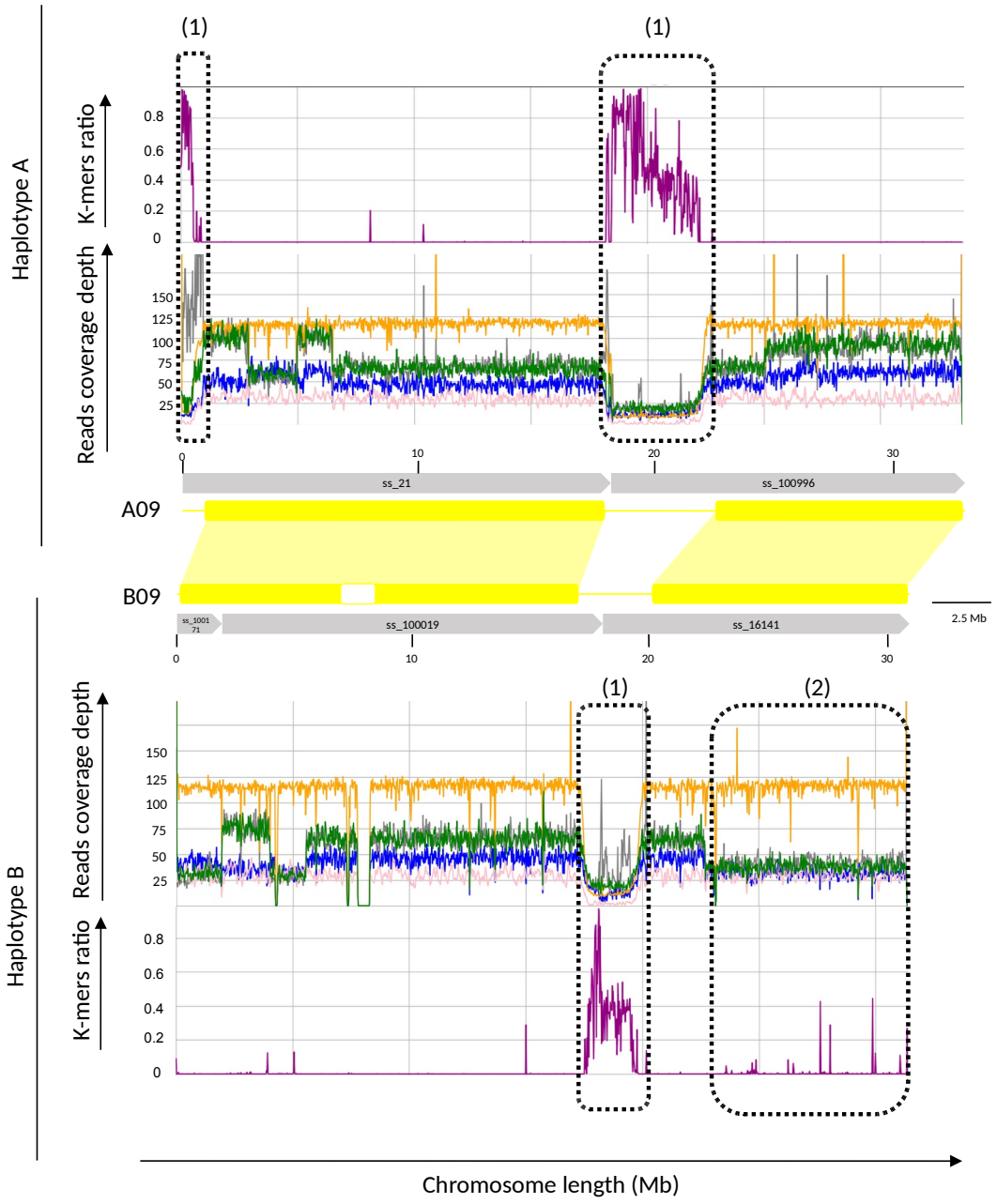
32

80 Supplemental Figure 10 continued



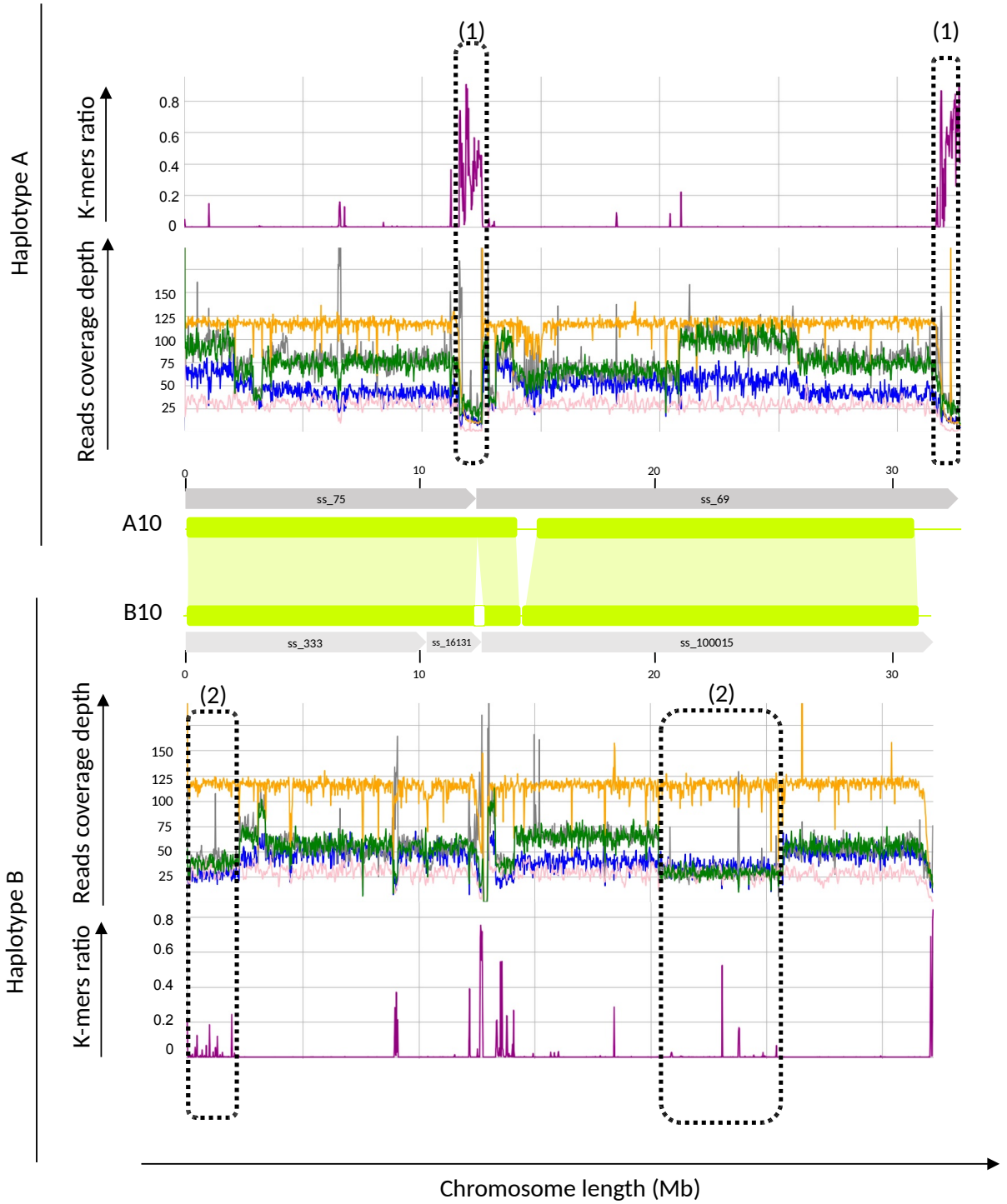
33

81 Supplemental Figure 10 continued



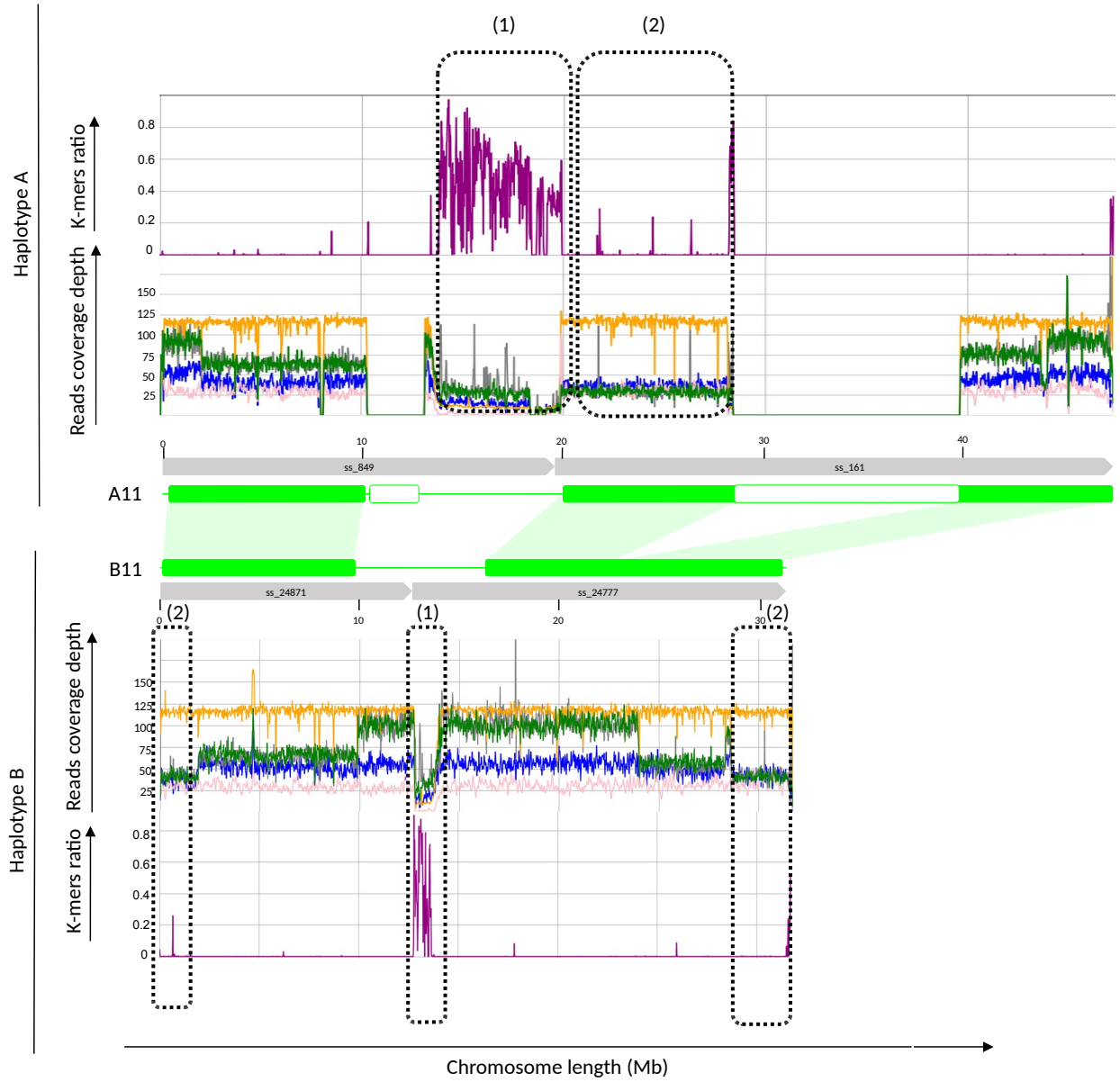
34

82 Supplemental Figure 10 continued



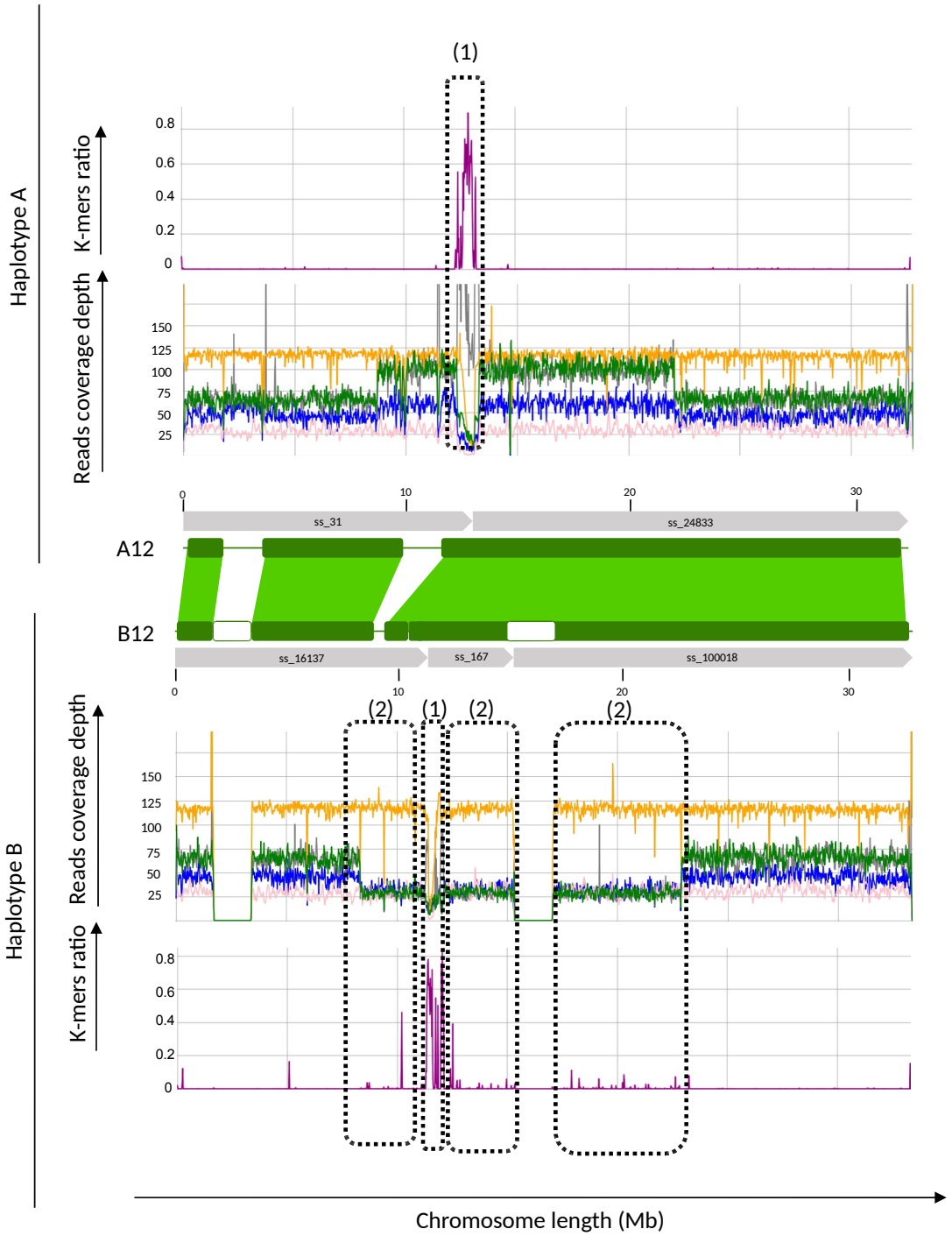
35

83 Supplemental Figure 10 continued



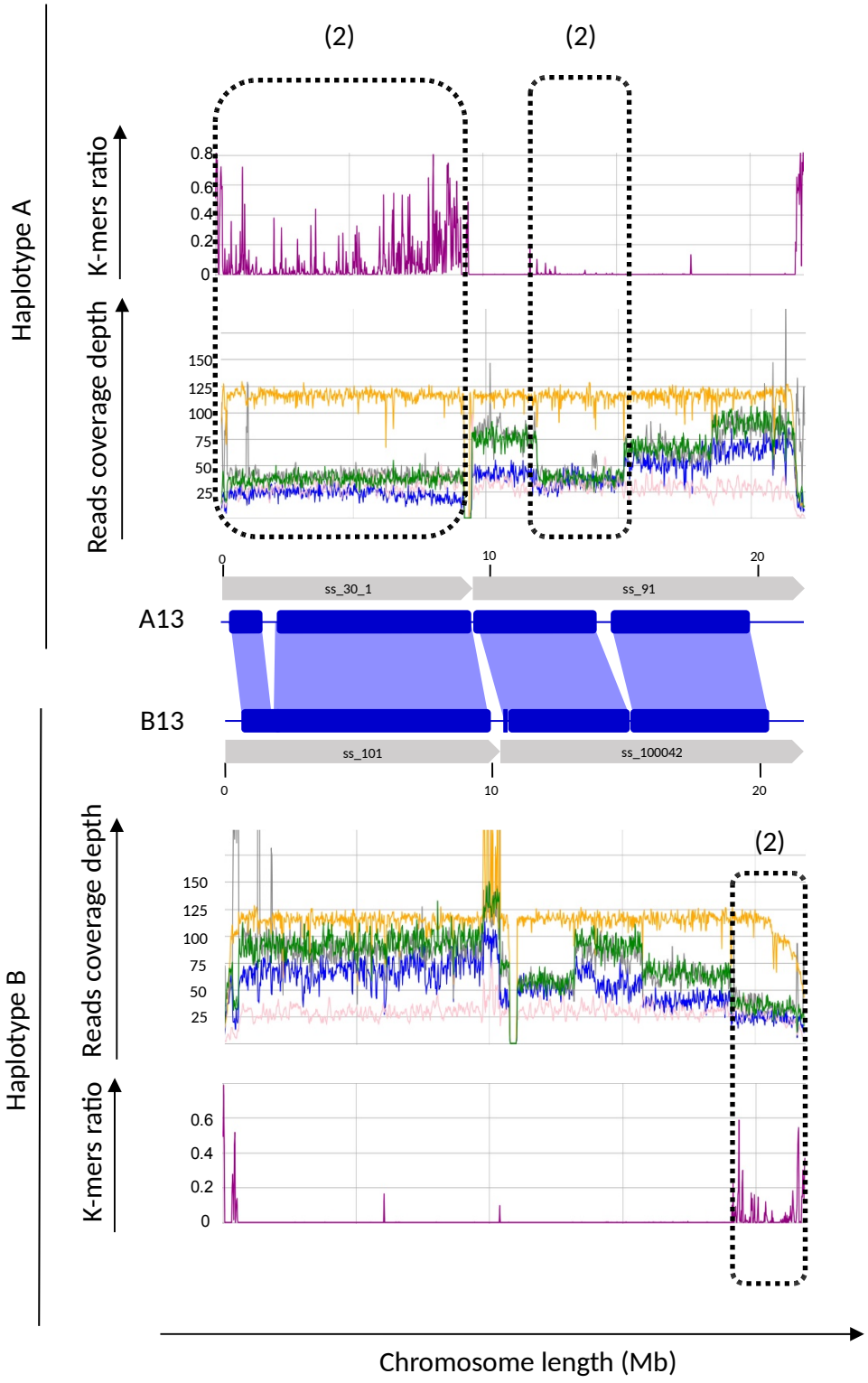
36

84 Supplemental Figure 10 continued



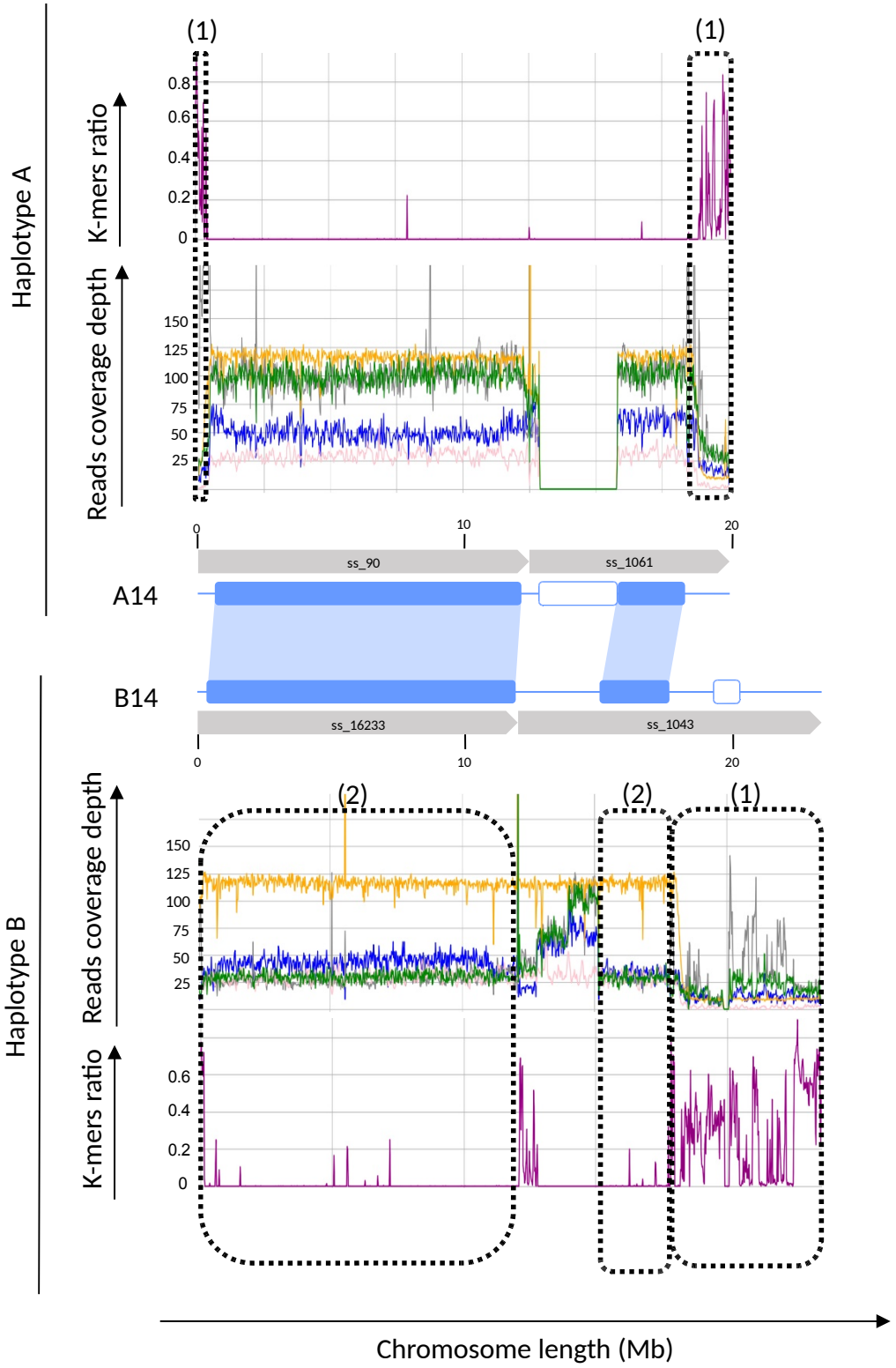
37

85 Supplemental Figure 10 continued

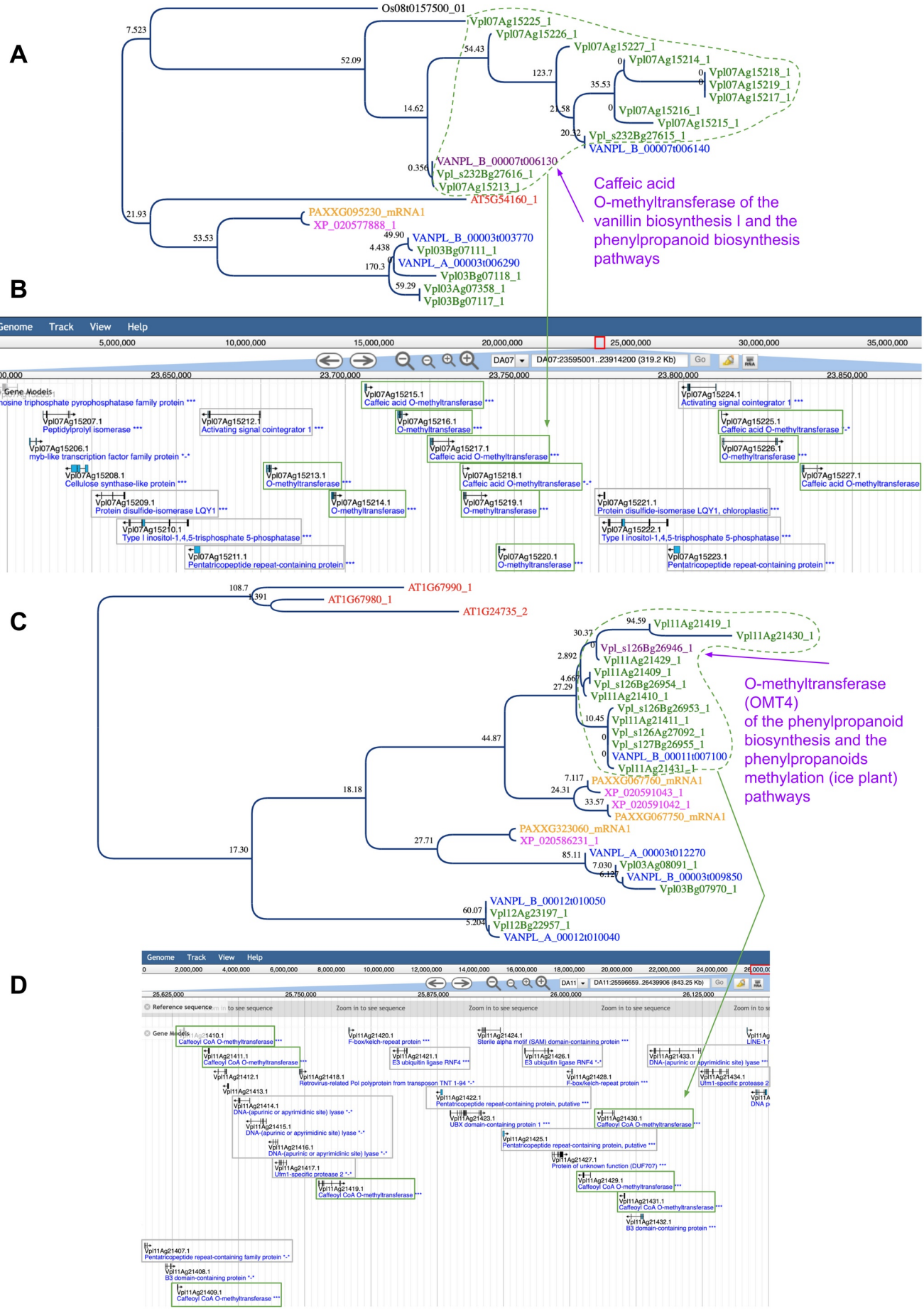


38

86 Supplemental Figure 10 continued



39


A
Os08t0157500_01
Vpl07Ag15225_1
Vpl07Ag15226_1
Vpl07Ag15227_1
Vpl07Ag15214_1
Vpl07Ag15218_1
Vpl07Ag15219_1
Vpl07Ag15217_1
Vpl07Ag15216_1
Vpl07Ag15215_1
