TITLE PAGE

Title:

Diploid genome assembly of the Malbec grapevine cultivar enables haplotype-aware analysis of transcriptomic differences underlying clonal phenotypic variation

Running Title:

Diploid assembly and clone variation in Malbec

Authors:

Calderón, Luciano[1*#]; Carbonell-Bejerano, Pablo[2,3*#]; Muñoz, Claudio[1,4]; Bree, Laura[5]; Sola, Cristobal[5]; Bergamin, Daniel[5]; Tulle, Walter [1]; Gomez-Talquenca, Sebastian[6]; Lanz, Christa[3]; Royo, Carolina[2]; Ibáñez, Javier[2]; Martinez-Zapater, José Miguel[2]; Weigel, Detlef[3] & Lijavetzky, Diego[1*]

Affiliations:

[1] Instituto de Biología Agrícola de Mendoza (CONICET-UNCuyo), Genetica y Genomica de Vid. (5505) Mendoza, Chacras de Coria, Argentina.

[2] Instituto de Ciencias de la Vid y del Vino, ICVV, CSIC - Universidad de La Rioja - Gobierno de La Rioja. (26007) La Rioja, Logroño, Spain

[3] Department of Molecular Biology, Max Planck Institute for Biology Tübingen. (72076) Tübingen, Germany.

[4] Facultad de Ciencias Agrarias (UNCuyo), Cátedra Fitopatología. (5505) Mendoza, Chacras de Coria, Argentina.

[5] Vivero Mercier Argentina. (5500) Mendoza, Perdriel, Argentina.

[6] Instituto Nacional de Tecnología Agropecuaria, Plant Virology Laboratory. (5534) Mendoza, Luján de Cuyo, Argentina.

Authors emails:

; lbree@viveromercier.com.ar; cjdsola@yahoo.com.ar; dbergamin@viveromercier.com.ar; wtulle@mendoza-conicet.gob.ar; gomez.talquenca@inta.gob.ar; christa.lanz@tuebingen.mpg.de; carolina.royo@icvv.es; javier.ibanez@icvv.es; zapater@icvv.es; weigel@tue.mpg.de

Corresponding Authors:

* Calderón, Luciano: lcalderon@mendoza-conicet.gob.ar ; tel: +54 2614054809 ext. 61307

* Carbonell-Bejerano, Pablo: pablo.carbonell@icvv.es; tel: +34 941 894 980

* Lijavetzky, Diego: dlijavetzky@conicet.gov.ar ; tel: +54 2614054809 ext. 61307

Symbols reference: #These authors contributed equally to this work; * Corresponding authors.

1. Abstract

To preserve their varietal attributes, established grapevine cultivars (*Vitis vinifera* L. ssp. *vinifera*) must be clonally propagated, due to their highly heterozygous genomes. Malbec is a France-originated cultivar appreciated for producing high-quality wines, and is the offspring of cultivars Prunelard and Magdeleine Noire des Charentes. Here, we have built a diploid genome assembly of Malbec, after trio binning of PacBio long reads into the two haploid complements inherited from either parent. After haplotype-aware deduplication and corrections, complete assemblies for the two haplophases were obtained with very low haplotype switch-error rate (<0.025). The haplophases alignment identified >25% of polymorphic regions. Gene annotation including RNA-seq transcriptome assembly and *ab initio* prediction evidence resulted in similar gene model numbers for both haplophases. The annotated diploid assembly was exploited in the transcriptomic comparison of four clonal accessions of Malbec that exhibited variation in berry composition traits. Analysis of the ripening pericarp transcriptome using either haplophases as reference yielded similar results, although some differences were observed. Particularly, among the differentially expressed genes identified only with the Magdeleine-inherited haplotype as reference, we observed an over-representation of hypothetically hemizygous genes. The higher berry anthocyanin content of clonal accession 595 was associated with increased abscisic acid responses, possibly leading to the observed overexpression of phenylpropanoid metabolism genes and deregulation of genes associated to abiotic stress response. Overall, the results highlight the importance of producing diploid assemblies to fully represent the genomic diversity of highly heterozygous woody crop cultivars and to unveil the molecular bases of clonal phenotypic variation.

56 **2. Introduction:**

57 The cultivated grapevine (*Vitis vinifera* L. subsp. *vinifera*) was one of the first plant species
58 accounting for a reference genome [1]. To facilitate genome assembly, this reference was produced from a
59 nearly homozygous line (i.e. PN40024) obtained after several rounds of selfing of the cultivar
60 Helfensteiner, the result of Pinot noir and Schiava grossa outcross [1,2]. The PN40024 reference genome has
61 been a key tool for the study of grapevine molecular biology and genetics and both the reference assembly
62 and annotation have been continuously improved [2–4].

63 The genomes of grapevine cultivars are characterized by enormous diversity, including the many
64 differences that typically distinguish the two haplotypes of each cultivar [7]. On average one single
65 nucleotide polymorphism (SNP) every 100 bp and one Insertion-deletion (InDel) difference every 450 bp
66 [7,8]. Also, larger structural variants (SVs) ≥50 bp have been reported to affect ~17% of the genome space
67 of each haplotype within the same cultivar [9]. Hemizygosity is also rampant in grapevines, since many SVs
68 cause presence-absence variation (PAV) between haplotypes [8], which may have severe phenotypic
69 impacts [10]. In addition to the PN40024 reference genomes, several additional cultivars have seen their
70 genomes assembled from Pacific Biosciences (PacBio) long reads using the haplotype-aware assemblers
71 FALCON-unzip [6,9,11–14]. FALCON-unzip outputs pseudo-haplotypes, which usually represent a single allele
72 at each position without preserving phase information across the genome [15], although tools such as
73 HaploSync have been developed to completely phase this type of assemblies [16]. In this context, the
74 combination of multiple sequencing technologies such as Oxford Nanopore Technologies (ONT) and short
75 reads[21], the use of PacBio HiFi and newer algorithms such as HiFiasm [17], as well as information from
76 parent-offspring trios [15] have notably improved assembly phasing and quality of grapevine genomes [4,18
77 [19].

78 The high and uneven heterozygosity of cultivated grapevine genomes is most likely the
79 consequence of the outcrossing nature of their dioecious wild ancestor (*V. vinifera* ssp. *sylvestris*) and
80 their vegetative way of propagation [6]. Moreover, cultivars relevant for wine making have originated from
81 complex processes including the outcrossing of preexisting genetically divergent cultivars [20] and eventual
82 post-domestication hybridization with local wild relatives of the subspecies *sylvestris* [21,22]. Irrespective of
83 their origin, progenies with desired phenotypic traits have been vegetatively propagated to retain those
84 attributes. Although vegetative propagation mostly ensures cultivar uniformity, it also favors the
85 accumulation of somatic mutations in the absence of outcrossing [23]. In fact, somatic variation is known to
86 exist at the phenotypic [24], genetic [13] and transcriptomic [25] levels. Moreover, within each grape cultivar
87 growers select somatic variants to propagate them, labelling these selections as clones. Therefore,
88 somatic diversity has been relevant for viticulture as a source of trait innovation and adaptation. Because

89 the global wine market is very slow at accepting new outcross varieties, a deep understanding of the
90 variation that exists within elite varieties is crucial for intra-cultivar improvement and to meet emerging
91 challenges from climate change [26].

92 Malbec, a red cultivar appreciated for its potential to produce high-quality wines, originated in the
93 Cahors region of France (where it is known as Cot) as the result of an outcross of cultivars Magdeleine
94 Noir des Charentes (hereafter Magdeleine) and Prunelard [27]. Malbec was introduced to Argentina in the
95 1850s, becoming the most relevant cultivar for the wine industry of this country [28]. Currently, Argentina
96 harbors the largest planted surface of Malbec in the world [29] and 85% of this production is concentrated
97 in Mendoza province [30]. Since its introduction, this cultivar has shown remarkable adaptability, producing
98 wines with properties that reflect the different growing conditions [31]. The observed adaptability may be
99 linked to the phenotypic plasticity exhibited by Malbec, as clonal variability has been reported for traits
100 relevant for the industry, such as polyphenolic composition of berries [32] and phenology [33]. Also molecular
101 plasticity has been described among clones, since epigenetic [34] and genetic [35] variability has been
102 reported.

103 The main objective of this study was to produce a phased and annotated genome assembly of
104 Malbec representing both haploid complements, or haplophases, to serve as reference for intra-cultivar
105 variation studies. We used our assembly in a transcriptome analysis to investigate the molecular bases of
106 Malbec clonal variants that differ in fruit traits relevant for wine making. The transcriptomic analysis was
107 performed following a haplotype-aware scheme and the results obtained using separately each
108 assembled haplophase as reference were compared. This allowed us to further evaluate potential
109 reference biases, and to identify haplotype-specific differentially expressed genes that could be correlated
110 with clonal variation in berry composition.

111 **3. <u>Results:</u>**

112 **The diploid *de novo* genome assembly of Malbec:** A total of 82.3 Gb of sequence coverage (168x
113 coverage of the haploid genome) was produced for Malbec in 4,117,439 PacBio CLR reads, with a mean
114 read length of 20 kb and read N50 of 33.7 kb (Table S1A). Illumina short reads were obtained for Malbec
115 (13.4 Gb) as well as the parental cultivars Prunelard (46.9 Gb) and Magdeleine (47.6 Gb) (Table S1B).
116 Malbec PacBio reads were binned into two sets of 35.9 Gb and 46.0 Gb, corresponding to the haplotypes
117 inherited from Magdeleine and Prunelard (Table S2A). The binned PacBio long reads were separately
118 assembled into two haplophases, named hereafter Malbec-Pru for the inherited from Prunelard and
119 Malbec-Mag for the inherited from Magdeleine. The draft raw assemblies of Malbec-Pru and Malbec-Mag
120 had good contiguity, with contig N50 = 6.2 Mb and 5.6 Mb, and gene completeness of 98.8% and 98.2%
121 of BUSCO core genes represented (Table S2B). However, the draft assemblies were larger than the

122 expected for a haploid grapevine genome of ~475 Mb [2], namely 608 Mb for Malbec-Pru and 558 Mb for
123 Malbec-Mag, and they had a high percentage of duplicated BUSCO genes, 21.7% and 10.4% (Table S2B).
124 After haplotype-aware deduplication, consensus polishing and miss-assembly corrections, the two
125 haplophases were closer to the expected size, 479.3 Mb and 479 Mb, with improved contiguity of N50 =
126 7.7 Mb and 6.6 Mb. Additionally, we obtained a smaller percentage of duplicated BUSCO genes (i.e., 2.7%
127 and 1.9%) without a significant change in completeness: 98.6% and 98.1% (Table 1A). The polished and
128 deduplicated haplophases were reference-based scaffolded, resulting in 19 pseudomolecules, as
129 expected for the haploid set of grapevine nuclear chromosomes (Fig. S1), with a scaffold N50 = 24.4 for
130 Malbec-Pru and 24.5 Mb for Malbec-Mag (Table 1B). Some unplaced contigs remained after the
131 scaffolding, for Malbec-Pru = 21 contigs (total length = 3.3 Mb) and for Malbec-Mag = 32 contigs (total
132 length = 5.74 Mb). Malbec-Mag corresponds to the maternally inherited haplotype [27], and included one
133 scaffold for a partial mitochondrial chromosome (length = 115,997 bp). Finally, a group of contigs of short
134 length (N50 = 0.2 Mb), that were slightly over-represented in repetitive sequences and under-represented
135 in BUSCO complete genes (avg. 17%) compared to the assembly size, were discarded from the final
136 assembly after the deduplication process (Table S2C).

137 **Table 1**. Summary statistics obtained for the Malbec diploid genome assembly after deduplication
138 pipeline. (A) Results obtained for Malbec-Pru and Malbec-Mag contigs. (B) Results obtained after
139 scaffolding process for both Malbec haplophases and the diploid (2n) assembly, compared to the
140 PN40024.v4 (REF) and PN40024.v5 (T2T) reference genome assembly versions. Percentages are
141 expressed over a total of 2,326 BUSCO genes in the eudicots lineage. (C) Evaluation of the phasing
142 process performed by comparing *k-mers* between the Malbec haplophase assemblies and raw
143 short-reads from the parental cultivars.
144

145 **A**

	Summary stats	Malbec-Pru	Malbec-Mag
Deduplicated final contigs	Total length (Mb)	479.36	479.06
	Number	163	179
	Longest (Mb)	18.3	17.6
	Mean length (Mb)	2.94	2.67
	N50 (Mb)	7.69	6.64
BUSCO genes (%)	Complete	98.6	98.1
	Single	95.9	96.2
	Duplicated	2.7	1.9
	Fragmented	0.8	0.6
	Missing	1.1	0.8

146

147 **B**

	Summary stats	Malbec-Pru	Malbec-Mag	Malbec-2n	PN40024v4	PN40024v5
Scaffolds	Total length (Mb)	479.35	479.05	958.4	475.6	504.6
	Number	47	52	99	22	20
	Longest (Mb)	38.82	38.89	38.89	34.9	36.6
	N50 (Mb)	24.42	24.48	24.89	24.38	26.89
	GC (%)	34.57	34.62	34.60	34.53	35.32
	Gaps	115	127	242	4019	169
	N counts	11,500	12,700	24,200	1,818,262	16,900
BUSCO genes (%)	Completeness	98.4	98.0	98.7	98.4	98.5
	Single	95.7	96.1	1.8	96.1	96.7
	Duplicated	1.9	1.8	96.9	2.3	1.8
	Fragmented	0.8	1.1	0.8	0.9	0.9
	Missing	0.8	0.9	0.5	0.7	0.6

148

149 **C**

Haplophase	Blocks Nº	Total bases In blocks	Block N50 size (bp)	Nº of switched makers	Total nº of markers	Switch error rate
Malbec-Pru	198	463,853,342	11,965,117	5,589	22,596,691	0.025%
Malbec-Mag	198	456,349,354	8,939,504	3,842	20,345,345	0.019%

150

151 Phasing accuracy and consensus quality was assessed on the final scaffolded assemblies. The
152 Malbec-Pru assembly comprised 97.3% of the Prunelard-specific *hap-mers* and only 0.13% of Magdeleine-
153 specific ones (Fig. 1A), whereas the Malbec-Mag assembly comprised 98.1% of the Magdeleine-specific
154 *hap-mers* and only 0.03% of the Prunelard-specific ones (Fig. 1B). The consensus quality score (QV) was
155 estimated for Malbec-Pru = 41.62 (error rate = 6.9 e-5) and Malbec-Mag = 42.06 (error rate = 6.2 e-5)
156 (Table 1C). Total bases placed in proper phase blocks and the phase blocks N50 for each scaffolded
157 haplophase assembly were for Malbec-Pru = 463.8 Mb (N50 = 11.96 Mb) and Malbec-Mag = 456.3 Mb
158 (N50 = 8.9 Mb) (Figs 1A and B, Fig. S2). Also, a very low haplophase switch error rate was observed, 0.025%
159 for Malbec-Pru and 0.019% for Malbec-Mag (Table 1C).

