

Integrative genomics reveals the polygenic basis of seedlessness in grapevine

Xu Wang^{1, 2, †}, Zhongjie Liu^{1, †}, Fan Zhang¹, Hua Xiao¹, Shuo Cao¹, Hui Xue¹, Wenwen Liu¹, Ying Su¹, Zhenya Liu¹, Haixia Zhong³, Fuchun Zhang³, Bilal Ahmad¹, Qiming Long¹, Yingchun Zhang¹, Yuting Liu¹, Yu Gan¹, Ting Hou¹, Zhongxin Jin¹, Xinyu Wu³, Yiwen Wang¹, Yanling Peng¹, and Yongfeng Zhou^{1, 4, *}

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

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

3 The State Key Laboratory of Genetic Improvement and Germplasm Innovation of Crop Resistance in Arid Desert Regions (Preparation), Key Laboratory of Genome Research and Genetic Improvement of Xinjiang Characteristic Fruits and Vegetables, Institute of Horticultural Crops, Xinjiang Academy of Agricultural Sciences, Urumqi, China

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

† These authors contributed equally to this work.

26 * Corresponding authors: zhouyongfeng@caas.cn

27 Abstract

28 Seedlessness is a crucial quality trait in table grape (*Vitis vinifera* L.) breeding.
 29 However, the development of seeds involved intricate regulations, while the
 30 polygenic basis of seed abortion remains unclear. Here, we combine comparative
 31 genomics, population genetics, quantitative genetics, and integrative genomics to
 32 unravel the evolution and polygenic basis of seedlessness in grapes. We generated
 33 four haplotype-resolved telomere-to-telomere (T2T) genomes for two seedless grape
 34 cultivars, ‘Thompson Seedless’ (TS, syn. ‘Sultania’) and ‘Black Monukka’ (BM).
 35 Comparative genomics identified a ~4.25 Mb hemizygous inversion on Chr10
 36 specific in seedless cultivars, with seedless-associated genes *VvTT16* and *VvSUS2*
 37 located at breakpoints. Population genomic analyses of 548 grapevine accessions
 38 revealed two distinct clusters of seedless cultivars, tracing the origin of the
 39 seedlessness trait back to ‘Sultania’. Introgression, rather than convergent selection,
 40 shaped the evolutionary history of seedlessness in grape improvement. Genome-wide
 41 association study (GWAS) analysis identified 110 quantitative trait loci (QTLs)
 42 associated with 634 candidate genes, including novel candidate genes, such as three
 43 *11S GLOBULIN SEED STORAGE PROTEIN* and two *CYTOCHROME P450* genes,
 44 and well-known genes like *VviAGL11*. Integrative genomic analyses resulted in 339
 45 core candidate genes categorized into 13 groups related to seed development.
 46 Machine learning based genomic selection achieved a remarkable 99% precision in
 47 predicting grapevine seedlessness. Our findings highlight the polygenic nature of
 48 seedless and provide novel candidate genes for molecular genetics and an effective
 49 prediction for seedlessness in grape genomic breeding.

50

51 **Keywords:** Viticulture, Grape breeding, Seed abortion, Structural variations,
 52 Genomic breeding, Genomic selection

54 Introduction

55 The production of seedless fruits leads to tremendous success in the global fruit
 56 market¹, such as bananas^{2,3}, citrus^{4,5}, watermelons^{6,7}, and table grapes⁸. Seed abortion
 57 in table grape has been a major focus of breeding efforts for decades, as seedless
 58 grapes are highly preferred by consumers owing to improved tastes and convenience.
 59 There are two primary methods employed to obtain seedless grapes. One involving
 60 the application of phytohormones, applying gibberellin acid (GA) and cytokinin
 61 analogs before the full bloom stage can effectively induce seed abortion in seeded
 62 grapes⁹⁻¹¹. Although this process has already become a common practice for
 63 producing seedless grapes, it raises concerns about food safety and labor costs¹².
 64 Another one is based on genetic breeding of seedless grape cultivars. Breeders have
 65 explored diploid and triploid breeding approaches and the embryo rescue strategy to
 66 obtain new seedless varieties in recent decades¹³⁻¹⁶.