Vpl_s232Bg27615_1
VANPL_B_00007t006140
VANPL_B_00007t006130
Vpl_s232Bg27616_1
Vpl07Ag15213_1
AT5G54160_1
PAXXG095230_mRNA1
XP_020577888_1
VANPL_B_00003t003770
Vpl03Bg07111_1
VANPL_A_00003t006290
Vpl03Bg07118_1
Vpl03Ag07358_1
Vpl03Bg07117_1
Caffeic acid O-methyltransferase of the vanillin biosynthesis I and the phenylpropanoid biosynthesis pathways
B
Genome
Track
View
Help
C
AT1G67990_1
AT1G67980_1
AT1G24735_2
Vpl11Ag21419_1
Vpl11Ag21430_1
Vpl_s126Bg26946_1
Vpl11Ag21429_1
Vpl11Ag21409_1
Vpl_s126Bg26954_1
Vpl11Ag21410_1
Vpl_s126Bg26953_1
Vpl11Ag21411_1
Vpl_s126Ag27092_1
Vpl_s127Bg26955_1
VANPL_B_00011t007100
Vpl11Ag21431_1
PAXXG067760_mRNA1
XP_020591043_1
XP_020591042_1
PAXXG067750_mRNA1
PAXXG323060_mRNA1
XP_020586231_1