160

Figure 1

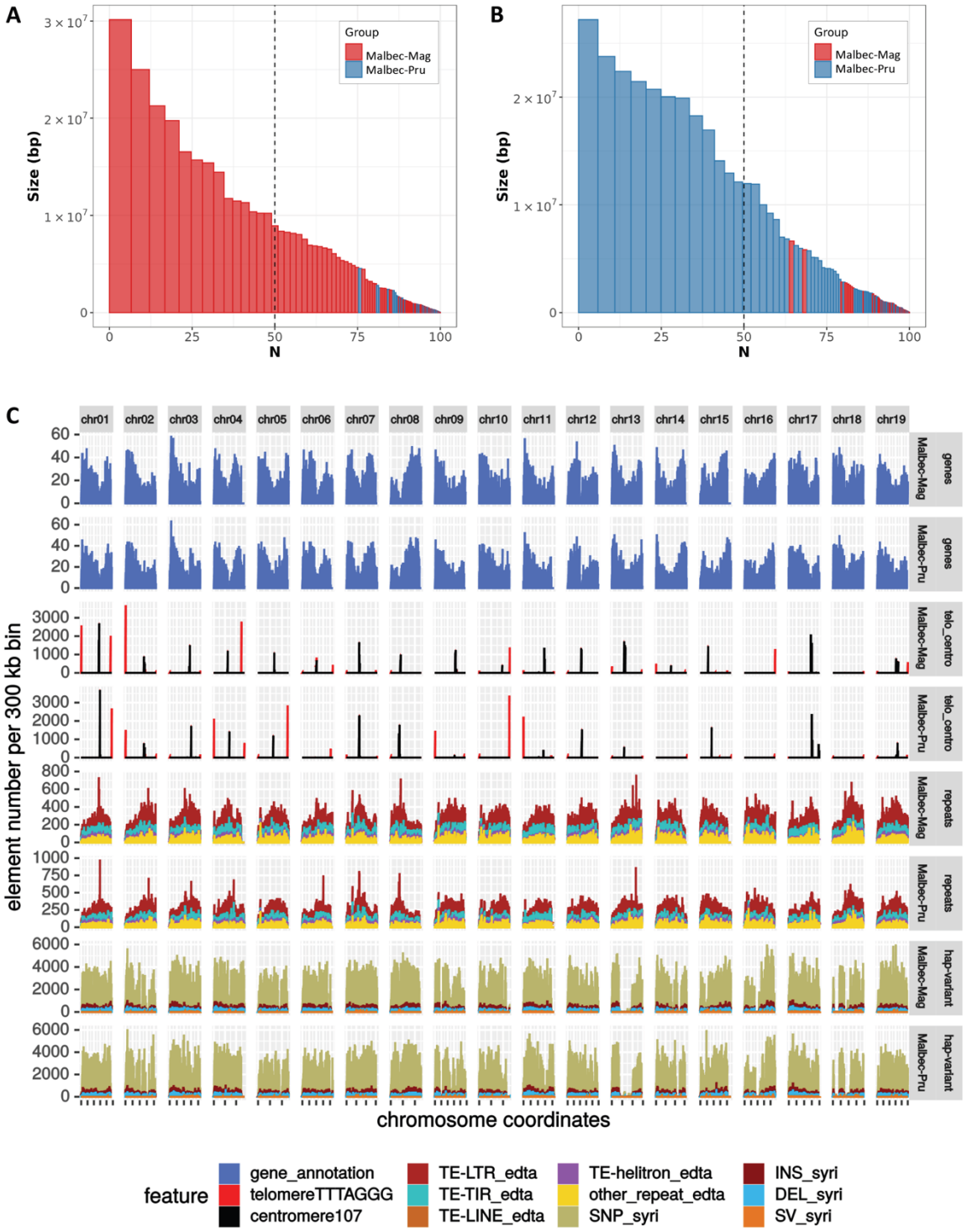
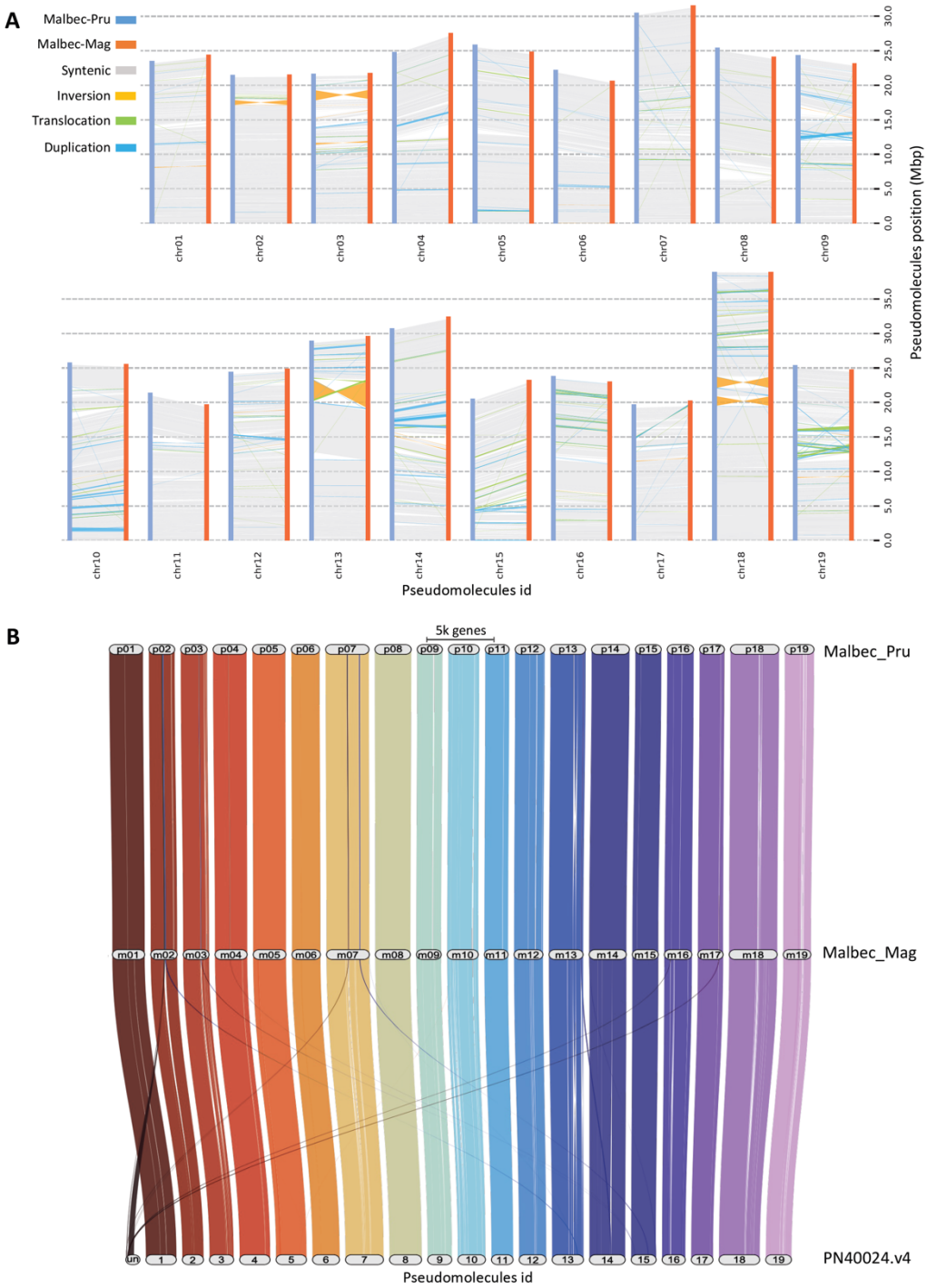


Figure 1. Barplots of the phased blocks sorted by size to visualize the efficiency of the phasing process for (A) Malbec-Mag and (B) Malbec-Pru haplophases. For the former the red bars are the own blocks and the blue bars are the switched blocks, for the latter the color code is inverted. Vertical bars represent phase blocks, defined as a continuous set of *k-mers* inherited from the same haplotype (i.e., *hap-mers*). The *X*-axis is the percentage of the genome size covered by blocks of this size of larger, expressed in base pairs (*Y*-axis). (B) Distribution of the annotated features along the 19 scaffolded pseudomolecules for Malbec haplophases, including: gene density, centromeres, telomeres, transposable elements (LTRs, TIR, LINE and Helitron) and haplotype variants (SNPs, INDELs, SVs).

170 The estimated divergence between the two haplophases did not change considerably based on
171 which one was used as reference, and we report the results obtained with Malbec-Pru as reference and
172 Malbec-Mag as query. Sequence differences affected 7.09% (33.55 Mb) of Malbec-Mag, this included
173 SNPs, InDels, copy number variations (CNVs), highly divergent regions (HDRs) and tandem repeats (TRs)
174 (Fig. 1C and Table S3). Considering the accumulative length, HDR was the most abundant type of sequence
175 variation and duplications (DUPS) were the most abundant type structural variation, affecting 4.9% (23.28
176 Mb) and 3.98% (18.84 Mbp) of Malbec-Mag (Fig. 2A). SNPs and InDels occurred every 147 and 1458 nt,
177 respectively; and SNPs distribution along the chromosomes was rather uneven (Fig. 1C). Several drops of
178 SNP density were detected, possibly as consequence of runs of homozygosity. The longest drop of SNPs
179 occurred on chr13 (Fig. 1C) and this overlapped with a large inversion between the haplophases (Fig. 2A),
180 which may appear as a SNP desert in the SyRI output [36]. Structural variations affected 9.97% (47.2 Mb) of
181 Malbec-Mag, more precisely inversions (INV) = 2.07%, translocations (TRANS) = 3.92% and duplications
182 (DUPS) = 3.98% (Fig. 2A and Table S3A). Finally, 76.5% (362.14 Mb) of the haplophases were identified as
183 syntenic and 14.2% (67.2 Mb) were classified as not aligned (Fig. 2A and Table S3A), altogether indicating
184 for a high heterozygosity of Malbec cultivar represented in the diploid assembly. Moreover, in both
185 haplophases an average of 29% of genes were affected by at least one SV (namely: TRANS, INV, DUP,
186 NOTAL) (Table S3C). Furthermore, genes overlapping with SVs were over-represented in biological
187 processes GO terms associated to 'defense response', 'external biotic stimulus' and 'DNA metabolic
188 processes' in both haplophases. At the same time, 'valine biosynthetic process' and 'phosphorylation'
189 were enriched solely in Malbec-Pru and 'glucan biosynthetic process' was enriched in Malbec-Mag (Table
190 S3D).
191

192 **Figure 2**



193 **Figure 2.** Synteny analysis based on the comparison of the 19 pseudomolecules assembled for Malbec
194 haplophases. (A) Structural variations (SVs) distribution along pseudomolecules, with Malbec-Pru (blue)
195 considered as the reference and Malbec-Mag (orange) as the query. Results are based on the percentages
196 estimated for the query, 76.5% (362.14 Mb) of the assembly resulted syntenic and 9.97% (47.2 Mb) was
197 affected by different types of SVs (inversions, translocations and duplications). (B) Riparian plots
198 representing the synteny of orthologous genes along the 19 scaffolded pseudomolecules for each haploid
199 complement of Malbec. A strong synteny was observed among the assembled pseudomolecules of
200 Malbec haplophases and the grapevine reference genome (PN40024.v4). Also, genes contained in the
201 "Unknown" (Un) chromosome of the PN40024.v4 exhibited orthologous genes assigned to different
202 pseudomolecules in Malbec haplophases.

203 **Gene model predictions for the assembled haplophases:** Annotation and soft-masking of
204 transposable elements (TEs) was performed on the Malbec diploid assembly, after concatenating the two
205 haplophases (958 Mb). Almost half, 48.4 % (464 Mb), of the genome accounted for TEs, including
206 structurally intact and fragmented TEs (Table S4). The great majority was classified as long terminal
207 repeats (LTRs) representing 28.8% (276 Mb) of the genome, whereas Gypsy was the most abundant type
208 among the LTRs class with 13.36% (128 Mb) (Fig. 1C, Table S4). In addition, telomeric and centromeric
209 repeats were identified in the diploid assembly. Out of the 19 scaffolded pseudomolecules for each
210 haplophase, 14 and 18 featured telomeric repeats in at least one end in Malbec-Mag and Malbec-Pru (Fig.
211 1C). Centromeric repeats were also detected (between 173 and 5817 clustered copies) in most
212 pseudomolecules, with the exception of chr16 and chr18 for both haplophases, and for chr06, chr10 and
213 chr14 for Malbec-Pru. The fact that the annotated position of centromeric repeats overlapped between
214 the two haplophases, as well as with LTR-TEs, suggest that these regions might correspond to the -at least
215 partially- assembled centromeres (Fig. 1C).
216 The final weighed consensus gene models obtained for the Malbec diploid genome assembly
217 yielded similar metrics to those of the highly curated PN40024.v4 [2] annotations, whereas the annotations
218 of the T2T PN40024.v5 assembly [4] differed in parameters like exon and UTR length (Table 2). The number
219 of protein coding loci was 35,739 for Malbec-Pru and 35,463 for Malbec-Mag, with an average of 1.2
220 transcript isoforms per locus (Table 2). Annotated genes were similarly distributed when comparing the
221 pseudomolecules of both haplophases. Drops of gene density overlapped with centromeric repeats, and
222 the former were also evident even in those pseudomolecules where peaks of centromeric repeats were
223 not detected, like chr16 and chr18 (Fig. 1C), further supporting the putative position of the centromeres.
224
225 **Table 2.** Coding protein sequences predicted for Malbec haplophases (considering isoforms),
226 compared to the PN40024.v4 (REF) and the PN40024.v5 (T2T) annotations.
227

	Malbec-Pru		Malbec-Mag		PN40024.v4 (REF)		PN4024.v5 (T2T)	
	Total number	Mean length (bp)	Total number	Mean length (bp)	Total number	Mean length (bp)	Total number	Mean length (bp)
Protein coding genes	35,739	5,086	35,463	5,071	35,197	4,742	37,476	6,133
Transcripts	44,590	5,025	44,183	5,017	41,160	4,689	40,994	6,494
Exons	236,022	319	233,585	320	208,719	283	179,096	645
CDS	44,590	1,182	44,183	1,180	41,160	1,126	40,994	1,235
5' UTRs	21,386	371	21,221	372	17,478	280	23,267	1,327
3' UTRs	22,138	670	21,818	676	18,344	440	23,278	1,459

228

229

230 Considering all annotated isoforms, the total number of predicted proteins for Malbec
231 haplophases were higher (Malbec-Pru = 44,590 and Malbec-Mag = 44,183) than the one reported for the
232 PN40024.v4 and v5 annotations (Table 3A). However, the size metrics such as the N50 of the predicted
233 proteins were more similar (Table 3A). BUSCO completeness for the proteins annotated on each
234 haplophase was close to 97%, whereas 98.6% completeness was achieved when all proteins annotated in
235 the diploid assembly (Malbec-2n) were analyzed (Table 3B).

236

237 **Table 3.** (A) Summary metrics and (B) BUSCO completeness analysis of the proteins predicted in
238 the genome assemblies. Results retrieved for each Malbec haplophase and for the diploid (2n)
239 assembly are compared to the PN40024.v4 (REF) and PN40024.v5 (T2T) annotations that were also
240 analyzed as quality control.