67 The development of various seed tissues in grapes involves intricate genomic
 68 regulations. Previous studies have identified specific genes associated with different
 69 tissues of seed development¹⁷. For instance, the formation of the seed coat
 70 (integument) has been affected by genes like *VviAGL11*¹⁸⁻²¹, *VvMADS28*²²,
 71 *VviINO*^{23,24}, and *VvHB63*²⁵⁻²⁷. Nutrient storage in the endosperm is controlled by
 72 genes such as 7S and 11S globulin-like seed storage proteins²⁸⁻³⁰, while normal
 73 embryo growth relies on several gibberellin (GA) genes^{31,32}. Moreover, the growth of
 74 ovules (young seeds) is influenced by genes such as *VvMADS39*³³, *VvMADS45*³⁴,
 75 *VviABCG20*³⁵⁻³⁷, *VvFUS3*³⁸, *VvNAC26*³⁹, *VvβVPE*^{40,41}, *VviASN1*⁴². In general,
 76 multiple genes involved in tissue development of grapevine seeds. Seed abortion in
 77 grapes can occur when any of the seed tissues fail to develop properly. However, the
 78 polygenic basis of seedlessness in grapes remains unclear.

79 Previous studies have mapped multiple QTLs in different progenies in grapevine. In
 80 the linkage map of ‘Dominga’ and ‘Autumn Seedless’, three QTLs for seed number

(SN) and six QTLs for seed fresh weight (SFW) were detected⁴³. Similarly, in the ‘Muscat of Alexandria’ and ‘Crimson Seedless’ progeny, six QTLs for SN and ten QTLs for SFW were detected⁴⁴. Recent comparative analyses, encompassing 28 different grape varieties (13 seeded and 15 seedless), have detected 34 candidate genes associated with the divergence between seeded and seedless lineages⁴⁵. However, the restricted genetic background in progenies and limited population samples hinders the investigation of the polygenic basis of seedlessness in grapes.

In this study, we first generated haplotype-resolved T2T genomes for two seedless grapes, ‘Thompson Seedless’ (TS) and ‘Black Monukka’ (BM), and we compared these haplotype genomes with 11 other grape genomes to detect structural variations (SVs) between seeded and seedless genomes. Population genetic analysis was conducted on whole-genome sequencing (WGS) of 548 accessions to investigate the evolutionary history of seedlessness, while quantitative genetic analysis involved 444 accessions to map QTLs and key genes associated with seedlessness. Integrative genomic analysis incorporated three transcriptomes with 14 development stages, homologous genes related to 34 Gene Ontology (GO) terms, and 451 family genes and seven molecular markers previously reported with significant effects on seed development processes. Finally, genomic selection, based on polygenic model and machine learning algorithms, were applied in predicting the seedlessness trait in grapes. Collectively, we aimed to address five sets of questions. First, at genome level, how do the seedless cultivars compare with cultivars with seeds? Can we detect big SVs related to seedless/seeded cultivars? Second, what evolutionary factors has driven the origin of seedlessness during grape improvement? i.e., introgression or convergent selection? Third, based on large natural populations, can we map genetic loci and candidate genes involved in seed abortion in table grapes? Fourth, can we integrate genomic analyses to identify core candidate genes underlying seed abortion? Finally, can we employ machine learning based genome selection to enhance prediction precision in grape breeding? Overall, our work contributes to improving

109 the understanding of the polygenic basis of seedlessness and facilitate genomic
110 breeding of grapes.

Results

Comparative genomics between seeded and seedless cultivars

To study the genetic basis of seedlessness, we generated haplotype-resolved T2T assemblies for two seedless cultivars: TS and BM (**Fig. 1b, c**), utilizing high-depth PacBio HiFi sequencing (~120× coverage) and Hi-C sequencing (~116× coverage; **Extended Data Fig. 2c**). The evaluation of K-mer heterozygosity in the TS and BM genomes, based on HiFi data, measured 1.51% and 1.41%, respectively. The quality of these genomes meets the assessment standards of the T2T level⁴⁶, with all centromeres regions and mostly telomeres regions marked (**Fig. 1a and Supplementary Table 1, 3**). Statistical analysis of variants revealed that TS and BM, between their two haplotype genomes, harbor 5.35 Mb and 5.04 Mb of SNPs, 5.01 Mb and 4.54 Mb of insertions and deletions (InDels, < 50 bp), and 33.42 Mb and 31.87 Mb of SVs (≥ 50 bp), respectively (**Supplementary Table 4**). A unique heterozygous inversion region (PN_T2T, Chr15: 10.7-12.0 Mb) specific to the TS hap2 genome was detected when aligning the four haplotypic genomes to PN_T2T, and several genes were found near the inversion breakpoints, such as *AGAMOUS* (*VvAG2*), *AGAMOUS-LIKE 62* (*VvAGL62*), *OIL BODY-ASSOCIATED PROTEIN 2B* (*VvOBAP2B*), and *GDSL esterase/lipase At1g29670*, which are involved in stamen and carpel determining, early endosperm development, and oil body synthesis (**Extended Data Fig. 3b and Supplementary Table 5**). The differential chromosomes between the two haplotype genomes explains the polymorphism of alleles and the variation in the number of genes (**Supplementary Table 1**).