VANPL_A_00003t012270
Vpl03Ag08091_1
VANPL_B_00003t009850
Vpl03Bg07970_1
VANPL_B_00012t010050
Vpl12Ag23197_1
Vpl12Bg22957_1
VANPL_A_00012t010040
O-methyltransferase (OMT4) of the phenylpropanoid biosynthesis and the phenylpropanoids methylation (ice plant) pathways
D
Genome
Track
View
Help

40

Supplemental Figure 11 Caffeic acid O-methyltransferase cluster comparison between CR0040 and Daphna. Phylogenetic trees and genomic positions of clusters of genes encoding for Caffeic acid O-methyltransferase (A, B; OG0000841; e.g. VANPL_B_00007t006130) and for O-methyltransferase (C, D; OG0000449; e.g. VANPL_B_00011t00710 Vpl_s126Bg26946.1). JBrowse screenshots were done from https://vanilla-genome-hub.cirad.fr/content/v-planifolia-daphna, DA07:23595079..23914278 (319.2 Kb) and DA11:25632407..26200506 (568.1 Kb).

41

Supplemental tables

Supplemental Table 1 Nuclear classes in flow cytometry samples from *Vanilla planifolia* CR0040 nodal tissues

Nuclear populations	2C	4E	8E	16E	32E
Proportion Frequency (%, mean (sd) of 13 independent measures)	17.31 (10.08)	18.52 (7.21)	33.95 (11.41)	21.5 (7.84)	8.72 (12.73)
Mean (sd) of previous interpeak ratio* (Brown et al., 2017)		1.28 (0.01)	1.44 (0.02)	1.62 (0.03)	1.75 (0.02)
Mean (sd) of interpeak ratio*		1.30 (0.03)	1.42 (0.03)	1.61 (0.02)	1.73 (0.08)

* Interpeak Ratio is the fluorescence intensity (I, arbitrary units) of peak n to peak (n — 1).

sd, standard deviation.

Supplemental Table 2 Main statistics of sequencing data production

Technology	Instrument	Sample type	Number of sequences	Raw data (Gb)	N50 (Kb)
Illumina	HiSeq3000	DNA 10X genomics	1,363,939,362	205	0.15
ONT	GridION	DNA	2,855,136	16	13
ONT	PromethION	DNA	18,044,495	131	18
Pacific BioSciences	Sequel II	DNA	15,306,231	1,178	15.2*
Illumina	NovaSeq6000	mRNA	857,606,502	129	0.15

Gb, giga bases ; Kb, kilo bases.

*N50 for 69.75 Gb of HiFi reads

2

Supplemental Table 3 Main metrics of CR0040 genome assemblies

	10X	ONT	PacBio HiFi	Meta-assembly
Input data	200 Gb	150 Gb	69 Gb	
Estimated coverage	100X	75X	35X	
Cumulative size (Gb)	1.68	1.9	3.4	3.3
Contig number	254,400	21,443	24,534	180,957
N50 (kb)	25	115	924	133
Max size (Mb)	9.6	23.4	31	28.7

Gb, giga bases ; Kb, kilo bases ; Mb, mega bases.