241 **A**

Predicted proteins	Malbec-Pru	Malbec-Mag	Malbec-2n	PN40024 v4 (REF)	PN40024 v5 (T2T)
Total proteins	44,590	44,183	88,773	41,160	40,994
Min. length (aa)	37	41	37	50	18
Average length (aa)	393	393	393	375	411
Max. length (aa)	5,480	5,488	5,488	5,481	66,180
N50 (aa)	532	534	533	519	543

242

243 **B**

BUSCO proteins (%)	Malbec-Pru	Malbec-Mag	Malbec-2n	PN40024 v4 (REF)	PN40024 v5 (T2T)
Completeness	96.8	96.7	98.6	98.7	95.8
Single	72.7	72.1	3.8	80.3	88.7
Duplicated	24.1	24.6	94.8	18.4	7.1
Fragmented	1.2	1.2	0.4	0.7	1.1
Missing	2.0	2.1	1.0	0.6	3.1

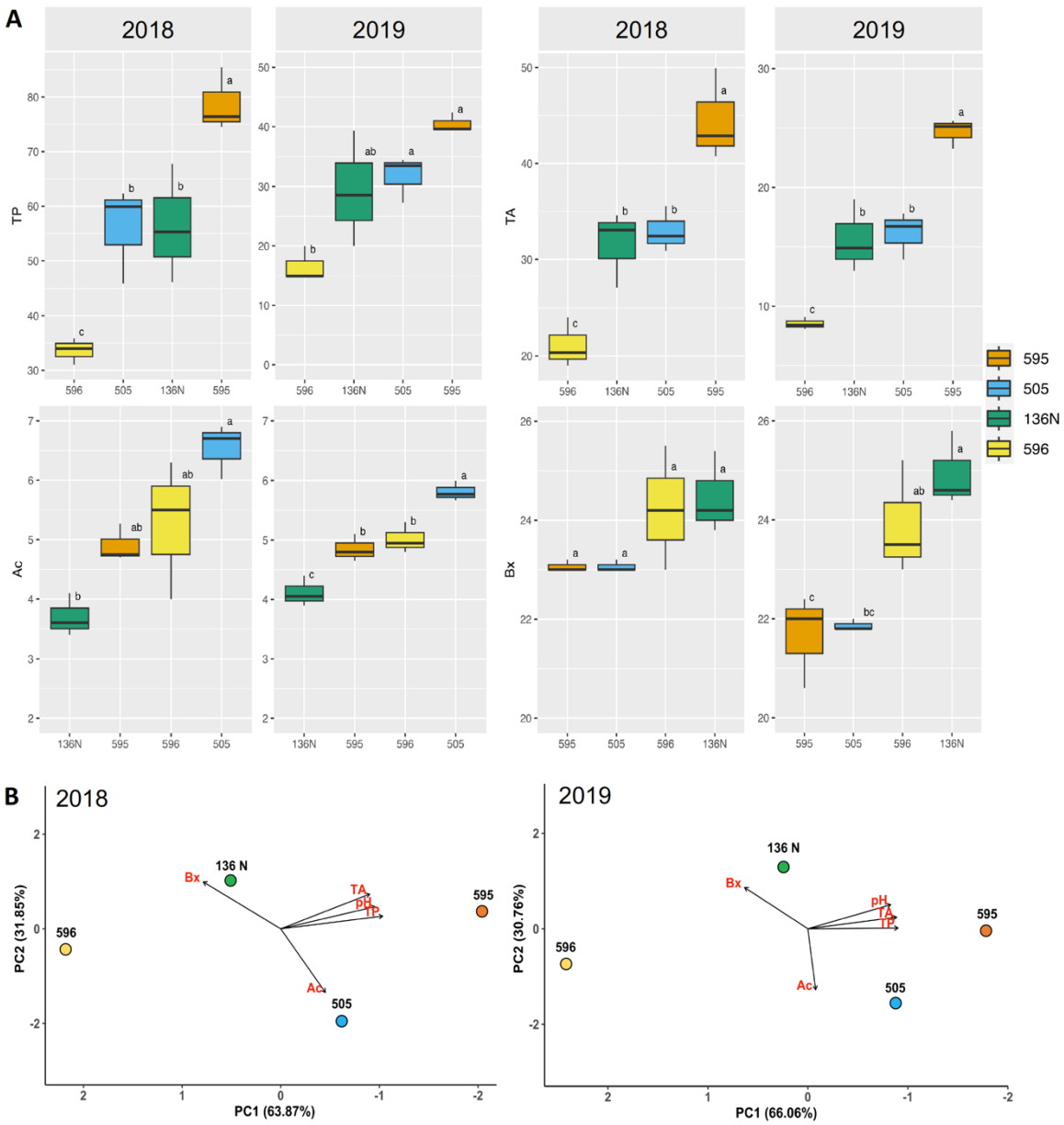
244

245 The orthology analysis based on predicted proteins showed, as expected [2,27,37], that there were
246 more orthologues between Malbec haplophases than with the more complete PN40024 v4 reference
247 genome predictions (Table S5A). Orthology was compared to the PN40024 v4 annotation, since it was
248 more complete than the annotations available for the v5 (Table 3B). When only Malbec haplophases were
249 compared ~96% of the proteins could be placed in orthogroups, on average 4.1% of the proteins remained
250 unassigned, and 2.45% were haplotype-specific orthogroups (Table S5B). The liftoff analysis, based on
251 annotated genes, between both haplophases showed that there were 1,915 genes of the Malbec-Mag
252 that were unmapped in Malbec-Pru and 2,296 unmapped genes for the inverse comparison. Interestingly,
253 the GO enrichment analysis of unmapped genes suggested that the trihydroxystilbene synthase activity
254 and the RNA (uridine-N3-)-methyltransferase molecular functions were overrepresented in Malbec-Mag

255 and Malbec-Pru, respectively (Fig. S3). The riparian plots for synteny visualization, showed that most gene
256 orthogroups matched in their physical location across the 19 pseudomolecules of Malbec haplophases
257 and PN40024.v4 (Fig. 2B). Furthermore, many genes located in the Unknown chromosome of the
258 PN40024.v4 had orthologues in genes assigned to chrMT, chr02, chr07, chr16 and chr17 in Malbec-Mag
259 and chr02, chr03, chr07 and chr17 in Malbec-Pru (Fig. 2B and Fig. S4).

260 **Malbec haplophases retrieved congruent -although not identical- transcriptomic differences,**
261 **that resembled the clonal phenotype variation:** The obtained diploid assembly was employed to further
262 understand the molecular bases of Malbec intra-cultivar phenotypic variation for berry composition traits.
263 Phenotyping of 27 clones of Malbec grown in the same plot showed significant diversity for traits
264 evaluated on mature berries. In a principal component analysis (PCA), the first component (PC1) explained
265 46.2% of the variation, mainly driven by total anthocyanins (TA), total polyphenols (TP) and must pH,
266 whereas PC2 explained 27.3% of the variation and appeared to be associated with sugar content (Bx) and
267 total acidity (Ac) (Fig. S5A). From the starting 27 phenotyped accessions, four clones (595, 596, 505 and
268 136N) were selected for further transcriptomic analysis (Fig. S5B). The greatest phenotypic differences
269 among the selected clones were associated with total polyphenols (TP) and total anthocyanin (TA)
270 concentrations, with accessions 595 and 596 having the highest and lowest values ($p<0.05$), whereas 505
271 and 136N had intermediate values (Fig. 3A). These differences among clones were consistent across
272 seasons, despite the inter-annual differences in the TP and TA absolute values (higher in 2018 than in
273 2019) (Fig. 3B). Overall, accession 595 showed the highest TP and TA concentrations in mature berries at
274 two consecutive harvest seasons.
275

276 **Figure 3**



277 **Figure 3.** Phenotypic analyses conducted on the four clones selected to perform RNAseq experiments. All
278 phenotypic data was obtained from mature berries during two consecutive years, 2018 and 2019.
279 Variables analyzed were total acidity (Ac), Brix (Bx), total anthocyanins (TA), total polyphenols (TP). (A)
280 Box-plots with median values and standard error bars. ANOVA results based on three biological replicates,
281 with the different letters representing significant differences, Tukey HSD (p<0.05). (B) PCAs summarizing
282 the phenotypic differences of the four selected clones for each analyzed year the greatest differences are
283 driven by pH, TP and TA represented by PC1.
284

285 The RNA-seq analysis of berry pericarp of the four selected clonal accessions produced on average
286 45.7 million raw reads (6.7 Gb) per replicate (14x raw coverage), with 93.4% of the reads having base QC
287 > 30 and an average GC content of 46% (Table S1C). After independent RNA-seq analysis, the sequenced
288 reads aligned to 37,807 and 37,556 annotated transcripts in the Malbec-Pru and the Malbec-Mag

haplophases, respectively. Despite this slight difference, the PCA based on global transcriptomic data showed that biological replicates of clone 595 were consistently differentiated from the other accessions (PC1 = 57%) (Figs 4A and B). According to their smaller phenotypic variation, the remaining three accessions appeared closer in the RNA-seq based PCA (PC2 = 17%) (Figs 4A and B). Differentiation of clone 595 was corroborated through clone pairwise comparisons, employing heatmaps of sample-to-sample distances (Fig. 4C and Fig. S6A to F). For all the described analyses the relations observed among the replicates of the four clonal accession was the same, using either of the haplophases as reference (Figs 4A-C and Fig S6A-F).

Figure 4

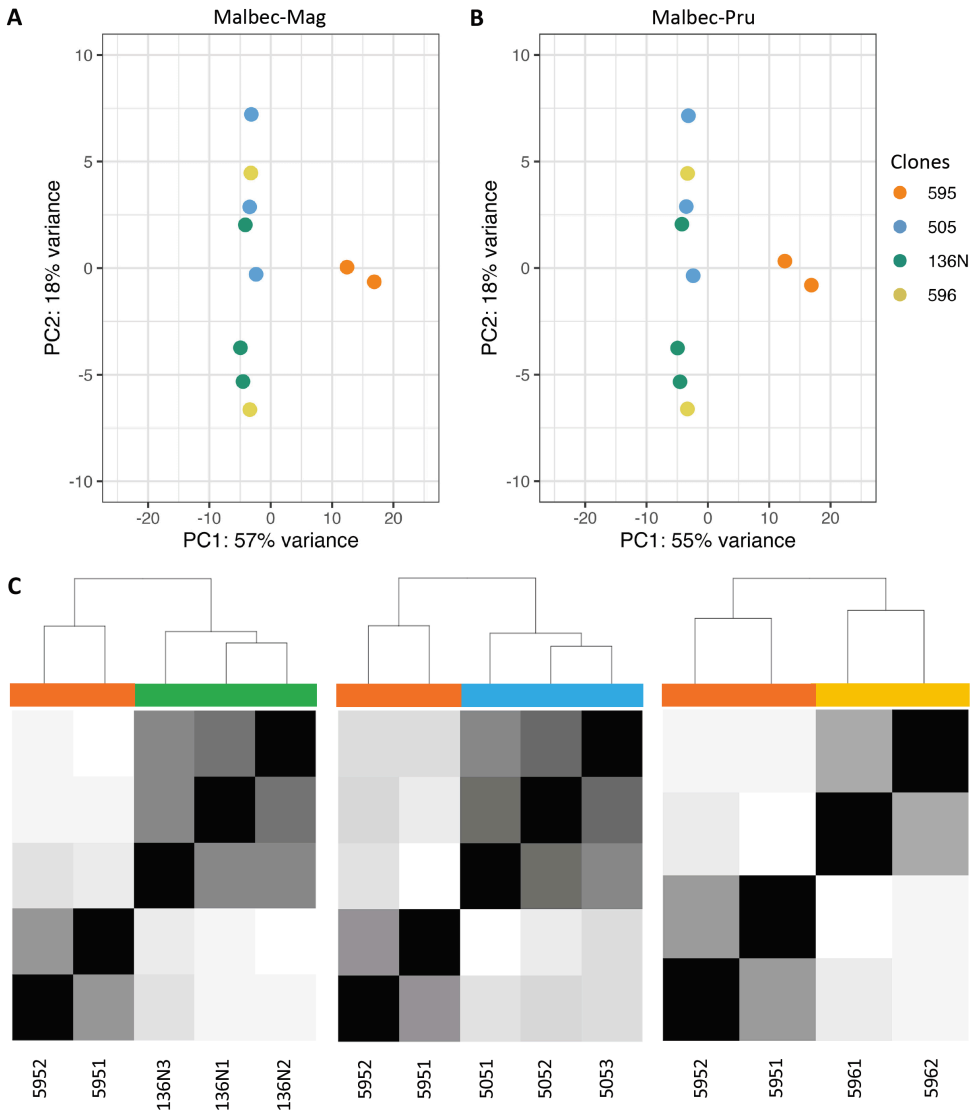


Figure 4. Principal components analysis based on global transcriptomic data based on total aligned transcripts to (A) Malbec-Mag and (B) Malbec-Pru. Both analyses showed consistent results, with the biological replicates of clone 595 differentiating from the rest based on PC1. The biological replicates of the other clones (136N, 505 and 506) differentiated less among each other based on PC2. (C) Pairwise

303 comparison of sample-to-sample distance between 595 and the other three clones also showed clear
304 differentiation. This analysis was performed with both haplophases and retrieved identical results;
305 therefore, only Malbec-Mag results are shown.

306 To further investigate the differentially expressed genes (DEGs) that were driving differentiation
307 of Malbec clone 595 from the others, pairwise comparisons were performed: 595vs136N, 595vs596 and
308 595vs505. For each comparison a very similar number of DEGs was detected with either haplophase as
309 reference (Figs 5A and B; Table S6). The number of DEGs ranged between 392 and 722, with 595vs505
310 comparison producing the highest number of DEGs and 595vs596 the lowest. Hereafter, downregulated
311 and upregulated DEGs are expressed in relative terms after comparing clone 595 to the other three. In all
312 comparisons there were more downregulated than upregulated DEGs (Figs 5A and B; Table S6). In each
313 comparison, on average 82% of the DEGs were detected for the orthologues of the two haplophases when
314 using in parallel each haplophase as reference for the differential expression analysis (Fig. 5C). Moreover,
315 236 DEGs were consistently detected with the two haplophases in the three pairwise comparisons
316 involving clone 595 (Fig. 5D). All these 236 DEGs showed the same direction of change using either of the
317 two haplophases as reference (Table S6). On the other hand, from DEGs detected uniquely with one
318 haplophase a higher number was detected with Malbec-Pru than with Malbec-Mag (Fig. 5C), whereas
319 there were 11 DEGs consistently detected in the three comparisons only with Malbec-Pru and six DEGs
320 detected uniquely with Malbec-Mag (Fig. 5D). On the other hand, for each pairwise comparison we
321 analyzed the subcategory of DEGs that did not account for an annotated orthologous gene in the other
322 haplophase (Table S6). We found that with Malbec-Mag an average of 7.4% of all the detected DEGs had
323 no ortholog annotated in the Malbec-Pru haplophase (Table S6), this was significantly above (Fisher's
324 exact test $p<0.05$) of the total number of unmapped genes (5.4% = 1,915 genes) in the other haplophase.
325 For Malbec-Pru the inverse scenario was observed, since an average of 3% of all the detected DEGs had
326 no orthologue in Malbec-Mag (Table S6) and this was significantly below (Fisher's exact test $p<0.05$) the
327 total percentage of unmapped genes (2,296 genes = 6.4%). We also found that among DEGs with no
328 ortholog annotated in the other haplophase, 20% in Malbec-Mag and 15% in Malbec-Pru were affected
329 by structural variations (Tables S3C and S6), adding support on the putative hemizygous condition of these
330 genes.
331