To further investigate the SVs associated with seedless and seeded grapes, we aligned a total of 15 genomes, including the five seedless genomes and ten seeded genomes, to the PN_T2T (**Extended Data Fig. 3a**). We detected a heterozygous inversion (PN_T2T, Chr10: 23.8-25.4 Mb) in seedless grape varieties (**Fig. 1d**). The authenticity of these inversions was confirmed by Hi-C heatmaps and IGV⁴⁷ (**Fig. 1e**

138 **and Extended Data Fig. 4).** A total of 210 genes (Chr10: 21.75-26.00 Mb) and 237
139 genes (Chr10:23.00-27.50 Mb) were identified in the inversion regions of TS hap1
140 and BM hap1, respectively (**Fig. 1f and Supplementary Table 7-8**). Three seed
141 development-related genes, *TRANSPARENT TESTA 16/ ARABIDOPSIS BSISTER*
142 (*TT16/ABS*) and two *SUCROSE SYNTHASE 2 (SUS2)* genes, were discovered near
143 the breakpoints of inversion region of the haplotype genomes. *TT16/ABS* controls the
144 formation of the maternal-derived endothelial cells by interacting with *AGL11/*
145 *SEEDSTICK (STK)* in the previous studies⁴⁸⁻⁵¹. *VvTT16* was found to be present in
146 both TS and BM haplotype genomes, while the two *VvSUS2* tandem duplication
147 genes were hemizygous, present only in the hap1 genome of TS and BM
148 (**Supplementary Table 6-7**). These findings suggest that the power of comparative
149 genomics of haplotype-resolved T2T genomes in uncovering overlooked new
150 candidate genes underlying crucial agronomic traits.

151 **Introgression rather than convergent evolution underlying the** 152 **evolvment of seedlessness in grapevine**

153 To explore the evolutionary history of seedlessness in grape improvement, we used
154 WGS data from 548 grapevine accessions, including 46 seedless grapes, for
155 population genetic analysis (**Supplementary Table 8**). A total of 4,462,797 SNPs,
156 443,812 InDels, 487,204 SVs were identified by aligning WGS data to ‘Cabernet
157 Sauvignon’ (CS) genome⁵². The phylogenetic tree split into six primary population
158 branches: European wild grapes (*V. vinifera* ssp. *sylvestris* EU population, EU, n = 69),
159 Middle East and Caucasus region wild grapes (*V. vinifera* ssp. *sylvestris* ME
160 population, ME, n = 23), domesticated grapes (*V. vinifera* ssp. *vinifera*, VV, n = 352),
161 American fox grapes (*V. labrusca*, VL, n = 5), hybrid of VV and VL grapes (VV×VL,
162 *V. vinifera* × *Vitis labrusca*, n = 92), and outgroup grapes (OG, n = 7; **Fig. 2c and**
163 **Extended Data Fig. 5**). The results showed two independent lineages of seedless
164 grapes nested in the VV×VL and VV branches, respectively, which is also supported

165 by PCA (**Fig. 2b, c**). The seedless grapes nested in two branches, which could be
166 driven by convergent artificial selection or introgression during grape improvement.

167 To distinguish convergent selection and introgression in generating seedless traits in
168 different grapevine lineages, we performed the introgression analyses, throughout the
169 whole genome using f_d statistics⁵³, following previous studies^{54,55}. Interestingly, we
170 detected significant genomic signals of introgression at seedless associated locus (see
171 the GWAS section) between VV and VV×VL seedless grapes (the upper 5th
172 percentile, $f_d = 0.266$, $P = 1.28e-39$), including the redefined *SEED DEVELOPMENT*
173 *INHIBITOR* (SDInew, 30.36-31.86 Mb) locus²⁰ and the newly detected QTL on Chr07
174 (SDI2, 8.85-8.86 Mb; **Fig. 3c, d**). These results suggested that introgression rather
175 than convergent artificial selection has driven the evolutionary history of seedlessness
176 in grapes.