11 Supplemental Table 4 Main metrics for the best genome assembly

	All contigs	Hybrid scaffolds	Contigs not scaffolded	Haplotype A	Haplotype B
Cumulative size (Gb)	3.4	1.9	1.5	1.5	1.9
Numbers	24,534	874	23,037	3,874	17,655
N50 (Mb)	0.92	6.2	0.09	3	0,4
Max Size (Mb)	31	32	10	73	69
N's (Mb)	0	79	0	37	40

12 Gb, giga bases ; Mb, mega bases.

13 Supplemental Table 5 Genic completeness for CR0040 genome

	Haplotype A	Haplotype B	Haplotypes A+B
Genes in Viridiplantae database	425	425	425
Complete single-copy	379 (89.2%)	367 (86.4%)	17 (4%)
Complete duplicated	17 (4.0%)	17 (4%)	395 (92.9%)
Fragment	1 (0.2%)	27 (6.4%)	1 (0.2%)
Missing	28 (6.6%)	14 (3.3%)	12 (2.9%)
Genes in Embryophyta database	1,614	1,614	1,614
Complete single-copy	1,436 (89%)	1,410 (87.4%)	182 (11.3%)
Complete duplicated	53 (3.3%)	86 (5.3%)	1,367 (84.7%)
Fragment	13 (0.8%)	20 (1.2%)	13 (0.8%)
Missing	112 (6.9%)	98 (6.1%)	52 (3.2%)
Genes in Liliopsida database	3,236	3,236	3,236
Complete single-copy	2,639 (81.6%)	2,589 (80%)	402 (12.4%)
Complete duplicated	111 (3.4%)	174 (5.4%)	2,475 (76.5%)
Fragment	223 (6.9%)	249 (7.7%)	230 (7.1%)
Missing	263 (8.1%)	224 (6.9%)	129 (4%)

14 Supplemental Table 6 Main characteristics of RNA samples

BioSample accession	Tissue	Nanodrop concentration (ng/μL)	RNA Integrity Number
SAMN20691786	Fruit	32.02	8.7
SAMN20691787	Leaf	119.00	8.1
SAMN20691788	Flower	121.50	8.7
SAMN20691789	Stem	47.55	8.8
SAMN20691790	Soil root	179.40	9.1
SAMN20691791	Aerial root	98.32	8.0
SAMN20691792	Bud	326.80	9.1
SAMN20691793	Flower bud	112.70	9.4
SAMN20691794	Ovary	91.66	9.0
SAMN20691795	Mix	57.28	8.5

Supplemental Table 7 Main statistics of RNA-seq experiment

	Number of reads	Alignment rate (%)	Number of transcripts	Mean transcript size (bases)
Bud	47,018,576	93.84	90,813	2,049
Flower bud	49,472,733	92.87	86,515	1,971
Aerial roots	40,920,676	82.19	81,970	1,905
Leaves	40,129,580	93.55	79,462	1,959
Flowers	33,920,088	92.68	75,827	1,966
Fruits	32,941,273	89.28	72,767	1,903
Ovary	43,083,064	93.53	83,388	2,101
Soil root	43,969,059	90.39	85,318	2,063
Stem	43,692,744	92.60	85,396	1,932
Mixed tissues	53,655,458	90.12	92,217	1,989

16 Supplemental Table 8 Genic completeness for transcriptomes (BUSCO)

	Complete single-copy	Complete duplicated	Fragment	Missing
Bud	5 (1.2%)	409 (96.2%)	1 (0.2%)	10 (2.4%)
Flower bud	4 (0.9%)	410 (96.5%)	1 (0.2%)	10 (2.4%)
Aerial roots	14 (3.3%)	392 (92.2%)	5 (1.2%)	14 (3.3%)
Leaves	7 (1.6%)	404 (95.1%)	1 (0.2%)	13 (3.1%)
Flowers	29 (6.8%)	362 (85.2%)	15 (3.5%)	19 (4.5%)
Fruits	37 (8.7%)	348 (81.9%)	15 (3.5%)	25 (5.9%)
Ovary	11 (2.6%)	400 (94.1%)	3 (0.7%)	11 (2.6%)
Soil root	9 (2.1%)	402 (94.6%)	2 (0.5%)	12 (2.8%)
Stem	7 (1.6%)	405 (95.3%)	2 (0.5%)	11 (2.6%)
Mixed tissues	7 (1.6%)	407 (95.8%)	1 (0.2%)	10 (2.4%)
Merged transcriptomes	2 (0.5%)	412 (96.9%)	1 (0.2%)	10 (2.4%)

Supplemental Table 9 Summary statistics of protein coding genes per haplotype

	Haplotype A	Haplotype B
Number of protein coding genes	26,392	32,736
Number of exons	140,181	158,279
Mean gene length (bp)	7,865	6,692
Mean exon length (bp)	269	269
Mean number of exons per mRNA	5.3	4.8
Number of single exon gene	5,097	7,460
Mean coding sequence length (bp)	1,072	995
Mean protein length (aa)	356	331
Orthologs 1 to 1	15,844	15,844
Orthologs 1 to many	1,335	1,999
Orthologs many to 1	6,052	4,721
Orthologs many to many	847	704
Haplotype A specific orthologs	742	-
Haplotype B specific orthologs	-	4,724
Orthogroups unassigned genes	985	4,515

bp, base pairs ; aa, amino acid.

9

19 Supplemental Table 10 Genic completeness for CR0040 annotation with BUSCO

BUSCO 5.0.0	Haplotype A	Haplotype B	Haplotype A+B
	26392	32736	59128
Genes in Viridiplantae database	425	425	425
Complete single-copy	343 (80.7%)	337 (79.3%)	65 (15.3%)
Complete duplicated	18 (4.2%)	27 (6.4%)	331 (77.9%)
Fragment	26 (6.1%)	29 (6.8%)	13 (3.1%)
Missing	38 (9.0%)	32 (7.5%)	16 (3.7%)
Genes in Embryophyta database	1,614	1,614	1,614
Complete single-copy	1,292 (80.0%)	1,271 (78.7%)	231 (14.3%)
Complete duplicated	56 (3.5%)	95 (5.9%)	1,245 (77.1%)
Fragment	88 (5.5%)	79 (4.9%)	54 (3.3%)
Missing	178 (11%)	169 (10.5%)	84 (5.3%)
Genes in Liliopsida database	3,236	3,236	3,236
Complete single-copy	2,383 (73.6%)	2,354 (72.7%)	539 (16.7%)
Complete duplicated	106 (3.3%)	187 (5.8%)	2,270 (70.1%)
Fragment	313 (9.7%)	310 (9.6%)	222 (6.9%)
Missing	434 (13.4%)	385 (11.9%)	205 (6.3%)