Figure 5

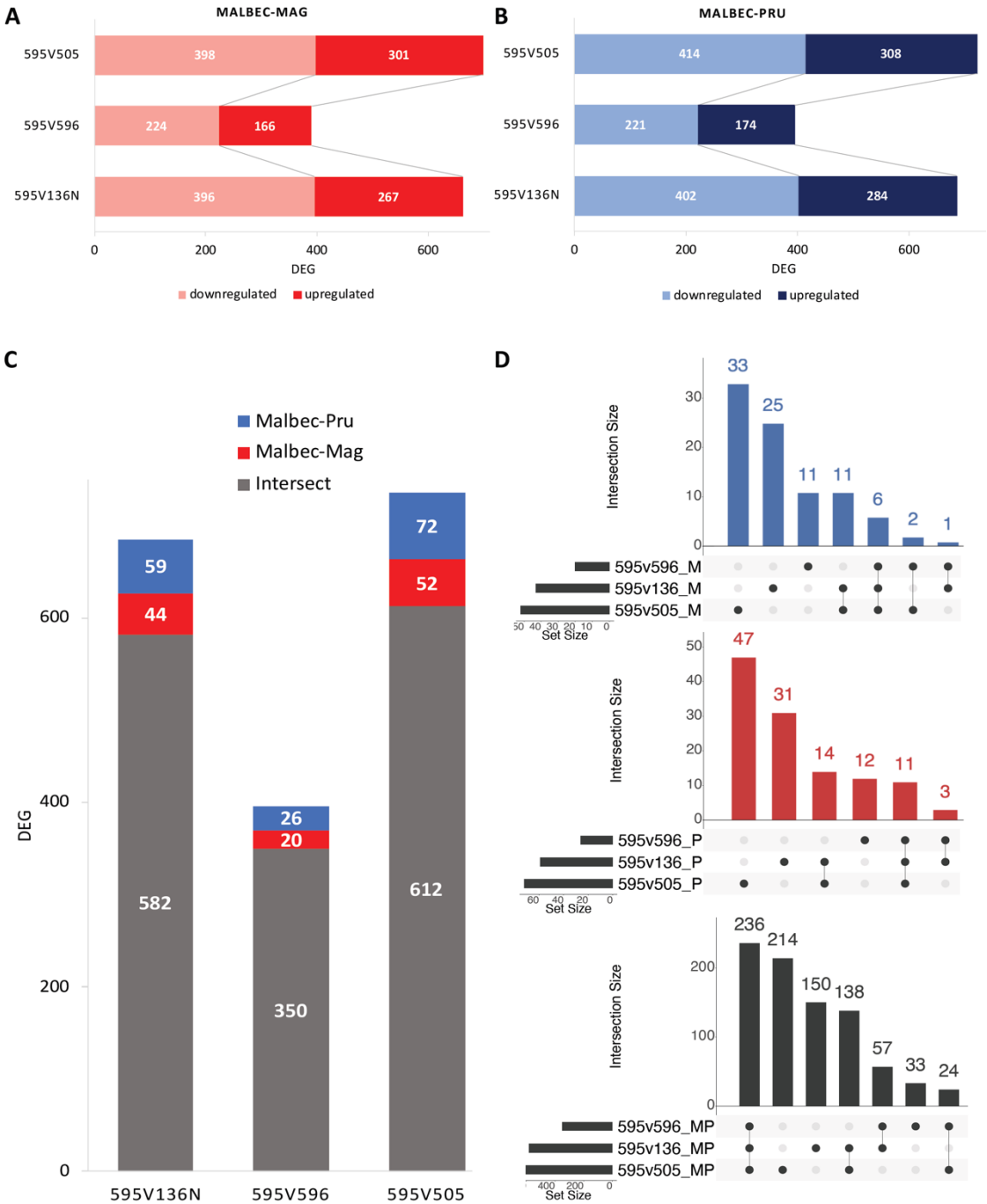
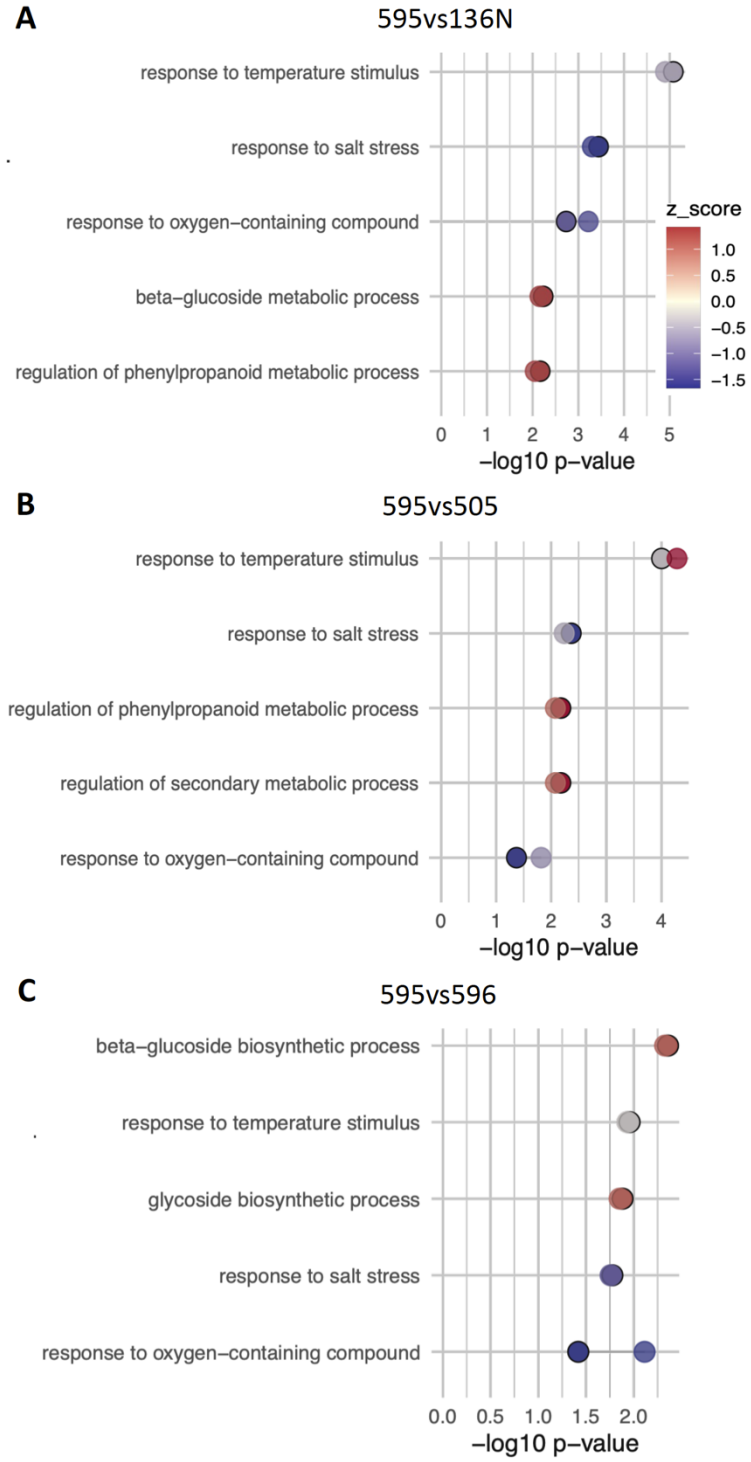


Figure 5. Total numbers of the statistically significant (p<0.01) differentially expressed genes (DEG). (A) and (B) Total number of DEGs detected in each clonal pairwise comparisons with Malbec-Mag and Malbec-Pru. Inside each bar downregulated and upregulated genes are distinguished based on the light and dark color code. (C) Intersection of the orthologous genes detected with both haplophases, for each clonal comparison. The grey portion of the bar represents the DEGs detected with both haplophases, the blue and red portions of the bar represent haplotype-specific DEGs detected with Malbec-Pru and Malbec-Mag, respectively. **d)** UpSet plots for the intersection among the three pairwise comparisons of the DEGs consistently detected with each haplophase, red bar-chart for Malbec-Mag, blue bar-chart for Malbec-Pru and black bar-chart for DEGs detected with both haplophases.

343 Gene Ontology (GO) enrichment analysis of biological processes (BPs) also produced similar results
344 using either of the haplophases as reference, both for the enriched terms and the direction of change
345 relative to clone 595. GO terms associated with 'response to abiotic stress', including 'response to salt
346 stress' and 'response to oxygen-containing compound' where downregulated, whereas genes related to
347 'secondary metabolism biosynthesis activity' were upregulated in clone 595 (Figs 6A-C). Even though
348 genes in the 'response to abiotic stress' GO term were mostly downregulated in 595, a number of DEGs
349 in the same category were also upregulated, indicating for a possible deregulation of stress responses in
350 berries of this clone (Fig. 6, Tables S6 and S8). In fact, a core of eight DEGs annotated in this category were
351 consistently identified in all comparisons. Five out of the eight core DEGs in the 'response to abiotic stress'
352 were consistently downregulated in 595 (Table S8), including: i) one gene (Mb_P_chr12g17550 |
353 Mb_M_chr12g17200 | Vitvi12g02138) coding for a small heat shock protein (sHSP) assigned to the three
354 enriched 'response to abiotic stress' child categories; ii) one gene coding for an ortholog of the DELLA
355 protein GAI1 (Mb_P_chr01g05780 | Mb_M_chr01g05800 | Vitvi01g00446) assigned to 'response to
356 osmotic stress' and 'response to oxygen-containing compound' child categories; and iii) three genes that
357 coded for a SRM1 transcription factor (Mb_M_chr10g18440 | Mb_P_chr10g18110 | Vitvi10g01533), a
358 Scarecrow-like protein 3 (Mb_P_chr06g12050 | Mb_M_chr06g12350 | Vitvi06g01133) and a C2H2-type
359 domain-containing protein (Mb_M_chr01g09990 | Mb_P_chr01g10150 | Vitvi01g00845) assigned to the
360 'response to oxygen-containing compound' child category. On the other hand, three of the 'response to
361 abiotic stress' core genes were consistently upregulated in 595 (Table S8). More precisely, two genes
362 coding for sHSPs (Mb_M_chr13g05600 | Mb_P_chr13g05620 | Vitvi13g00491 and Mb_M_chr02g00360
363 | Mb_P_chr02g00330 | Vitvi02g00025), and one for a ZAT11 zinc finger protein (Mb_M_chr06g05190 |
364 Mb_P_chr06g05230 | Vitvi06g01682). Overall, the function of the mentioned proteins is related to
365 protein-folding homeostasis in response to stress [38] or regulation of transcription in the gibberellic acid
366 signaling in response to osmotic stress [39].
367

368 **Figure 6**



369

370 **Figure 6.** Biological processes enrichment analysis for the three pairwise comparisons (A) 595 vs 136N, (B)
371 595 vs 505 and (C) 595vs 596. For each process two dots are shown, indicating that the same process has
372 been detected with both haplophases. Lighter dots represent the process detected with Malbec-Pru and
373 darker dots with hard outline Malbec-Mag, in some cases dots are overlapped (and seen as a single dot)
374 because the process has been detected with very similar *p-value*. Dots are colored based on its z-score
375 value, blue indicates downregulation of the process, red is for upregulation and grey indicate values close
376 to zero. The significance of the detected processes is expressed as the -Log_{10} p-value.

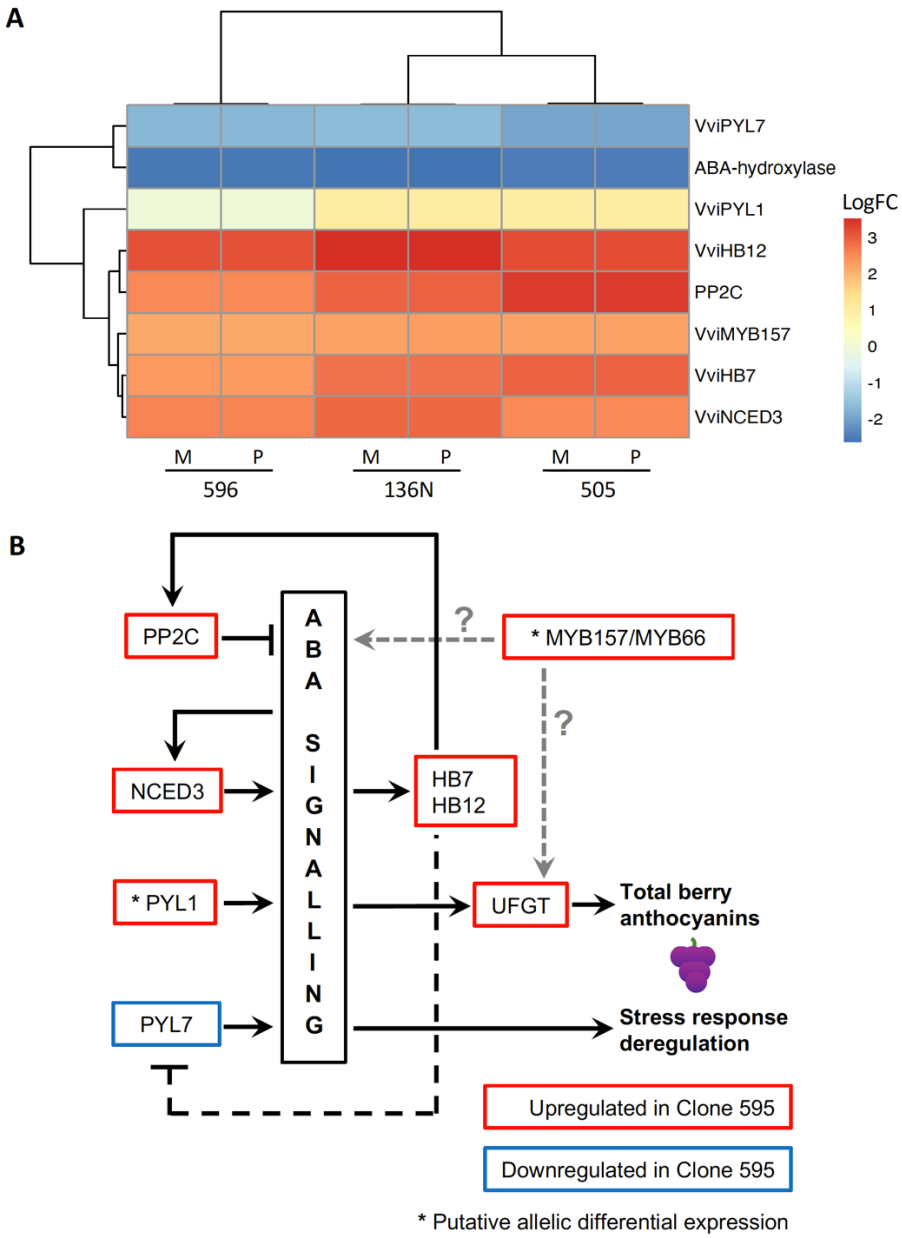
377

378 GO terms enriched with DEGs that were upregulated in clone 595 were associated with secondary
379 metabolism, more precisely to the phenylpropanoid and beta-glucoside biosynthesis (Fig. 6A-C). A core of
380 four DEGs assigned to these categories was consistently upregulated in the three pairwise comparisons
381 and with both haplophases as reference (Tables S6 and S8), including i) two genes that were assigned to
382 the 'phenylpropanoid biosynthesis' category, one coding for an Ultraviolet-B photoreceptor (UVR8-2)
383 gene (Mb_M_chr13g04770 | Mb_P_chr13g04840 | Vitvi13g00350) and another coding an F-box/kelch
384 repeat protein (Mb_M_chr05g19550 | Mb_P_chr05g19310 | Vitvi05g01703) (Table S8) and ii) two genes
385 that were assigned to 'beta-glucoside biosynthesis' category, one annotated as putative UDP-
386 glycosyltransferase in the reference gene catalogue (https://integrape.eu/resources/),
387 (Mb_P_chr12g12840 | Mb_M_chr12g12580 | Vitvi12g01718) and a *VviUFGT* gene (Mb_P_chr16g02060
388 | Mb_M_chr16g01980 | Vitvi16g00156) encoding the UDP-glucose:anthocyanidin:flavonoid
389 glucosyltransferase (UFGT) enzyme, that limits anthocyanin accumulation in grape berries [40]. Another
390 relevant gene in the anthocyanin biosynthetic pathway, *VvAOMT* (Mb_P_chr01g17010 |
391 Mb_M_chr01g16440 | Vitvi01g04438), was upregulated in the 595vs136N and 595vs596, but not in the
392 595vs505 comparison (Table S6).