177 To validate the genetic relationship among seedless varieties, a network was
178 constructed comprising 46 seedless grapes (35 VV and 11 VV×VL) based on the
179 results of Identity-by-Descent (IBD) analysis (**Supplementary Table 8**). The result
180 revealed that TS group and BM group serve as bridges for gene flow between VV and
181 VV×VL clusters (**Fig. 2d, e**). In fact, BM grape traces its ancestry back to ‘Sultania’
182 (TS) and ‘Ichkimar’⁵⁶, with an IBD score of 0.50 supporting this observation
183 (**Supplementary Table 9**). Additionally, we identified three grapes belonging to the
184 ‘Sultania’ somatic variants or synonym group (‘Jingfeng seedless’, TS1, and TS2,
185 $IBD > 0.95$), seven grapes classified as parent-offspring relationship ($IBD > 0.50$),
186 and 32 grapes ($0.50 > IBD > 0.09$) that were closely connected to ‘Sultania’ variety
187 (**Supplementary Table 9**). These findings provide evidence that ‘Sultania’ had been
188 extensively employed in crossbreeding with local grape varieties to enhance quality
189 and develop new seedless grape cultivars^{8,57}. Furthermore, the seed abortion could
190 also be caused by cytoplasmic male sterility (CMS)⁵⁸. We analyzed the chloroplast
191 and mitochondrial genomic variation and found nuclear inheritance, rather than CMS
192 inheritance, played a crucial role in controlling seed abortion (**Extended Data Fig. 6**).

193 These findings corroborated IBD results that frequent introgression has facilitated the
194 formation of seedlessness the VV and VV×VL populations (**Fig. 3e**). As a result, the
195 origin of the seedlessness trait could be traced back to ‘Sultania’, and continuous
196 introgression, rather than convergent evolution, led to seed abortion in seedless grape
197 varieties.

198 **Genome-wide Association Study for Seed Abortion Trait**

199 To detect the QTLs and genes associated with seed abortion, we used three population
200 for GWAS analysis, considering the effects of population structure in GWAS analyses:
201 VV (35 seedless and 317 seeded grapes), VV×VL (11 seedless and 81 seeded grapes),
202 and an admixed population (46 seedless and 398 seeded grapes; **Supplementary**
203 **Table 8**). We identified a total of 110 QTLs (634 genes), including 20 QTLs (126
204 genes) specific to the VV population, 18 QTLs (106 genes) specific to the VV×VL
205 population, and 72 consensus QTLs (402 genes) in admixed population, respectively
206 (**Extended Data Fig. 7 and Supplementary Table 10, 11**). GO analysis revealed that
207 genes specific to VV×VL were enriched in defense response and lignin catabolic
208 process ($P < 0.05$), while genes specific to VV population were enriched in embryo
209 development ending in seed dormancy and xylan metabolic process ($P < 0.05$;
210 **Extended Data Fig. 8 and Supplementary Table 12**).

211 Remarkably, two consensus regions exhibited high consistence in three populations:
212 Chr07: 8.85-8.86 Mb and Chr18: 29.40-35.54 Mb (**Fig. 3a, b**). In Chr07 locus (SDI2),
213 two genes, *REVERSE TRANSCRIPTASE ZINC-BINDING DOMAIN-CONTAINING*
214 *PROTEIN (LOC104880636, Vitvi011893)* and *STRUCTURAL MAINTENANCE OF*
215 *CHROMOSOMES PROTEIN (SMC1, Vitvi011891)*, were positioned within a tightly
216 linked region with high linkage disequilibrium (LD) values (**Fig. 3c**). The
217 *LOC104880636* gene is annotated as the regulation of seed growth on UniProt, and
218 the *smc1* mutants produced arrested early embryo development and blocked
219 cellularization of the endosperm⁵⁹ (**Supplementary Table 11**). Additionally, within

the 50 kb upstream region of SDI2, we identified a closely linked cluster of three tandem-duplicated genes, *11S GLOBULIN SEED STORAGE PROTEIN*^{28-30,60} (**Fig. 3c**), and designated them as *11S globulin G1*, *G2*, and *G3* based on their genomic positions. Notably, two nonsynonymous mutations and one deletion related to seedlessness were detected across 14 grape genomes (**Fig. 4a and Extended Data Fig. 9**). Among them, both the heterozygous Asp-to-Val and Leu-to-Val mutations were specific in seedless grapes, except for the somatic mutations of ‘Black Corinth’ (BC) seeded grapes. However, the heterozygous deletion of the 18 amino acids in *11S globulin G3* was only detected in seedless grapes (**Fig. 4a**). The relative expression values of these genes showed a significant correlation with seed phenotypes, especially from 40-50 days after flowering (DAF; **Fig. 4b**).