Supplemental Table 11 RepeatMasker summary report

	Haplotype A (nb)	Haplotype B (nb)	Haplotype A (Mb)	Haplotype B (Mb)	Haplotype A (%bp)	Haplotype B (%bp)	Haplotype A+B (%bp)
Total interspersed repeats	914,993	1,270,877	659.98	930.76	46.6	47.3	47.0
Total retroelements	690,136	992,412	577.09	827.14	40.7	42.0	41.5
Total LTR	378,082	563,026	229.77	344.86	16.2	17.5	17.0
LTR/Copia	134,148	173,060	89.02	118.81	6.3	6.0	6.1
LTR/Gypsi	209,894	346,287	123.85	203.4	8.7	10.3	9.7
LINES	136,139	170,984	168.91	236.34	11.9	12.0	12.0
Unclassified retroelements	170,906	252,117	177.45	244.74	12.5	12.4	12.5
SINE	5,009	6,285	0.96	1.2	0.1	0.1	0.1
DNA transposons	64,797	79,501	21.8	26.44	1.5	1.4	1.4
Unclassified repeats	159,752	198,616	61.03	77.12	4.3	3.9	4.1
Simple repeats	306,532	477,099	191.41	330.64	13.5	16.8	15.4
Low complexity regions	58,671	77,463	129.99	197.09	9.18	10.0	9.7

nb, number ; Mb, mega bases ; bp, base pairs.

Supplemental Table 12 Number and repartition of protein coding genes and repeats per chromosome and per haplotype of CR0040

	CR0040 Haplotype A					CR0040 Haplotype B				
Chr	w N	wo N	genes	cds wo N (%bp)	TEs wo N (%bp)	w N	wo N	genes	cds wo N (%bp)	TEs wo N (%bp)
1	73.50	70.63	3,532	6.12	30.16	68.78	62.17	3,293	6.55	27.95
2	42.59	42.57	1,426	4.14	39.09	37.53	37.47	1,375	4.73	35.03
3	41.87	41.79	1,341	3.95	43.63	40.03	35.97	1,095	3.72	43.54
4	36.44	36.24	1,179	4.05	40.06	44.10	43.14	1,429	4.12	40.53
5	44.52	44.52	1,907	5.36	31.34	42.22	38.80	1,777	5.66	30.64
6	42.39	42.39	1,346	3.98	36.93	24.06	23.18	933	5.08	35.12
7	39.34	38.93	1,455	4.50	40.82	33.37	33.15	1,307	4.85	38.37
8	31.38	31.38	1,229	4.89	35.05	28.94	28.93	1,225	5.26	33.21
9	33.51	33.50	1,336	4.79	37.56	31.37	30.67	1,354	5.21	36.04
10	32.65	32.65	1,240	4.58	35.05	32.20	32.01	1,184	4.62	34.77
11	47.22	32.47	1,228	4.48	40.81	31.72	31.72	1,179	4.63	37.93
12	32.78	32.76	1,276	4.89	34.45	33.34	29.92	1,279	5.27	33.50
13	22.01	21.81	1,007	5.48	36.55	21.85	21.59	981	5.57	36.18
14	20.09	17.17	695	4.85	36.22	23.58	23.33	977	4.75	39.58
1-14	540.27	518.80	20,197	4.77	36.61	493.10	472.06	19,388	5.03	35.56
0	876.47	860.02	6195	0.41	47.33	1,476.30	1,455.48	13,348	0.60	46.97
1-14-0	**1,416.74**	**1,378.82**	26,392	2.05	43.30	1,969.40	1,927.54	32,736	1.68	44.19

Pseudomolecules lengths with (w N) and without (wo N) the Ns (any base) nucleotide in Mbp.

The percentage of CDS/repeats is calculated as the percentage of bases covered by the CDS/repeats sequences per chromosome.

Supplemental Table 13 Number and repartition of protein coding genes per chromosome and per haplotype of Daphna

	Daphna Haplotype A			Daphna Haplotype B		
Chr	w N	wo N	genes	w N	wo N	genes
1	88.30	88.28	4,619	83.42	83.40	4,177
2	72.75	72.74	1,962	76.19	76.18	2,126
3	50.16	50.15	1,670	54.87	54.86	1,781
4	50.01	50.00	1,817	52.15	52.14	1,781
5	52.22	52.21	2,571	52.41	52.40	2,391
6	48.33	48.32	1,739	43.88	43.87	1,673
7	42.01	42.00	1,722	40.07	40.06	1,555
8	35.52	35.52	1,538	34.78	34.77	1,583
9	35.31	35.31	1,681	37.56	37.56	1,661
10	35.05	35.04	1,314	37.45	37.44	1,494
11	37.57	37.56	1,410	40.79	40.78	1,504
12	41.05	41.04	1,712	40.14	40.13	1,738
13	27.55	27.55	1,213	25.81	25.80	1,210
14	23.25	23.25	1,066	22.80	22.80	1,093
1-14	639.09	638.96	26,034	642.33	642.19	25,767
0	97.73	97.66	3133	101.93	101.87	3413
1-14-0	736.82	736.62	29,167	744.26	744.06	29,180

Pseudomolecules lengths with (w N) and without (wo N) the Ns (any base) nucleotide in Mbp.