393 Among the hundreds of detected DEGs some have key regulatory functions that could be driving
394 the described gene expression pattern. Furthermore, these genes appeared in the three comparisons with
395 consistent change directions (Fig. 7). More precisely, two homeodomain-leucine zipper transcription
396 factor (TF) encoding genes, *VviHB7* (Mb_P_chr15g12080 | Mb_M_chr15g11550 | Vitvi15g00912) and
397 *VviHB12* (Mb_P_chr02g02730 | Mb_M_chr02g02810 | Vitvi02g00228), were upregulated in 595 in all
398 comparisons (Fig. 7A). Orthologs of these genes play a role in stress-responsive co-expression networks
399 mediated by abscisic acid (ABA) signaling [41]. In fact, several ABA-related DEGs were consistently identified
400 here in all comparisons, including the upregulation in Malbec 595 of the ABA biosynthesis gene *VviNCED3*
401 (Mb_P_chr19g13140 | Mb_M_chr19g12630 | Vitvi19g01356), coding for a 9-*cis*-epoxycarotenoid
402 dioxygenase that is limiting for ABA biosynthesis, and one PP2C gene (Mb_P_chr16g13850 |
403 Mb_M_chr16g14410 | Vitvi16g01985); along with the downregulation of the *VviPYL7* ABA receptor
404 (Mb_P_chr15g13250 | Mb_M_chr15g13250 | Vitvi15g00997) and an ABA-8 hydroxylase coding gene
405 involved in ABA catabolism (Mb_P_chr18g15370 | Mb_M_chr18g15400 | Vitvi18g01304) (Fig. 7A).
406 Another ABA receptor, *VviPYL1* (Mb_P_chr02g09580 | Mb_M_chr02g09810 | Vitvi02g00695), was
407 upregulated in 595vs136 and 595vs505, whereas read counts were also higher in 595 when compared to
408 596 despite not being classified as DEG in the 595vs596 comparison (Fig. S7A and B). In addition,
409 *VviMYB157* (Mb_P_chr17g10170 | Mb_M_chr17g09800 | Vitvi17g00822), a homolog of the *WEREWOLF*

410 / *AtMYB66* TF, was upregulated in 595 in all comparisons (Fig. 7A). The allelic frequency of one SNV at
411 *VviMYB157* suggested downregulation of the Magdeleine allele in the other three clones (Fig. S7C and D).
412 Overall, the detected transcriptome variation during veraison correlated with the higher TA
413 concentrations on mature berries of 595 (Figs 3, 6 and 7).

414 **Figure 7**



415

416 **Figure 7**. (A) Heatmap showing the change direction of DEGs with regulatory functions associated to the
417 ABA signaling pathways. Log fold change of each gene is expressed in relative terms to 595. Results
418 obtained with both haplophases (M = Malbec-Mag, P = Malbec-Pru) for the three pairwise comparisons
419 are shown. (B) Proposed model based on the identified DEGs for the origin of the higher berry anthocyanin
420 content in Malbec clone 595**.** Clonal variant phenotypes observed in 595 berries are indicated in bold.
421 Genes upregulated in clone 595 compared to the other three are represented in red boxes, whereas the
422 downregulated gene is in a blue box. Arrow and flat end lines indicate for positive and negative regulation.

423 Relations that are suggested but not proven are indicated with dashed lines. Genes with possible allele-
424 specific overexpression in Malbec 595 suggesting for *cis*-acting regulatory somatic mutations are labelled
425 with asterisks as candidate genes triggering all other downstream responses.
426
427 **4. Discussion**

428 One of the main challenges for accurate genomic and transcriptomic clonal variation analyses in
429 grapevine is to start with an appropriate reference genome for the cultivar of interest. Here we present
430 the haplotype-resolved and annotated genome for cultivar Malbec, composed of the haploid
431 complements inherited from its parental cultivars Prunelard and Magdeleine. We also employed Malbec
432 diploid assembly as reference genome to perform a haplotype-aware evaluation of the transcriptomic
433 differences underlying clonal phenotypic variation.

434 **The use of parental reads improved the phasing and evaluation of the assembly process.**

435 A haplotype-resolved genome assembly for Malbec was obtained in a trio-based approach
436 applying prior knowledge of Malbec pedigree [27]. Trio phasing and haplotype-aware deduplication yielded
437 two haplophases that were close to the expected genome size of ~475 Mbp and with low duplicated
438 BUSCO genes (Table 1B). In this regard, our results were comparable to the obtained with the Merlot
439 genome assembly, also based on trio binning [19]. Trio binning approach differs from the applied for
440 assemblies of other grapevine cultivars generated as pseudo-haplotypes, which exhibit inflated primary
441 assemblies compared to the haplotigs [6,9,11–14,42]. Nonetheless, recently developed bioinformatic tools help
442 to improve the proper assignment of primary contigs to the two haplophases in those cases [16].

443 We analyzed the assembly consensus and the phasing accuracy based on *k-mers* from the
444 parentals' short-reads whole genome sequencing (Figs 1A and B). This had only been reported before in
445 grapevine for the PN40024 reference genome [2]. The estimated assembly accuracy for Malbec (QV > 40)
446 indicates that error *k*-mers in the haplophase assemblies are < 0.00001%, which is slightly above the
447 accuracy (QV=37.43) reported for the reference genome [2]. Parental-specific *k-mers* (*hap*-mers) from short
448 read sequencing of Malbec parents identified very few blocks that were apparently switched (<0.025%,
449 Fig. 1 and Table 1C), in some cases these might be true homozygous regions in Malbec that were
450 misassigned as switched. These results emphasize the importance of grapevine pedigree analysis and the
451 value of using the parental reads, considering that some commercially relevant cultivars still lack this
452 information despite the invested efforts (e.g. Nebbiolo) [43].

453 The diploid assembly also enabled a genome-wide study of the variation between the two Malbec
454 haplotypes that showed ca. 25% of polymorphic regions (Fig. 2A and Table S3). The fraction of Malbec
455 haplophases affected by SVs (12.1%) is much higher than in Zinfandel (6.5%) [13], but lower than in Nebbiolo
456 (17.8%) [9] and close to Chardonnay (15.1%) [6], with the caveat that different SV evaluation methodologies

457 were used, and that a thorough validation of reported SVs was performed only for Nebbiolo [9]. In this
458 direction, the percentage of genes overlapping with SVs between the two haplotypes in Malbec (29%,
459 Table S3C) was in the range to the reported for Nebbiolo (26.2%). Overall, the fraction of the Malbec
460 genome and genes affected by SVs is in agreement with the general observation that grapevine genomes
461 can be unbalanced and that hemizygosity is widespread [6].

462 Almost half of Malbec genome consists of repetitive sequences and transposable elements (TE),
463 with long terminal repeats (LTRs) representing the most abundant class of TEs (Fig. 1C and Table S4). This
464 is in agreement with other grapevine cultivars [9,18,42,44], including the T2T assembly of the PN40024
465 genotype [4]. Annotated gene number was similar in the two Malbec haplophases (Tables 2 and 3), and to
466 the PN40024 v4 [2] and v5 (T2T) [4] versions of the reference genome. Nonetheless, close to 6% of genes did
467 not have a counterpart mapping on the other haplophase. Furthermore, an enrichment of the resveratrol
468 biosynthetic pathway in Malbec-Mag specific genes was observed (Fig. S3), suggesting for a putative
469 higher contribution of this haplotype to a key secondary metabolite in grapevines, both from an industrial
470 and biological perspective [45].

471 Orthologous genes between the two Malbec haplophases were mostly located in syntenic regions,
472 whereas many genes assigned to the "Unknown" chromosome in PN40024 v4 could be placed on specific
473 chromosomes across Malbec haplophases (Fig. 2B and Fig. S4). This suggests that the Malbec diploid
474 assembly is improved in contiguity and representation of previously difficult-to-assemble regions.
475 Moreover, when compared to the T2T HiFi sequencing-based assembly version of the PN40024 [4], Malbec
476 haplophases have shown both strong structural correspondence and a congruent placement of most
477 centromeres along the scaffolded pseudomolecules (Figs. 1C, S1C and S1D).

478 **Clone transcriptome variation in ripening berries could explain differences in berry anthocyanins**
479 **content at harvest**

480 We searched for consistent phenotypic and transcriptomic variation among Malbec clones that
481 were cultivated in the same vineyard for over 20 years. Four clones were selected for analysis, based on
482 their consistent variation in berry composition during 2018 and 2019 seasons (Fig. 3 and Fig S3A).
483 According to a previous intra-cultivar genetic variation analysis, the four selected clones represent
484 different derived genotypes from the same clonal genetic lineage (Group-Fr), with clones 595, 596, 136N
485 and 505 belonging to genotypes I, B, H and E, respectively [35]. Clone 595 had reproducible higher levels of
486 total polyphenols and anthocyanins berry content at harvest (Fig. 3), whereas it also showed a clear
487 transcriptome differentiation from the other clones (Figs 4 and 5). Functional enrichment of DEGs
488 indicated for abiotic stress response deregulation in 595 compared to the other three clones (Fig. 6; Tables
489 S6 and S8). On the other hand, phenylpropanoid and anthocyanin biosynthesis genes were consistently

490 upregulated in 595 (Fig. 6; Tables S6 and S8), in agreement with the observed phenotypic differences. In
491 our analysis, statically enriched processes (Fig. 6) correlated with the phenotype of the mature berry (Fig.
492 3). Similar relations were previously observed among Malbec clones diverging in berry anthocyanin
493 content, when the transcriptional activity of post-veraison berries was studied [32].

494 Particularly, we identified ABA-related DEGs in Malbec clone 595 that might be major drivers of its
495 global transcriptional profile and phenotype (Fig. 7A). ABA treatments have been reported to induce the
496 gene *VviNCED3* and PP2C encoding genes and also to downregulate ABA receptor genes in different
497 grapevine tissues including berry skin [46]. Similarly, *VviNCED3* and an homolog of the ABA-inducible
498 repressor PP2CA phosphatase [47] were upregulated in clone 595 compared to the other three, whereas
499 the *VviPYL7* receptor gene was downregulated (Fig. 7A, Table S6). *VviHB7* and *VviHB12* class I HD-ZIP TF
500 genes were also upregulated in 595. Their homologs in *Arabidopsis thaliana* have been implicated in a
501 negative feedback loop induced by ABA and repressing ABA receptor genes [48]. These expression patterns
502 suggest that ABA over-accumulation or hypersensitivity could be a feature of clone 595. However, even
503 though ABA usually represses *PYL* receptor genes [46], *VviPYL1* was upregulated in 595 (Figs 7A, S7 and Table
504 S6). We hypothesize that somatic mutations in *VviPYL1* of clone 595 could trigger its over-expression and
505 the hyperactivated ABA gene expression responses compared to other clones. To test this hypothesis, we
506 inspected RNA-seq reads alignments for possible allele-specific over-expression of *VviPYL1* in 595, that
507 could indicate the presence of a *cis*-acting somatic mutation in a regulatory region of this gene triggering
508 its overexpression. However, both the RNA-seq reads and haplophase assembly comparison by SyRI
509 support that Malbec is homozygous for *VviPYL1*, making it difficult to identify allele-specific differences
510 among clones (Fig. S7). Therefore, further functional analyses are required to identify the underlying
511 reasons explaining clone 595 differential response to ABA signaling.

512 Both the higher accumulation of anthocyanins and the differential expression of abiotic stress
513 genes detected in clone 595 (Fig. 6, Tables S6 and S8) could be explained by an alteration in ABA
514 homeostasis (Fig. 7B) [49]. The overexpression of *VviUFGT* in 595 compared to the other three clones (Tables
515 S6 and S8) could ultimately result in high berry anthocyanin content of this clone, since UFGT plays a
516 critical role in the synthesis and accumulation of anthocyanins in grape berries [50]. ABA treatment induces
517 *VviUFGT* in grape berry skins, increasing its anthocyanin content [46]. Therefore, ABA signal hyper-activation
518 in 595 could potentially over-activate the flavonoid pathway leading to the higher phenolic character in
519 595 fruits (Fig. 7B). Transient overexpression of *VlPYL1* in grape berries induces ABA-responsive genes like
520 NCEDs, but also *VviUFGT* [51], suggesting that the overexpression of *VviPYL1* in 595 could be an upstream
521 trigger of both the ABA hypersensitivity and the high berry anthocyanin content in this clone (Fig. 7A-B).
522 In addition, the putative allele-specific overexpression of *VvMYB157/MYB66* in 595 (Fig. S7) could be

523 involved in the deregulation of the ABA signaling and the upregulation of the anthocyanin biosynthesis
524 pathways characterizing this clonal line (Fig. 7B). *VvMYB157* might be particularly relevant in Malbec
525 clonal anthocyanin content variation, since this cultivar is homozygous for *VvMAYBA1* and *VvMYBA2* loci
526 and no evidence of differential gene expression was detected here (Fig. S7). A previous study of mature
527 berries also reported the absence of variation for *VvMYBA1* expression among Malbec clones with
528 different anthocyanins content [32]. On the other hand, overexpression of a *MYB157/MYB66* homolog in
529 the fruit skin of Russet apple somatic variants was associated with activation of ABA responses and the
530 phenylpropanoid pathway [52]. Our results warrant further functional genomic analyses to confirm the role
531 of *VvMYB157* in the regulatory mechanism responsible of the high phenolic content found in Malbec clone
532 595.

533 **Phasing haplotypes helps to identify haplotype-specific gene expression.**

534 Analysis of transcriptomic data with either of the two Malbec haplophases as reference retrieved
535 similar results, however some differences were observed. In this direction, 0.58% more transcripts aligned
536 when Malbec-Pru was used as reference than with Malbec-Mag. This is in line with the fact that Malbec-
537 Pru assembly turned 0.06% larger in size (Table 1B), accounting with 0.78% more protein coding
538 annotated genes (Table 2) and 0.92% more transcripts predicted (Table 3), than Malbec-Mag. The analysis
539 of cultivar Merlot haplotype-resolved genome also reported a similar difference, regarding annotated
540 gene number between the two haplophases [19]. In our analysis, the number of detected DEGs was 2.7%
541 (±1.2% SD) higher with Malbec-Pru than with Malbec-Mag, on average for the three comparisons
542 involving Malbec clone 595 (Fig. 5A and 5B). This observation is in agreement with the expected by chance
543 based on the assembly and annotation features of both haplophases. Regarding the haplotype-specific
544 DEGs, we found that six and 11 genes were consistently detected in all pairwise comparisons only with
545 Malbec-Mag and Malbec-Pru (Fig. 5D), from which four and six had no ortholog in the other haplophase
546 (Table S7), respectively. There is indeed a significant enrichment of Malbec-Mag putative hemizygous
547 DEGs that were unmapped in the other haplophase, in contrast there was underrepresentation of Malbec-
548 Pru DEGs that were unmapped in Malbec-Mag. Furthermore, two of the haplotype-specific DEGs detected
549 only with Malbec-Pru coded for fruit development relevant functions. More precisely, an acyltransferase
550 protein involved in fruit ripening (Mb_P_chr11g08440 | Vitvi11g04184) that was upregulated in clone
551 595, and an F-box like domain protein involved in the regulation of phenylpropanoids biosynthesis
552 (Mb_P_chr08g18140 | Vitvi08g01787) that was downregulated (Table S7). Visual inspection of RNA-seq
553 alignments on Malbec-Pru reference indicate putative hemizygosity for *Mb_P_chr11g08440*. In
554 agreement, it has been shown that phased assemblies enhance the capture of molecular variation
555 between clonal variants in highly heterozygous genomic regions, and that allele-specific expression is also

556 associated to clonal phenotypic variation in other woody crops like citrus [53]. On the other hand, two
557 consistent heterozygous SNPs in *Mb_P_chr08g18140* indicate that Malbec might not be hemizygous for
558 this locus, even though its ortholog was not predicted in Malbec-Mag haplophase (Fig. S7).

559 Overall, our findings are in agreement with the general observation that the genomes of grapevine
560 cultivars are unbalanced and that hemizygosity is widespread [6,10]. However, caution must be taken since
561 miss-assembly and/or miss-annotation might be overestimating the reported putative hemizygosity. In
562 this direction, the diploid assembly presented here could be useful not only to further corroborate
563 putative allele-specific expression differences in highly heterozygous genomic regions[53]; but also for
564 future introgression experiments, for which accounting with phased assemblies and transcript
565 annotations has shown to be pivotal in grapevines [54].