Another consensus region is located on the Chr18: 29.40-35.54 Mb, encompassing 17 QTLs and the reported SDI locus (Chr18: 29.83-31.34 Mb; **Fig. 3d**). Due to the narrow genetic background of the samples used in the previous study²⁰, we redefined the SDI locus (SDInew, Chr18: 30.36-31.86 Mb) through GWAS analysis, revealing a total of eight QTLs (**Fig. 3d and Extended Data Fig. 7**). In this region, the population differentiation was lower than the genomic background between VV (n = 35) and VV×VL (n = 11) seedless grapes, as showed by the fixation indices (F_{ST}) results, suggesting a relatively close genetic distance between the two populations (SDInew F_{ST} = 0.073 vs. genome-wide F_{ST} = 0.126; **Supplementary Table 13**). This finding was also supported by genetic diversity (π) statistics. The two seedless populations showed similar genetic diversity on the SDInew locus (**Fig. 3d**). The f_d statistics⁵³ revealed numerous sites showing evidence of introgressions (f_d = 0.266, P = 1.28e-39), including numerous genes related to seedlessness, such as *CYP716A94*, *CYP716A17*, *VviAGL11*, etc. (**Fig. 3d**). Notably, several SNPs and InDels were highly associated with these candidate genes, especially in the promoter and coding sequence region (**Fig. 4c, d and Extended Data Fig. 10**), while several published molecular markers for seedlessness prediction, including e7_VviAGL11⁴⁴, 5U_VviAGL11⁴⁴,

P3_VviAGL11¹⁸, and VMC7f2⁶¹, showed low predictive accuracy due to the absence of significant genotyping quality (**Fig. 4d and Supplementary Table 14**). We selected the top 10% of associated sites based on $-\log_{10}[P]$ values within the SDInew locus (**Fig. 4e**), and the most promising site for seedlessness prediction is Chr18_30874059, exhibiting a predicted accuracy of 97.8% for seedless grapes and 94.22% for seeded grapes in natural population.

Interestingly, we also detected two genomic regions for specific to each population (**Extended Data Fig. 7**). In the VV population, a specific region Chr01: 17.85-20.42 Mb harbored 13 QTLs and 67 genes, including seven primary genes involved in seed development, such as *NON-SPECIFIC LIPID TRANSFER PROTEIN GPI-ANCHORED 1 (LTGP1)*, *ARABIDOPSIS HISTIDINE KINASE 3 (AHK3)*, *B3 DOMAIN-CONTAINING TRANSCRIPTION FACTOR FUS3 (FUS3)*, *XYLOGLUCAN ENDOTRANSGLUCOSYLASE PROTEIN (XTH)*, *11-BETA-HYDROXYSTEROID DEHYDROGENASE B (SOP3)*, as well as previously reported VvMADS4⁶² and VvARF2-1⁶³ (**Fig. 3a**). In the VV×VL population, a specific region Chr18: 15.14-20.57 Mb harbored eight QTLs and 33 associated genes. Among these genes, *SERINE DECARBOXYLASE 1 (SDC1)*, *ABC TRANSPORTER G FAMILY MEMBER 22 ABCG22* or VvPNWBC22.2(TANG, 2018 #20), *SUS2*, *LACCASE-14 (LAC14)* and *TT10/LAC15* were highly associated with seed development (**Fig. 3a**). Our results suggest that seed abortion can be regulated by the collaborative effects of multiple genes, and the mapped candidate genes and variable sites hold valuable potential applications in seedless grapes breeding.

Integrative Genomic Analysis Identified 339 Seedless Candidate Genes

To elucidate the polygenic basis of seed abortion, we further utilized an integrative genomic analysis using transcriptomic analyses, seed development associated GO term genes, previously reported family genes and molecular markers, and GWAS

275 candidate genes, to identify the core candidate genes associated with seed abortion
 276 (**Supplementary Table 17**). Among these, three transcriptomic groups, including 76
 277 samples and 14 time points, were employed to detect differentially expressed genes
 278 (DEGs) between seeded and seedless grapes at each development stage
 279 (**Supplementary Table 15**). For the ‘Italia’ and ‘Hongju Seedless’ (HS) groups, we
 280 identified a total of 2,680 significantly upregulated genes and 1,835 significantly
 281 downregulated genes from the six time points (**Extended Data Fig. 13c**). Similarly, in
 282 the ‘Pinot Noir’ (PN) and TS groups, we identified a total of 3,969 significantly
 283 upregulated genes and 2,695 significantly downregulated genes (**Extended Data Fig.**
 284 **13c**). ‘Himrod Seedless’ (Himrod) and ‘Jinzao Wuhe’ (Jinzao) were used serve as a
 285 control for cross-validation during four fruit development stage. Interestingly, we
 286 found *VviAGL11* was exclusively in the downregulated DEGs in the PN and TS
 287 comparison, but not in the ‘Italia’ and HS comparison (**Extended Data Fig. 13a, b**).
 288 In addition, 1,301 core upregulated genes and 616 core downregulated genes were
 289 only identified in transcriptomic analyses using integrative genomic analysis, such as
 290 *DORMANCY-ASSOCIATED PROTEIN HOMOLOG 4* (*DRM1* homolog 4),
 291 *NON-SPECIFIC LIPID-TRANSFER PROTEIN 2* (*LTP2*), *VICILIN-LIKE SEED*
 292 *STORAGE PROTEIN*, *7S SEED STORAGE PROTEIN* (*7S GLOBULIN*), *TT10*,
 293 *LAC17*, etc. (**Fig. 5a and Supplementary Table 11**).