Supplemental Table 14 Conserved and specific gene groups

	Gene number	Genes in orthogroups	Number of conserved orthogroups	Number of specific orthogroups	Number of specific singleton genes
CR0040 Haplotype A	26,392	25,779	15,916	88	613
CR0040 Haplotype B	32,736	29,689	16,211	899	3,047
Daphna Haplotype A	29,167	26,506	17,131	135	2,661
Daphna Haplotype B	29,180	26,478	17,060	109	2,702
Phalaenopsis equestris	20,081	19,662	13,045	79	419
Phalaenopsis aphrodite	28,903	27,946	13,322	288	957
Oryza sativa	43,770	32,466	12,081	2,731	11,304

14

Supplemental Table 15 Sequencing depth of CR0040 genome per technology

	Haplotype A			Haplotype B		
	PacBio HiFi	ONT	Illumina	PacBio HiFi	ONT	Illumina
Chr0	10.97	19.08	50.55	9.62	15.74	45.26
Chr1	45.53	64.55	73.13	38.93	55.94	62.44
Chr2	51.98	73.30	75.01	34.81	49.41	51.10
Chr3	51.47	81.00	83.68	33.91	43.69	45.16
Chr4	56.02	74.51	76.93	32.98	55.89	58.90
Chr5	53.17	82.57	87.21	34.24	43.53	50.84
Chr6	71.49	103.11	102.67	17.30	24.86	26.50
Chr7	50.15	34.58	55.71	54.54	74.65	76.21
Chr8	47.15	68.15	68.65	56.91	41.25	62.25
Chr9	46.44	69.27	74.45	37.00	49.75	53.92
Chr10	46.74	74.06	80.02	41.63	51.10	53.22
Chr11	24.79	36.80	42.83	49.90	73.66	74.02
Chr12	50.19	77.09	87.36	35.28	44.44	44.44
Chr13	36.78	52.40	55.33	53.44	75.91	76.66
Chr14	40.99	78.69	86.60	35.22	32.76	37.31

32 Supplemental Table 16 Sequencing depth of CR0040 genome using Daphna data sequences

	Haplotype A		Haplotype B	
	ONT	Illumina	ONT	Illumina
Chr0	2.50	17.93	3.18	24.05
Chr1	28.40	106.18	27.12	101.70
Chr2	26.85	104.49	28.96	112.48
Chr3	28.71	114.75	25.02	105.23
Chr4	27.21	108.87	27.74	112.54
Chr5	29.20	110.89	28.37	107.20
Chr6	31.17	116.08	27.51	106.37
Chr7	26.00	100.27	28.48	110.39
Chr8	27.39	108.01	29.08	115.18
Chr9	26.45	102.54	26.89	104.81
Chr10	28.53	110.27	28.93	113.21
Chr11	17.10	66.85	29.10	111.80
Chr12	29.42	114.87	26.51	103.46
Chr13	29.32	112.36	28.49	111.60
Chr14	22.67	90.23	23.43	91.57

33 Supplemental Table 17 Cell cycle regulation genes found in CR0040 genome

		A. thaliana	CR0040 haplotype A	CR0040 haplotype B	*P. aphrodite*
CDKs	CDKA	At3g48750	VANPL_A_00001t012310	VANPL_B_00001t011840	PAXXG231470
			VANPL_A_00014t005470	VANPL_B_00014t007380	
	CDKB	At3g54180	VANPL_A_00007t008490	VANPL_B_00007t007640	PAXXG089720
		At2g38620	VANPL_A_00009t001400	VANPL_B_00009t001350	PAXXG113920
		At1g76540			
		At1g20930			
CYCs					
	CYCA	At1g80370	VANPL_A_00008t011460	VANPL_B_00004t012020	PAXXG008040
		At1g15570	VANPL_A_00005t000960	VANPL_00313t000620	PAXXG183010
		At5g25380		VANPL_B_00008t011460	PAXXG054020
		At5g11300		VANPL_B_00005t001350	
		At5g43080			
		At1g47210			
		At1g47220			
		At1g47230			
	CYCB	AT4G34160			
		At4g37490	VANPL_A_00001t011770	VANPL_B_00001t011320	PAXXG000750
		At5g06150	VANPL_A_00001t024950	VANPL_B_00001t019080	PAXXG156550
		At3g11520	VANPL_00050t000750	VANPL_B_00001t011320	PAXXG228890

		At2g26760	VANPL_A_00001t019600	VANPL_B_00005t000920	PAXXG028470
		At1g34460	VANPL_A_00001t011770		PAXXG000750
		At4g35620	VANPL_A_00001t024950		PAXXG156550
		At2g17620	VANPL_00050t000750		PAXXG219870
		At1g16330	VANPL_A_00005t000640		
			VANPL_A_00005t000650		
			VANPL_A_00005t000650		
	CYCD3-1	At4g34160			
FZR (CCS52A)		At4g22910	VANPL_A_00010t004300	VANPL_B_00010t004420	PAXXG076980
		At4g11920	VANPL_A_00001t012550	VANPL_B_00001t012090	PAXXG084080
		At5g13840		VANPL_13258t000030	PAXXG082520
WEE1		At1g02970	VANPL_A_00011t010580	VANPL_B_00011t010080	PATC144561
KRP		At2g23430	VANPL_A_00008t003440	VANPL_B_00006t003720	PAXXG048550
		At3g50630	VANPL_A_00004t006200	VANPL_B_00008t003330	PAXXG111330
		At5g48820	VANPL_A_00002t006250	VANPL_B_00004t006400	PAXXG290150
		At2g32710	VANPL_A_00001t031310	VANPL_B_00002t006550	PAXXG115580
		At3g24810		VANPL_B_00001t028840	
		At3g19150		VANPL_B_00001t020090	
		At1g49620		VANPL_00053t001550	
ILP1		At5g08550	VANPL_A_00012t005730	VANPL_B_00012t005680	PAXXG094170
ADF		At3g46010	VANPL_A_00010t001900	VANPL_B_00009t008980	PAXXG101780
		At3g46000	VANPL_A_00011t003650	VANPL_B_00010t001860	PAXXG066610
		At5g59880	VANPL_A_00009t008690	VANPL_B_00002t003560	PAXXG194280

At5g59890	VANPL_A_00001t019380	VANPL_B_00009t002390	PAXXG119700
At2g16700	VANPL_A_00002t003620	VANPL_B_00001t012150	PAXXG241790
At2g31200	VANPL_A_00014t004200	VANPL_B_00011t003680	PAXXG043520
At4g25590	VANPL_A_00005t008770	VANPL_B_00001t018850	PAXXG173900
At4g00680	VANPL_A_00003t000430	VANPL_B_00014t004330	PAXXG146670
At4g34970	VANPL_A_00009t002490	VANPL_00170t000390	PAXXG125360
At1g01750	VANPL_A_00006t002570	VANPL_01089t000240	PAXXG160710
At3g45990			PAXXG009590
At5g52360			PAXXG386760
			PAXXG302270

CDKs : Cyclin-Dependent Kinases ; CYCs : Cyclins ; FZR : Fizzy-Related protein ; WEE1 : WEE1 like protein kinase ; KRP : cyclin-dependent kinase inhibitor ; ILP : Transcriptional repressor ILP1 ; ADF : Actin-Depolymerizing Factor ;