566 **Conclusions**

567 We have assembled the diploid genome of Malbec in two haplophases of the expected size and
568 phase. Similar results were obtained in our RNA-seq comparison of Malbec clonal variants with either
569 haplophase used as reference, nonetheless some haplotype-specific gene expression differences could be
570 spotted thanks to the phased genome assembly. In this regard, genes with no ortholog annotated in the
571 other haplophase of the Malbec assembly were enriched in relevant biological processes for grapevines,
572 whereas a significant number of all the detected DEGs with Malbec-Mag had no ortholog in Malbec-Pru.
573 SVs between the two assembled haplophases also suggested for hemizygosity of some DEGs contrasting
574 between the studied clonal variants. Finally, we proposed that a somatic variant causing the
575 hyperactivation of ABA signaling and the differential activity of a MYB transcription factor in a Malbec
576 clone, might be responsible for its increased berry anthocyanin content.

577 **5. Materials and Methods:**

578 **6.1 Genome *de novo* assembly and annotation:**

579 **Plant material:** For genomic sequencing and assembly, samples were collected from the Malbec
580 clone accession 136N, representing a clonal lineage with short propagation history in Argentina (i.e.,
581 Group-Fr) [35]. Accession 136N is implanted at the Vivero Mercier Argentina experimental vineyard located
582 at Lujan de Cuyo, Mendoza, Argentina (-33.09° S, -68.87° W). Samples of Malbec parental cultivars [27] were
583 requested to the *Institut Français de la Vigne et du Vin* (VInopôle Sud-Ouest, France), corresponding to
584 clone 1232 of Prunelard (reference number: 08-04-00004) and the unique available accession of
585 Magdeleine Noire des Charentes (reference number: 08-04-00075).

586 **DNA extractions, library preparation and genomic sequencing:** Single node wood samples of the
587 mentioned accessions were hydroponically grown, until the development of roots and shoots with young
588 leaves. For PCR-free Illumina sequencing, DNA was obtained from roots for all three cultivars. For the

589 parental cultivars solely roots were used, because this tissue derives from the L2 meristem cell layer (as
590 gametes do in grapevine) and its genotype should be closer to the inherited by Malbec, in case chimeric
591 somatic mutations existed. For Prunelard and Magdeleine, DNA extractions were performed with the
592 DNeasy Plant mini kit (Qiagen). For Malbec, high molecular weight (HMW) DNA was obtained with the
593 Nanobind plant nuclei kit (Circulomics-PacBio), either from roots or young leaves. Leaves were used for
594 long read sequencing because it was the only tissue yielding the amount of HMW DNA needed for library
595 preparation. In brief, nuclei were isolated from the source tissue according to Workman *et al*. [55]. Isolated
596 nuclei were the input for HMW DNA extraction with the Nanobind nuclei kit following manufacturer's
597 instructions. The HMW DNA was sheared with two strokes using a 26 G needle in a 1 mL syringe, with an
598 average fragment size of 70 kb.

599 Library preparation and genomic sequencing were performed at the Max Planck Institute for
600 Biology (Tübingen, Germany). A large insert PacBio gDNA library was prepared from the sheared HMW
601 DNA of Malbec leaves using the SMRTbell Express Template Preparation Kit 2.0. The library was size-
602 selected for >30 kb using BluePippin with a 0.75% agarose cassette (Sage Science) resulting in a 48 kb peak
603 size final library. Five PacBio Sequel I SMRT cells were loaded with the same SMRTbell library and were
604 sequenced with the PacBio Sequel I sequencer. Illumina PCR-free sequencing libraries were produced
605 from gDNA of roots, using the NxSeq® AmpFREE Low DNA Library Kit (Lucigen) and sequenced to obtain
606 paired-end short reads (150 bp), as described in Rabanal *et al.* [56].

607 **Genome de novo assembly:** The assembler Canu v1.8 [57] was employed for *de novo* genome
608 assembly using the TrioBinning module [15]. Here, parental-specific *k-mers* from the short-reads of the
609 parental cultivars (Prunelard and Magdeleine) were used as template to split long reads of the child
610 (Malbec) into the haploid complements. After adapters trimming with cutadapt v2.3 [58], Magdeleine and
611 Prunelard short-reads were loaded into Canu assembler as *-haplotype1 and -haplotype2*, respectively,
612 whereas Malbec PacBio CLR reads were loaded with option *-pacbio-raw*. The child haplotypes inherited
613 by Malbec cultivar from each parent were separately assembled from the PacBio read partition,
614 theoretically without the interference of inter-haplotype variation [15]. Canu was run in all cases with option
615 *genomeSize=490m*. No contaminant contig from non-plant genera or phyla was found from the draft
616 assemblies by running blobtools v.1.1.1 [59], comparing to the non-redundant (nr) protein database using
617 diamond v2.0.8.146 [60]. The post-processing of the raw assembled contigs representing each haplotype of
618 Malbec (hereafter: haplophases) involved a first round of polishing based on the long reads raw data using
619 the ArrowGrid wrapper [57,61] of Arrow (http://github.com/PacificBiosciences/GenomicConsensus/)
620 consensus framework, within the SMRT Analysis Software (http://github.com/PacificBiosciences/SMRT-

621 Analysis). ArrowGrid was run on the primary assembly output of Canu with the corresponding haplotype-
622 binned Malbec PacBio CLR reads for each haplophase, separately.
623 The assembly size metrics were tested along with the quality of the phasing process, to reduce
624 haplotype switch errors [15,62]. This was performed with Merqury v1.3 [62], based on the count of haplotype
625 specific *k-mers* (i.e., *hap-mers*) from the short reads of Magdeleine and Prunelard as compared to Malbec
626 short reads and the assembly *k-mers*. *Hap-mers* were used to determine phase blocks, defined as sets of
627 markers inherited from the same parental haplotype [62]. Each assembled haplophase genomic composition
628 was tested using the default values, allowing 100 marker switches in a 20 kb bounded range (0.5% switch
629 rate). Afterwards, complete contigs showing haplotype switch errors all along the contig and no-matches
630 to the own haplotype in the Merqury output, were removed using a custom script (based on Unix/Linux
631 *awk* and *grep* commands). Then, another haplotype-aware deduplication was performed based on
632 minimap2 [63] alignments of one haplophase assembly to the other using *-x asm5* alignment option. Average
633 similarity between haplotypes was computed for each contig from the *dv* value of the output PAF file.
634 Candidate haplotype-switched contigs with high similarity to the other haplotype (average dv < 0.005)
635 were extracted and used as reference to align the rest of contigs from the same haplophase with
636 minimap2 (*-ax asm20 -secondary=no*). To check for possible duplicated contigs from that alignment, the
637 coverage utility from samtools v1.10 [64] was run on the resulting *bam* file. High similarity contigs to the
638 other haplotype were removed if they showed >70% bases covered, with depth ≥1 and mean depth of
639 coverage >0.7 from the other contigs in the same haplophase. The resulting haplophases were further
640 deduplicated with purge_dups v1.2.5 [65], read depth was used to remove duplicated contigs, by mapping
641 the haplotype-binned Malbec PacBio reads to the corresponding haplophase assembly separately. The
642 purge_dups command get_seqs was run with *-s* option to break contigs instead of adding Ns at internal
643 breaks.
644 To perform miss-assembly corrections, RaGOO v1.1 [66] was used to split contigs that may generate
645 spurious structural variations using the PN40024 12X.v2 [3] as reference, with options *-b -r 300000 -c*
646 *15000000 -d 30000000 -g 100 -s -C*. Afterwards, another round of assembly consensus correction based
647 on long reads data was performed with ArrowGrid, in this case using all the Malbec PacBio CLR subreads
648 to correct deduplicated contigs of the two haplophase assemblies concatenated. To overcome the high
649 error rate of long reads, PCR-free Illumina short-reads of Malbec were used to polish the concatenated
650 haplophases contigs, using Pilon v1.23 [67] with options *--fix snps,indels --minmq 3*. Then, haplophases were
651 split according to their original contig IDs; and another round of haplotype-switch error detection was
652 performed based on Merqury *hap-mer* analysis.

653 A final manual deduplication was conducted based on the gene content of the contigs spotted as
654 haplotype switches by Merqury, and on BUSCO duplicated genes identified using the orthologue plant
655 core genes database (eudicotyledons_odb10) and BUSCO v3.0.2 [68]. When ≥2 BUSCO genes with different
656 IDs were duplicated in the same order in another region of the same haplophase assembly, the duplicated
657 copy spotted by Merqury as haplotype-switch error was considered as an artifactual haplotype
658 duplication leaked to the wrong haplophase. Therefore, while the duplicated copy with correct haplotype
659 assignment was kept in the final assembly, the duplicated and switched counterpart region was removed
660 from the affected contig. This removal of duplicated and switched blocks was performed by running the
661 *complement* and the *getfasta* applications of bedtools v2.27.1, using the contigs assembly (*fasta* file) and
662 the duplicated coordinates (*bed* file). The boundaries of the coordinates of the duplicated regions to be
663 removed were refined according to minimap2 alignments, with option *--secondary=no -x asm20 --MD* of
664 each pair of duplicated contigs to the other haplophase contigs assembly. When no duplicated BUSCO
665 genes were observed, then the contig spotted as switched by Merqury was consider as a likely
666 homozygous region and was kept in the final assembly. In those cases, no corresponding haplotype
667 switched block to be used for phase correction was present in the other haplophase assembly. After
668 running the deduplication pipeline, another round of RaGOO miss-assembly correction was run as
669 described before leading to only one break between chr2 and chr7 in Prunelard haplophase and none in
670 Magdeleine haplophase.
671 RagTag v1.0.1 [69], using the PN40024 12X.v2 assembly as reference (including the 19 scaffolded
672 chromosomes of grapevine, as well as the plastid and mitochondrial genomes), was employed for the final
673 scaffolding of the deduplicated haplophases into the pseudo-molecules that represent the 19
674 chromosomes of grapevine haploid complement. The ragtag.py script was run with options *--remove-*
675 *small -f 20000 -q 10 -d 500000 -i 0.5 -a 0.05 -s 0.3 -u --mm2-params "-x asm20*. Seqkit [70] was used to obtain
676 assembly summary statistics of size and contiguity. SyRI v1.3 [36] was used to quantify the synteny and
677 diversity between Malbec assembly haplophases. To run SyRI, the two haplophase assemblies were
678 aligned with nucmer using options *-maxmatch -l 100 -c 500 -b 500* and delta-filter with options *-m -l 88 -l*
679 *100*. Nucmer and delta-filter were used from mummer v4.0.4 [71]. Plots were produced with plotsr from
680 SyRI. Completeness and duplication of contig and scaffolded assemblies of Malbec and of the PN40024
681 reference genome v4 and v5 versions was assessed running BUSCO v5.6.1[68] on euk_genome_min mode,
682 and using the eudicots_odb10 (Creation date: 2024-01-08) dataset as lineage and miniport [72] as gene
683 predictor. Dot plots between the scaffolded Malbec haplophases and the 12Xv2 and v5 reference genome
684 assembly versions were obtained running minidot on assembly-to-assembly alignments produced with
685 minimap2 using *-x asm20* option.

686 **Genome assembly annotation:** Prior to gene annotation, EDTA v1.9.6 [73] was employed for the
687 annotation of repeats and transposable elements (TEs), and to soft-mask the repeated regions. A repeat
688 library was generated by running the EDTA.pl script with the options *--sensitive 1 --anno 1* and non-TE
689 plant proteins were excluded from the repeat library. This analysis was run on the concatenated assembly
690 (Malbec scaffolded haplophases) and separately, on the discarded contigs of each of the two haplophases
691 that were not included in the final assembly after the deduplication pipeline. Firstly, blastx from blast
692 v2.2.29 was run with options *-evalue 1e-10 -num_descriptions 10* to identify overlaps between the repeat
693 library and a curated plant protein database obtained from the MAKER-P manual [74]. Then, the overlaps
694 were excluded from the EDTA repeat library by running ProtExcluder.pl [74]. Finally, the curated repeat
695 library was used to annotate TEs in the diploid assembly by running again EDTA.pl with options *--step*
696 *anno --sensitive 1 --anno 1*.
697 For gene annotation a pipeline combining gene evidence obtained from different sources based
698 on EVidenceModeler v1.1.1 [75] was implemented, using as input genome the soft-masked Malbec diploid
699 assembly (concatenated and scaffolded haplophases). Here, *de novo* assembled transcripts, *ab initio*
700 predictions and gene lift-over data were combined, and obtained as follows.
701 Transcriptome assembly was produced from an RNA-seq dataset obtained from berry pericarp of
702 eight different Malbec clones and their biological replicates. Clones 225, 228, 59 and 53, were sampled
703 during 2016 and employed solely for the annotation pipeline (Table S1C). For these samples total RNA
704 extractions were performed from mature berries pericarp using a CTAB based protocol [76]. Three to four
705 berries were grounded with liquid nitrogen discarding the seeds, to obtain the desired amount of RNA
706 (>3,5 ug). Afterwards, a column-based purification step was conducted with the kit *SV total RNA isolation*
707 *System* (Promega). Extracted RNA integrity, concentration and purity were tested in 2% agarose gel and
708 by spectrophotometry (AmpliQuant AQ-07). Samples were shipped to CRG facilities (Barcelona, Spain) for
709 library preparation and sequencing. Library preparation was performed with a TruSeq Stranded mRNA LT
710 Sample Prep Kit (Illumina) and included rRNA depletion using the Ribo-Zero kit (Illumina). Sequencing was
711 performed with Illumina HiSeq 2500 and paired-end fragments of 125 bp were obtained. The four clones
712 employed for transcripts *de novo* assembly, were also used for the clonal transcriptomic variability
713 analysis (see that section for further details on total RNA obtention). Overall, a batch of 22 RNA-seq
714 samples was employed for transcripts *de novo* assembly, adding ~150 Gb of raw sequence (Table S1C).
715 The Illumina sequencing adaptors and low quality reads were trimmed with TrimGalore v0.6.4 (Babraham
716 Bioinformatics) that includes cutadapt v2.3 for read trimming and FastQC [77] for sequence quality checks.
717 Two rounds of TrimGalore were run, the first to remove adaptors and low-quality reads with options *-q*
718 *20 --paired --length 30 --max_n 1 --trim-n* and the second to remove polyA tails and run the final QC with

options *--paired --length 30 --polyA --fastqc*. For transcriptome assembly, the trimmed RNA-seq reads were aligned to the diploid Malbec genome assembly using HiSat2 v2.2.1 [78] with options *--rna-strandness RF --dta-cufflinks --max-intronlen 20000*. The obtained alignments (*bam* files) were merged into a single *bam* using the merge application of samtools v1.10 [64] and assembled into potential transcripts with StringTie v2.2.1 [79], using the options *-t -c 2 -f 0.1 -s 4.75 -M 0.95 -m 80 -j 2 -A --rf*. Starting from a genome-based transcript structure obtained with StringTie, the software TransDecoder v5.7.0 (Haas, BJ. https://github.com/TransDecoder/TransDecoder) was used to identify the likely coding sequences (CDs), adding them to the annotation files.