294 To include more important genes involved in seed development, we performed a
 295 protein sequence similarity alignment for all genes associated with 34 GO terms
 296 related to seed development against the PN_T2T reference genome (**Extended Data**
 297 **Fig. 12a, b**). This yielded 6,529 homologous genes (see Methods), with 5,061 genes
 298 only present in the GO pathway, including the *TT16/FBP24* gene identified in the
 299 comparative genomics (**Fig. 5a**). Notably, the intersection between the RNA-seq
 300 related genes and the GO homologous genes revealed 163 downregulated genes and
 301 294 upregulated genes (**Extended Data Fig. 13b**), such as *LTP*, *11S GLOBULIN*,
 302 *OLEOSIN*, *CELLULOSE SYNTHASE A CATALYTIC SUBUNIT* (*CESA4*, *CESA7* and

303 *CESA8*), etc. (**Supplementary Table 11**). The consensus GWAS genes and GO
 304 homologous genes exhibited nine candidate genes, such as previously mentioned
 305 *LOC104880636* (**Fig. 3c, 5a**). Furthermore, we integrated a total of 451 family genes
 306 and seven molecular markers related to seed abortion from previous studies
 307 (**Supplementary Table 16**). All these elements were aligned against the PN_T2T
 308 reference genome (e-value < 0.1) and exhibited overlap with candidate genes in other
 309 analyses, including previously reported genes such as *VvβVPE*⁴⁰, *HD-ZIP PROTEINS*
 310 *ATHB-1/HAT5*, *ATHB-12/VvHB56* and *ATHB40/VvHB18*²⁵, *VviASN1*⁴², *VvMJE1*⁶⁴,
 311 *VvLEC1*⁶⁵, *VvMADS2/VvSEP1*⁶⁶, etc. (**Extended Data Fig. 13d and Supplementary**
 312 **Table 11**).

313 Through integrative genomic analysis, we screened 339 core candidate genes,
 314 categorized into 13 groups, by condition-based filtering that exhibited significant
 315 differential expression between seedless and seeded grape cultivars (see Methods,
 316 **Supplementary Table 11**). Among them, 77 genes were directly associated with seed
 317 development-related GO homologous genes, and three groups deserve our attention:
 318 Firstly, the differential expression of candidate genes in the endosperm development
 319 impacts nutrient storage in seedless grapes (**Fig. 5b**). Nutrient deficiency could be
 320 primary factor leading to seed abortion in later embryo development; Secondly, genes
 321 involved in the regulation of lignin and cellulose synthesis/degradation in seed coat
 322 exhibit higher activity levels in seeded grapes (**Fig. 5c**); Thirdly, candidate genes
 323 manage the synthesis and transport of oil bodies ensuring efficient lipid accumulation
 324 and utilization during seed development (**Fig. 5d**). Overall, our findings indicate that
 325 multiple genes have an accumulative effect in the process of seed abortion, resulting
 326 varying degrees of seedlessness⁶⁷. This complexity highlights the challenge of
 327 accurately distinguish seed abortion using a single gene or variant site.

328 **Machine Learning based Genomic Selection for Seedless Grape** 329 **Breeding**

330 Given the polygenic nature of the seedlessness trait, genomic prediction could greatly
331 improve the speed and accuracy for seedless grape breeding. We extracted the
332 information of all 794 high-quality variant sites from GWAS analysis, including 77
333 InDels and 717 SNPs (**Supplementary Table 18**). Using these variations, an unrooted
334 phylogenetic tree was constructed based on admixed populations ($n = 444$), revealing
335 that the majority of seedless individuals clustered together (**Fig. 6a**). However, some
336 seeded samples, like seeded hybrid progeny²⁰ and Rizamat⁴⁴ were mixed in seedless
337 grapes, as well as seedless samples, such as ‘Dawn Seedless’, ‘Bronx Seedless’,
338 ‘Cheongsoo’, ‘Ruby Seedless’, and ‘Jingkejing’, were mixed in seeded grapes.
339 Interestingly, the mutations in the top two QTLs were heterozygous: Chr07: 8.85-8.86
340 Mb (SDI2 locus) and Chr18: 30.36-31.86 Mb (SDInew locus; **Fig. 6d**). These results
341 suggest the complexity of seedless in grapes.

342 Therefore, to address this problem, we employed genomic selection based on machine
343 learning to enhance predictive accuracy (**Extended Data Fig. 14**). We used 794
344 variant sites and phenotypic data from the admixed populations (444 samples) as the
345 training dataset, and evaluated the ability of nine different classical models to predict
346 the seedlessness trait based on 100 rounds of random cross-validations (**Fig. 6b**).
347 Among these models, machine learning based methods, including SVR-poly and
348 ElasticNetCV, demonstrated a strong performance in predicting the phenotype,
349 yielding prediction accuracies of 85.36% and 84.57%, respectively (**Fig. 6b**). As a
350 result, we applied SVR-poly and ElasticNetCV to genomic prediction on the testing
351 set data (39 samples not included in the model building) (**Supplementary Table 8**).
352 We observed that the SVR-poly model and the ElasticNetCV model yielded high
353 levels of accuracy, with correlation coefficient R values of 0.99 ($P < 2.2e-16$) and
354 0.96 ($P < 2.2e-16$), respectively (**Fig. 6c**), suggesting the efficiency of machine
355 learning based genomic selection of seedless in grapes.

356 Discussion

357 Seedless is an important quality trait in table grape breeding. Previous genetic
 358 investigated the functions of multiple genes, including *VvAGL11*, however,
 359 quantitative genetic analyses of natural population were not conducted in grapes. In
 360 this study, we conducted integrative genomic analyses to investigate the polygenic
 361 basis of seedlessness in grapes: (1) comparative analysis of 15 genomes allowed us to
 362 discover a heterozygous inversion in Chr10 associated with the seedless trait; (2) the
 363 evolutionary genomic analyses showed that seedless grapes were closely related with
 364 an origin from the well-known seedless grape ‘Sultania’ (TS). Introgression rather
 365 than convergent evolution was associated with the evolution of seedlessness in grapes;
 366 (3) a total of 110 QTLs associated with 634 candidate genes were identified through
 367 GWAS analysis within a large natural population, including four significant linkage
 368 regions such as Chr01: 18.49-19.96 Mb (specific to VV population), Chr07: 8.85-8.86
 369 Mb (shared, SDI2 locus), Chr18: 11.7-20.0 Mb (specific to VV×VL population), and
 370 Chr18: 23.9-35.5 Mb (shared, SDInew locus); (4) a total of 339 core genes associated
 371 with seedlessness was detected through integrative genomics analyses; (5) machine
 372 learning based genome selection were built to accurately predict the seedless
 373 phenotypes. Importantly, these findings could efficiently save the cost and time in
 374 table grape breeding.

375 The polygenic nature for seedlessness in grapes

376 The occurrence of seedlessness phenotypes from a cumulative polygenic effect
 377 associated with different tissues and development stages¹⁷. The limited observations
 378 employing single methods, such as transcriptomic analysis, are insufficient. As a
 379 result, numerous important candidate genes were ignored by previous studies. For
 380 example, *TT16/ABS*, found at the inversion boundary through comparative genomics
 381 (**Fig. 1d**), is one of promising candidate genes. Due to the redundant function between
 382 *STK* and *SHATTERPROOF* (*SHP1* and *SHP2*), the double *abs stk* mutants and triple

383 *tt16 shp1 shp2* mutants all induced fewer seeds and exhibited defects in seed coat
 384 formation^{17,50}. Candidate genes mapped by GWAS, such as *SMC1*, *11S globulin*
 385 *G1/2/3*, *LTG1*, *SUS2*, *LAC14*, *TT10/LAC15*, *MANNAN*
 386 *ENDO-1,4-BETA-MANNOSIDASE 5 (MAN5)*, *PROBABLE FRUCTOKINASE-5*, *E3*
 387 *UBIQUITIN-PROTEIN LIGASE (DA2)*, *AGL62*, etc., and well-known gene *VviAGL11*,
 388 play crucial roles in pollen, endosperm and seed coat development (**Fig. 3a, b and**
 389 **Supplementary Table 11**). Additionally, three transcriptomic analyses, GO
 390 homologous genes, and previously reported family genes provide us with numerous
 391 significant candidate genes (**Fig. 5a and Supplementary Table 11**).