On the other hand, *ab initio* gene predictions were performed with Augustus v3.5.0 [80]. Before running Augustus, Braker2 v2.1.6 [81] was run for the Malbec diploid assembly using as strand-aware hints the *bam* files produced by HiSat2 from the malbec RNA-seq data. This was performed to generate the retraining configuration file to be used as input in the *--species* option of Augustus. The EDTA soft-masked diploid assembly was used as input genome to run both Augustus and Braker2. To run Augustus, *extrinsic M.RM.E.T* information configuration was used as described in the *extrinsic.cfg* file (File S1) and the UTR prediction was inactivated *--UTR=off*. The hints input for Augustus was a *gff* file with exons annotated by TransDecoder from the StringTie transcriptome assembly of the merged Malbec RNA-seq (*gtf* file). Gene models resulting from Augustus were kept as the final *ab initio* gene predictions, but gene models were only kept for loci that showed some overlap with gene models that were also predicted in both, the final output of Braker2 and the output of Genemark that Braker2 uses for its retraining. The intersection between the three gene prediction version *gff* files was obtained by running a custom script based on *comm*, *grep* and the *intersect* -c command of bedtools v2.27.1 [82].

Also, a lift-over of the PN40024.v4 gene annotation to the Malbec haplophases assemblies was conducted with Liftoff v1.6.1 [83], with options*: -a 0.5 -s 0.5 -copies -sc 0.99 -exclude_partial -polish*. This included a final polishing step and transferred genes were only kept if similarity and coverage were >50% for the primary gene copy, whereas for secondary copies the similarity threshold was increased to 99%. The PN40024.v4 gene annotations of the improved grapevine reference genome assembly [2] were retrieved on 15.11.2021, from the Integrape Cost Action CA17111 portal (https://integrape.eu).

The weighed consensus gene models for Malbec diploid and unmasked genome assembly were obtained with EVidenceModeler, using the evidence obtained from the described sources. The final combination of evidence inputs and weights was selected after testing more than 50 EVidenceModeler runs and comparing the obtained gene models to the deeply curated PN40024.v4 [2]. For EVidenceModeler, the RNA-seq transcriptome assembly obtained with StringTie was used as TRANSCRIPT input with the maximum weight = 4, the lifted-over of PN40024.v4 reference annotations obtained with Liftoff was

752 loaded as OTHER PREDICTION (weight = 3), and the gene model produced by Augustus was used as AB
753 INITIO PREDICTION with the minimum weight = 1.
754 Finally, the *annotation update* function of PASA v2.4.1 [75,84] was used to polish exon boundaries
755 and to update the 5' and 3' UTRs to the genes structures and alternative isoform annotations, previously
756 predicted by EVidenceModeler. Here, the input transcripts were the ones assembled with StringTie from
757 the Malbec RNA-seq samples and the transcripts annotated in the primary (REF) and alternative (ALT)
758 haplotypes of the PN40024.v4. All these transcripts were cleaned from polyA, low quality or vector
759 sequences using the *seqclean* script of PASA. Cleaned transcripts were aligned to the Malbec diploid
760 assembly with Blat [85] and Gmap [86] aligners using the *Launch_PASA_pipeline.pl* script with –
761 *transcribed_is_aligned_orient* option and the following options in the alignAssembly.config file:
762 validate_alignments_in_db.dbi:--MIN_PERCENT_ALIGNED=90; validate_alignments_in_db.dbi:--
763 MIN_AVG_PER_ID=95; subcluster_builder.dbi:-m=50; run_spliced_aligners.pl:-N=5;
764 validate_alignments_in_db.dbi:--MAX_INTRON_LENGTH=25000. Two consecutive rounds of PASA
765 updates from the aligned clean transcripts were run using again the *Launch_PASA_pipeline.pl* script, by
766 loading firstly the EVidenceModeler annotations and secondly the updated annotations resulting from
767 the first round. Genes were renamed according to assembly chromosome coordinates using the script
768 retrieved from https://github.com/andreaminio/AnnotationPipeline-EVM_based-DClab [87].
769 To perform QC at each annotation step, the general metrics of the predicted gene models were
770 computed using the *agat_sp_statistics.pl* script from AGAT v0.8.0 [88]. Transcripts and translated proteins
771 were extracted from *gff* annotation files with Gffread v0.11.7 [89] and the general stats were computed
772 with Seqkit v0.12.0 [70]. Completeness and duplication levels for the annotated protein sets were obtained
773 running BUSCO v5.6.1 [68](on protein mode), using the eudicots_odb10 (Creation date: 2024-01-08) dataset
774 lineage.
775 Functional annotation to predict gene ontology (GO) classes and gene functional description (DE)
776 was performed with Panzzer2 [90], using as input the *fasta* files containing all the protein sequences of each
777 haplophase. The correspondence between the obtained gene model predictions for each haplophase,
778 PN40024.v4 and VCostv3 annotations was performed using the proteins amino acid sequences, with
779 OrthoFinder [91]. Afterwards, GENESPACE [92] was used to evaluate the order and location of the predicted
780 gene models along the scaffolded pseudomolecules of the assembled haplophases, compared to the
781 PN40024.v4 as the most complete gene predictions for the reference genome (Table 3B). The co-location
782 of annotated genes with genome variation between the two assembled haplophases of Malbec was
783 performed using the *intersect* command of bedtools v2.27.1 to compare the final gene predictions of each
784 haplophase to the genome variation output of the comparison of the two haplophases produced by SyRI.

785 Genes overlapping with each variation feature assigned by SyRI were analyzed for enrichment in Biological
786 Process GO terms using g:Profiler [93], with the same settings described below for the functional enrichment
787 analysis of DEGs.

788 To annotate centromeres and telomeres, a search for the 107 nt monomer (CEN107=
789 AGTACCGAAAAAGGGTCGAATCAGTGTGAGTACCGAAAAATGGTAGAATCCGGGCGAGTACCGGGAAAAGGTA
790 GAATCCGTGCGAGTATCGAAAAACTGTCCGGGCG) of centromeric repeats and the 7 nt monomer
791 (TTTAGGG) of telomeric repeats was performed [94]. EDTA v1.9.6 was used to annotate centromeric
792 repeats on the Malbec diploid assembly by running the EDTA.pl command with *--sensitive* 1 *--anno* 1
793 parameters from a repeat library containing only the CEN107 monomer. Telomeric repeat positions were
794 identified on the diploid assembly using BLAST v2.2.29+ [95]. Blast alignment was run with oligonucleotide
795 query parameters (*blastn* -task *blastn-short*) using as query two consecutive copies of the 7-mer
796 monomer. Finally, the density of all the annotated elements was plotted along the assembled
797 pseudomolecules using ggplot2, by taking the coordinates of the start position of each annotated element
798 as well as of variants detected between Malbec haplophases.

799 **6.2. Analysis of the phenotypic and transcriptomic clonal variation:**

800 **Clonal phenotypic variation analysis:** A survey focused on berry composition variation was
801 performed on 27 Malbec clones, with three biological replicates, resulting in 81 plants. All analyzed plants
802 were located in the same plot at Vivero Mercier Argentina experimental vineyard (Luján de Cuyo,
803 Mendoza, Argentina), implanted since 2002. Therefore, the analyzed plants have been exposed to the
804 same cultural treatments and environmental conditions ever since. Phenotypic measurements were
805 conducted during 2018 on mature berries. For each biological replicate a random sample of 50 berries
806 was obtained and the pulp was separated from the skin. The pulp was used for analytical measures (pH,
807 total acidity and total soluble solids) and the skin was used for spectrometry analyses (total anthocyanins
808 and total polyphenols). The pH was directly measured with a Denver Instrument (UB-10, pH/mV Meter).
809 Total acidity (expressed in g/L of tartaric acid) was measured through acid-base titration with 0.1 N NaOH
810 (pH = 8.2), using 5 mL of juice diluted with 10 mL of distilled H_2O and containing bromothymol blue
811 (0.04%). Total soluble solids were measured in wt/wt of sucrose, expressed in Brix degrees (Bx), using an
812 ATAGO refractometer (PAL-1). The skins were dried and weighed to perform total polyphenols extraction
813 using a ethanol and water solution (12:88, v/v), containing 5 g/L of tartaric acid, according to Muñoz *et al.*
814 [32]. Total polyphenols (TP) and total anthocyanins (TA) were measured with a UV-VIS spectrophotometer
815 (Spectrum SP-2000UV) at wavelengths of 280 and 520 nm respectively, and expressed in AU/g of skin. The
816 same procedures were repeated for the phenotypic analyses in 2019, only for the clonal accessions
817 selected for transcriptomic analyses.

818 **Plant material selection for clonal diversity transcriptomic experiments:** Four Malbec clones
819 (595, 596, 136N and 505) were selected after a ranking-based procedure, based on phenotypic data
820 obtained for the 27 clones in 2018. The selected samples combined high, mid and low values for the
821 surveyed traits, except for sugar content trait where the selected clones had either high or low values,
822 aiming to represent as much as possible of the observed variation (Fig. S4B). ANOVAs and Tukey HSD tests
823 were performed for pairwise comparisons between the selected clones, to test for significant differences
824 among the analyzed traits (p<0.05). Berries used for total RNA extraction were sampled during *veraison*
825 stage (switch point from berry development to ripening, when 75% of berries were colored); a
826 transcriptionally active stage for genes involved in secondary metabolism pathways [96]. Sampling was
827 performed previous to 2019 harvest, within the same day (01.15.2019) and during morning time (between
828 8-10 a.m.); the collected berries were kept in liquid nitrogen until final storage in ultra-freezer (-80°C).
829 **Total RNA extraction, library preparation and sequencing:** Total RNA extractions and quality
830 checks were performed as previously described (annotation section). RNA samples were shipped to
831 Novogene (Beijing, China) for sequencing. At arrival, samples were tested before library preparation and
832 all passed the QC performed for quantitation (NanoDrop, ThermoFisher), degradation (agarose gel) and
833 integrity (Agilent 2100, Bioanalyzer systems). Library preparation was performed with a TruSeq Stranded
834 mRNA LT Sample Prep Kit (Illumina) and included rRNA depletion using the Ribo-Zero kit (Illumina). All
835 samples passed the library QC performed for concentration (Qubit 2.0, ThermoFisher), insert size (Agilent
836 2100) and library effective concentration (qPCR). An Illumina HiSeq 2500 instrument was used for
837 sequencing, to obtain 150 bp paired-end reads.
838 **Differentially expressed genes and functional enrichment analyses:** The base quality of the raw
839 reads was checked with FastQC [77]. Trimming of sequencing adaptors and low quality sequences was
840 performed with Trimmomatic [97]. rRNA and tRNA sequences were removed from the trimmed Illumina
841 reads using Bowtie2 [98] aligner, by mapping the reads to a specific grapevine database [99]. All forthcoming
842 bioinformatic procedures were performed following a haplotype-aware scheme, using separately and in
843 parallel both Malbec haplophases as reference genomes. Alignment of the trimmed reads was conducted
844 with STAR [100], the assembly of the aligned reads with StringTie [79], and GffCompare [89] was used for
845 extraction and annotation of transcripts. The read counting of the aligned *bam* files was performed with
846 FeatureCount [101], according to Chialva *et al*. [102].
847 Heatmaps and PCAs displaying the global gene expression were built in R, based on the regularized
848 log transformation of the normalized counts (rld), obtained with DESeq2 [103]. To detect differentially
849 expressed genes (DEG) with DESeq2 [103], pairwise comparisons among the four clones were performed.
850 The threshold to consider a gene to be differentially expressed (DEG) was set to a p-value<0.01 (FDR-

adjusted) and an absolute value of log_2 fold-change |LFC|≥1. Functional enrichment analyses were conducted using the list of DEGs with g:Profiler2 [93], to find statistically significant gene ontology (GO) terms related to Biological Processes (BP) and Molecular Functions (MF). For all GO enrichment analysis the g:GOSt function of g:Profiler was implemented using the PN40024.v4 functional annotation as reference. Aiming to compare DEGs and BPs identified with each of Malbec haplophases, the orthologue genes obtained with Orthofinder were employed. For all performed GO terms enrichment analyses, only manually curated annotations of experimental and computational studies were considered, and electronically inferred annotations were dismissed to avoid sporous results. This was performed based on a *p*-value <0.05 (g:SCS-corrected for multiple comparisons [93]). The aggregate scores and z-scores where obtained and graphically represented with the R package GeneTonic [104]. Z-scores summarize the direction of change based on the LFC values of the gene set involved in the enriched process. Positive z-scores (color red) indicate that a BP GO term is upregulated, whereas negative z-scores (color blue) indicate downregulation of a BP [104]. GeneTonic function *gs_overlap* = 0.9 was employed to summarize redundant GO terms and *gs_summarize_overview_pair* was employed to make dumbbell plots, to visualize simultaneously the enriched GO terms detected with both haplophases. The web tools UpSetR (https://gehlenborglab.shinyapps.io/upsetr/) and jVenn [105] were employed to perform upset plots and Venn diagrams for DEGs pairwise comparisons.

6. Acknowledgements

This work was supported by Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT): PICT2018-02381, MINCyT/ANPCyT (FONTAR)-CDTI IBEROGEN; CONICET (Bilateral PCB-II, CONICET-CSIC; MINECO BIO2017-86375-R, and the project PID2020-120183RB-I00 funded by MCIN/AEI/10.13039/501100011033), and the Max Planck Society. L.C. was supported by a CONICET travel grant for a short-term staying at the Max Planck (Tübingen). We are grateful to Ilja Bezrukov for answering bioinformatics question, and to Silvina van Houten for collaboration with field work. We thank Olivier Yobregat (Institut Français de la Vigne et du Vin, IFV) for providing plant material from Malbec parental cultivars. This project received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 797460. This study benefited from the networking activities and resources produced within the COST action Integrape (CA17111) and COST innovators grant Grapedia (IG17111).

7. Author contributions

L.C. coordinated the project and wrote the article. Produced and analyzed phenotypic, transcriptomic and genomic data. P.C.B. laboratory work for genomic sequencing and designed the bioinformatic pipelines for genome assembly and annotation, analyzed genomic data, and global data

884 interpretation. C.M. phenotypic and transcriptomic diversity analysis. L.B. field and laboratory work for
885 phenotypic diversity analysis. D.B. and C.S. project design, provided insight and access to the analyzed
886 clonal accessions. W.T. bioinformatic analysis, set-up and maintenance of genome browse at IBAM-
887 CONICET web-page. C.L. laboratory work for genomic data obtention. S.G.T., J.I. and J.M.M.Z. project
888 design for genomic and transcriptomic experiments. C.R., J.I. and J.M.M.Z. obtained parental cultivars
889 plant material and provided resources for whole DNA extractions. D.W. provided resources, facilities and
890 advice for genomic laboratory work and bioinformatic analyses. D.L. designed and coordinated the entire
891 project, performed transcriptomic analysis. All authors read and improved this manuscript

892 **8. Data availability statement**

893 Malbec-Mag and Malbec-Pru assemblies are available at the NCBI public repository under
894 BioProjects: PRJNA1036636 and PRJNA1036637, respectively. All raw data is available at NCBI under
895 BioProject: PRJNA1037531, including genomic Illumina short-reads obtained for cultivars Malbec,
896 Prunelard and Magdeleine, PacBio long reads for Malbec and a batch of 22 RNA-Seq from eight Malbec
897 clones and replicates. Malbec annotated assemblies are also available for visualization and gene browse
898 at two public repositories: IBAM-CONICET (http://ibam.mendoza-conicet.gob.ar/resources/genome-
899 visualizer/) and Gramene (https://www.gramene.org/). A gene correspondence between Malbec-Pru and
900 Malbec-Mag annotated genes, alongside with PN40024.v4 and VCost.v3 annotations is provided,
901 therefore gene queries can be performed using any of the mentioned formats.

902 **9. Conflicts of interests**

903 The authors declare no conflicts of interest.

904 **10. Supplementary data**

905 Supplementary data is available at Horticulture Research online.

906 **11. References**

907 1. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in
908 major angiosperm phyla. *nature* **449**, 463 (2007).
909 2. Velt, A. *et al.* An improved reference of the grapevine genome reasserts the origin of the
910 PN40024 highly homozygous genotype. *G3 Genes|Genomes|Genetics* jkad067 (2023)
911 doi:10.1093/g3journal/jkad067.
912 3. Canaguier, A. *et al.* A new version of the grapevine reference genome assembly (12X.v2)
913 and of its annotation (VCost.v3). *Genomics Data* **14**, 56–62 (2017).
914 4. Shi, X. *et al.* The complete reference genome for grapevine (Vitis vinifera L.) genetics and
915 breeding. *Horticulture Research* uhad061 (2023) doi:10.1093/hr/uhad061.
916 5. Da Silva, C. *et al.* The High Polyphenol Content of Grapevine Cultivar Tannat Berries Is
917 Conferred Primarily by Genes That Are Not Shared with the Reference Genome. *Plant Cell* **25**,
918 4777–4788 (2013).
919 6. Zhou, Y. *et al.* The population genetics of structural variants in grapevine domestication.
920 *Nat. Plants* **5**, 965–979 (2019).
921 7. Velasco, R. *et al.* A High Quality Draft Consensus Sequence of the Genome of a

Heterozygous Grapevine Variety. *PLoS ONE* **2**, e1326 (2007).
8. Lijavetzky, D., Cabezas, J., Ibáñez, A., Rodríguez, V. & Martínez-Zapater, J. M. High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* **8**, 424 (2007).
9. Maestri, S. *et al.* 'Nebbiolo' genome assembly allows surveying the occurrence and functional implications of genomic structural variations in grapevines (Vitis vinifera L.). *BMC Genomics* **23**, 159 (2022).
10. Carbonell-Bejerano, P. *et al.* Catastrophic Unbalanced Genome Rearrangements Cause Somatic Loss of Berry Color in Grapevine. *Plant Physiol.* **175**, 786–801 (2017).
11. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**, 1050–1054 (2016).
12. Roach, M. J. *et al.* Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLoS Genet* **14**, e1007807 (2018).
13. Vondras, A. M. *et al.* The genomic diversification of grapevine clones. *BMC Genomics* **20**, 972 (2019).
14. Massonnet, M. *et al.* The genetic basis of sex determination in grapes. *Nat Commun* **11**, 2902 (2020).
15. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* **36**, 1174–1182 (2018).
16. Minio, A., Cochetel, N., Vondras, A. M., Massonnet, M. & Cantu, D. Assembly of complete diploid-phased chromosomes from draft genome sequences. *G3 (Bethesda)* **12**, jkac143 (2022).
17. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
18. Onetto, C. A., Ward, C. M. & Borneman, A. R. The phased diploid genome assembly of Vitis vinifera cv. Shiraz. 2022.12.12.520163 Preprint at https://doi.org/10.1101/2022.12.12.520163 (2022).
19. Sichel, V. *et al.* Chimeras in Merlot grapevine revealed by phased assembly. (2022).
20. This, P., Lacombe, T. & Thomas, M. Historical origins and genetic diversity of wine grapes. *Trends in Genetics* **22**, 511–519 (2006).
21. Freitas, S. *et al.* Pervasive hybridization with local wild relatives in Western European grapevine varieties. *Science Advances* **7**, eabi8584 (2021).
22. Xiao, H. *et al.* Adaptive and maladaptive introgression in grapevine domestication. *Proceedings of the National Academy of Sciences* **120**, e2222041120 (2023).
23. Carbonell-Bejerano, P., Royo, C., Mauri, N., Ibáñez, J. & Miguel Martínez Zapater, J. Somatic Variation and Cultivar Innovation in Grapevine. in *Advances in Grape and Wine Biotechnology* (eds. Morata, A. & Loira, I.) (IntechOpen, 2019). doi:10.5772/intechopen.86443.
24. Arrizabalaga, M. *et al.* Tempranillo clones differ in the response of berry sugar and anthocyanin accumulation to elevated temperature. *Plant Science* **267**, 74–83 (2018).
25. Grimplet, J., Ibáñez, S., Baroja, E., Tello, J. & Ibáñez, J. Phenotypic, Hormonal, and Genomic Variation Among Vitis vinifera Clones With Different Cluster Compactness and Reproductive Performance. *Front. Plant Sci.* **9**, 1917 (2019).
26. Wolkovich, E. M., García de Cortázar-Atauri, I., Morales-Castilla, I., Nicholas, K. A. & Lacombe, T. From Pinot to Xinomavro in the world's future wine-growing regions. *Nature Clim Change* **8**, 29–37 (2018).
27. Boursiquot, J.-M. *et al.* Parentage of Merlot and related winegrape cultivars of southwestern France: discovery of the missing link. *Australian Journal of Grape and Wine Research* **15**, 144–155 (2009).
28. Lacoste, P. El Malbec de Francia: la Denominación de Origen Controlada 'Cahors'. Historia y perspectivas. *Idesia* **33**, 113–124 (2015).

971 29. OIV. Distribution of the world's grapevine varieties. (2017).
972 30. Instituto Nacional del Vino (Argentina). Informe variedad Malbec. (2019).
973 31. Urvieta, R., Jones, G., Buscema, F., Bottini, R. & Fontana, A. Terroir and vintage
974 discrimination of Malbec wines based on phenolic composition across multiple sites in Mendoza,
975 Argentina. *Scientific Reports* **11**, 2863 (2021).
976 32. Muñoz, C. *et al.* Relationships among Gene Expression and Anthocyanin Composition of
977 Malbec Grapevine Clones. *J. Agric. Food Chem.* **62**, 6716–6725 (2014).
978 33. van Houten, S. *et al.* Natural Genetic Variation for Grapevine Phenology as a Tool for
979 Climate Change Adaptation. *Applied Sciences* **10**, 5573 (2020).
980 34. Varela, A. *et al.* Vineyard environments influence Malbec grapevine phenotypic traits and
981 DNA methylation patterns in a clone-dependent way. *Plant Cell Rep* **40**, 111–125 (2021).
982 35. Calderón, L. *et al.* Whole genome resequencing and custom genotyping unveil clonal
983 lineages in 'Malbec' grapevines (Vitis vinifera L.). *Scientific Reports* **11**, 7775 (2021).
984 36. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and
985 local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277 (2019).
986 37. Dong, Y. *et al.* Dual domestications and origin of traits in grapevine evolution. *Science* **379**,
987 892–901 (2023).
988 38. Waters, E. R. & Vierling, E. Plant small heat shock proteins – evolutionary and functional
989 diversity. *New Phytologist* **227**, 24–37 (2020).
990 39. Shu, K., Zhou, W., Chen, F., Luo, X. & Yang, W. Abscisic Acid and Gibberellins
991 Antagonistically Mediate Plant Development and Abiotic Stress Responses. *Frontiers in Plant*
992 *Science* **9**, (2018).
993 40. Kobayashi, S. Retrotransposon-Induced Mutations in Grape Skin Color. *Science* **304**, 982–
994 982 (2004).
995 41. Wong, D. C. J., Lopez Gutierrez, R., Gambetta, G. A. & Castellarin, S. D. Genome-wide
996 analysis of cis-regulatory element structure and discovery of motif-driven gene co-expression
997 networks in grapevine. *DNA Res* **24**, 311–326 (2017).
998 42. Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A. & Cantu, D. Diploid Genome
999 Assembly of the Wine Grape Carménère. *G3* **9**, 1331–1337 (2019).
1000 43. Raimondi, S. *et al.* DNA-based genealogy reconstruction of Nebbiolo, Barbera and other
1001 ancient grapevine cultivars from northwestern Italy. *Sci Rep* **10**, 15782 (2020).
1002 44. Minio, A. *et al.* Iso-Seq Allows Genome-Independent Transcriptome Profiling of Grape
1003 Berry Development. *G3 Genes|Genomes|Genetics* **9**, 755–767 (2019).
1004 45. Chialva, C. *et al.* Differential Expression Patterns Within the Grapevine Stilbene Synthase
1005 Gene Family Revealed Through Their Regulatory Regions. *Plant Mol Biol Rep* **36**, 225–238 (2018).
1006 46. Pilati, S. *et al.* Abscisic Acid Is a Major Regulator of Grape Berry Ripening Onset: New
1007 Insights into ABA Signaling Network. *Front Plant Sci* **8**, 1093 (2017).
1008 47. Tischer, S. V. *et al.* Combinatorial interaction network of abscisic acid receptors and
1009 coreceptors from Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* **114**,
1010 10280–10285 (2017).
1011 48. Valdés, A. E., Övernäs, E., Johansson, H., Rada-Iglesias, A. & Engström, P. The
1012 homeodomain-leucine zipper (HD-Zip) class I transcription factors ATHB7 and ATHB12 modulate
1013 abscisic acid signalling by regulating protein phosphatase 2C and abscisic acid receptor gene
1014 activities. *Plant Mol Biol* **80**, 405–418 (2012).
1015 49. Koyama, K., Sadamatsu, K. & Goto-Yamamoto, N. Abscisic acid stimulated ripening and
1016 gene expression in berry skins of the Cabernet Sauvignon grape. *Funct Integr Genomics* **10**, 367–
1017 381 (2010).
1018 50. Walker, A. R. *et al.* White grapes arose through the mutation of two similar and adjacent
1019 regulatory genes: White grape genes. *The Plant Journal* **49**, 772–785 (2007).

1020 51. Gao, Z. *et al.* Characterization of the ABA Receptor VlPYL1 That Regulates Anthocyanin
1021 Accumulation in Grape Berry Skin. *Front Plant Sci* **9**, 592 (2018).
1022 52. Falginella, L. *et al.* Differential regulation of triterpene biosynthesis induced by an early
1023 failure in cuticle formation in apple. *Hortic Res* **8**, 1–15 (2021).
1024 53. Wang, N. *et al.* Phased genomics reveals hidden somatic mutations and provides insight
1025 into fruit development in sweet orange. *Horticulture Research* **11**, uhad268 (2024).
1026 54. Zou, C. *et al.* A multitiered haplotype strategy to enhance phased assembly and fine
1027 mapping of a disease resistance locus. *Plant Physiology* **193**, 2321–2336 (2023).
1028 55. Workman, R. *et al.* High molecular weight DNA extraction from recalcitrant plant species
1029 for third generation sequencing. (2018).
1030 56. Rabanal, F. A. *et al.* Pushing the limits of HiFi assemblies reveals centromere diversity
1031 between two Arabidopsis thaliana genomes. *Nucleic Acids Research* **50**, 12309–12327 (2022).
1032 57. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer
1033 weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
1034 58. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
1035 *EMBnet.journal* **17**, 10–12 (2011).
1036 59. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. Preprint at
1037 https://doi.org/10.12688/f1000research.12232.1 (2017).
1038 60. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
1039 using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
1040 61. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT
1041 sequencing data. *Nat Methods* **10**, 563–569 (2013).
1042 62. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality,
1043 completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).
1044 63. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
1045 3100 (2018).
1046 64. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
1047 2079 (2009).
1048 65. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome
1049 assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
1050 66. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes.
1051 *Genome Biol* **20**, 224 (2019).
1052 67. Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection
1053 and Genome Assembly Improvement. *PLOS ONE* **9**, e112963 (2014).
1054 68. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update:
1055 Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for
1056 Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647–
1057 4654 (2021).
1058 69. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato
1059 system for high-throughput genome editing. *Genome Biology* **23**, 258 (2022).
1060 70. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q
1061 File Manipulation. *PLoS ONE* **11**, e0163962 (2016).
1062 71. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS*
1063 *Computational Biology* **14**, e1005944 (2018).
1064 72. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
1065 73. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a
1066 streamlined, comprehensive pipeline. *Genome Biology* **20**, 275 (2019).
1067 74. Campbell, M. S. *et al.* MAKER-P: A Tool Kit for the Rapid Creation, Management, and
1068 Quality Control of Plant Genome Annotations. *Plant Physiology* **164**, 513–524 (2014).

1069 75. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler
1070 and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
1071 76. Reid, K. E., Olsson, N., Schlosser, J., Peng, F. & Lund, S. T. An optimized grapevine RNA
1072 isolation procedure and statistical determination of reference genes for real-time RT-PCR during
1073 berry development. *BMC Plant Biol* **6**, 27 (2006).
1074 77. Andrews, S. FastQC a quality-control tool for high-throughput sequence data http://www.
1075 *Bioinformaticsbabraham. ac. uk/projects/fastqc* (2014).
1076 78. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment
1077 and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
1078 79. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-
1079 seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
1080 80. Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and
1081 genomic alignments for improved gene prediction in the human genome. *Genome Biology* 8
1082 (2006).
1083 81. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic
1084 eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein
1085 database. *NAR Genomics and Bioinformatics* **3**, lqaa108 (2021).
1086 82. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
1087 features. *Bioinformatics* **26**, 841–842 (2010).
1088 83. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics*
1089 **37**, 1639–1643 (2021).
1090 84. Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A. & Wortman, J. R. Approaches to Fungal
1091 Genome Annotation. *Mycology* **2**, 118–141 (2011).
1092 85. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res* **12**, 656–664 (2002).
1093 86. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA
1094 and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
1095 87. Cochetel, N. *et al.* Diploid chromosome-scale assembly of the Muscadinia rotundifolia
1096 genome supports chromosome fusion and disease resistance gene expansion during Vitis and
1097 Muscadinia divergence. *G3 Genes|Genomes|Genetics* **11**, jkab033 (2021).
1098 88. Dainat, J., Hereñú, D., LucileSol & pascal-git. NBISweden/AGAT: AGAT-v0.8.1. Zenodo
1099 https://doi.org/10.5281/zenodo.5834795 (2022).
1100 89. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**, ISCB Comm J-
1101 304 (2020).
1102 90. Törönen, P. & Holm, L. PANNZER—A practical tool for protein function prediction. *Protein*
1103 *Science* **31**, 118–128 (2022).
1104 91. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
1105 genomics. *Genome Biol* **20**, 238 (2019).
1106 92. Lovell, J. T. *et al.* GENESPACE: syntenic pan-genome annotations for eukaryotes.
1107 2022.03.09.483468 Preprint at https://doi.org/10.1101/2022.03.09.483468 (2022).
1108 93. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and
1109 conversions of gene lists (2019 update). *Nucleic Acids Research* **47**, W191–W198 (2019).
1110 94. Di Gaspero, G. & Foria, S. 2 - Molecular grapevine breeding techniques. in *Grapevine*
1111 *Breeding Programs for the Wine Industry* (ed. Reynolds, A.) 23–37 (Woodhead Publishing, Oxford,
1112 2015). doi:10.1016/B978-1-78242-075-0.00002-8.
1113 95. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
1114 (2009).
1115 96. Massonnet, M. *et al.* Ripening Transcriptomic Program in Red and White Grapevine
1116 Varieties Correlates with Berry Skin Anthocyanin Accumulation. *Plant Physiol.* **174**, 2376–2396
1117 (2017).

1118 97. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
1119 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
1120 98. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**,
1121 357–359 (2012).
1122 99. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic
1123 sequence. *Nucleic Acids Res* **37**, D93–D97 (2009).
1124 100. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
1125 101. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for
1126 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
1127 102. Chialva, C., Blein, T., Crespi, M. & Lijavetzky, D. Insights into long non-coding RNA
1128 regulation of anthocyanin carrot root pigmentation. *Scientific Reports* **11**, 4093 (2021).
1129 103. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
1130 RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
1131 104. Marini, F., Ludt, A., Linke, J. & Strauch, K. GeneTonic: an R/Bioconductor package for
1132 streamlining the interpretation of RNA-seq data. *BMC Bioinformatics* **22**, 610 (2021).
1133 105. Bardou, P., Mariette, J., Escudié, F., Djemiel, C. & Klopp, C. jvenn: an interactive Venn
1134 diagram viewer. *BMC Bioinformatics* **15**, 293 (2014).
1135