392 Importantly, to define the interconnections among multiple genomic analyses,
 393 integrative genomic analysis was applied in this case, screened out 339 core genes
 394 with high significance from thousands of candidate genes (**Fig. 5a and**
 395 **Supplementary Table 17**). Most of these genes are associated with three main tissues
 396 related to seed coat, endosperm, and embryo, as shown in **Fig. 5b-d**, and ten other
 397 development processes (**Supplementary Table 11**). For example, we identified
 398 numerous candidate genes involving in seed hormone regulation, such as *MOTHER*
 399 *OF FT AND TFL1 (MFT)* in ABA and GA pathways, *VvABI3-1*⁶⁸ in ABA pathway,
 400 *VvGH3.9*⁶⁹ in auxin pathways, and *VvMJE1*⁶⁴ in jasmonate pathway. Additionally,
 401 genes controlling the development of floral organs, especially pollen and stigma,
 402 significantly influence the subsequent ovule development. This includes candidate
 403 genes like *CYP78A5*, *VvMADS27*, and metal ion transport genes
 404 *METALLOTHIONEIN-LIKE PROTEINS MT1* and *MT3*⁷⁰. Except for lipid
 405 accumulation (**Fig. 5d**), the processes of sugar and amino acid synthesis also
 406 contribute to seed development, such as *BGLU15*, *VviGAPDH*⁷¹, glycine-rich protein
 407 (*Vitvi021557*, *Vitvi036421*), and 36.4 kDa proline-rich protein (*Vitvi002605*). Overall,
 408 these findings not only emphasize the polygenic nature of seedlessness but also
 409 provide novel candidate genes for functional genetics, highlighting the complexity of
 410 seed abortion regulation.

411 The implications of genomic breeding of seedless table grapes

412 Both individual markers for Marker-Assisted Selection (MAS)^{44,72-74} and marker set
 413 for genomic selection generated in this study could be efficiently used in table grape
 414 breeding (**Fig. 6d**). Interestingly, in the variation map around *VviAGL11*, we observed
 415 that all the previously developed molecular markers based on lineages with a narrow
 416 genetic background for marker-assisted in table grape breeding were filtered away,
 417 due to either the low genotyping quality or a high missing rate, such as SCF27⁷⁵,
 418 e7_VviAGL11⁴⁴, 5U_VviAGL11⁴⁴, P3_VviAGL11¹⁸, VMC7f2⁶¹, and VrSD10⁷⁶ (**Fig.**
 419 **4d**). Luckily, we designed a set of 12 markers, based on our GWAS analyses of
 420 natural population with species-wide genetic diversity. These markers could
 421 accurately delimit seeded and seedless grapes, achieving a precision rate >90% in
 422 nature populations (**Fig. 4e**).

423 Quantitative genetics analysis revealed 110 high-quality QTLs associated with
 424 seedlessness in grapevines, including 634 candidate genes (**Supplementary Table 11**).
 425 Through extracting GWAS significant variants from admixed population, we obtained
 426 detail information on all 794 significant variant sites for training models. Genome
 427 selections, employing machine learning algorithms, achieved an impressive precision
 428 of 99%. This approach could facilitate early genomic selection of natural germplasms
 429 and hybrid progeny. Many crops, such as tomato⁷⁷, potato^{78,79}, cereal⁸⁰, rice^{81,82}, wheat
 430 and maize^{83,84}, have successfully applied genomic selection and prediction in their
 431 breeding programs. However, genomic selection has rarely been used on grape
 432 breeding. In the future, numerous agronomic traits like fruit aroma, disease resistance,
 433 soluble solids content, and so on, could be integrated into a single GS chip, offering a
 434 powerful genomic tool for genomic design of grapevine breeding.

435

436 **Methods**

437 **Plant materials and genome sequencing**

438 To enrich the genetic diversity of seedless grape varieties, we collected fresh tissues
 439 of two seedless grape varieties, ‘Thompson Seedless’ (TS) and ‘Black Monukka’
 440 (BM), from the Anningqu Experimental Station (87°28′00″E, 45°56′00″ N) at the
 441 Xinjiang Academy of Agricultural Sciences in China. Genomic DNA was extracted
 442 from grape leaves using CTAB method, followed by purification with the QIAGEN
 443 Genomic kit (CAT#13343). For each sample, a total of 15 µg DNA was utilized for
 444 HiFi (SMRTbell) library preparation. The sheared DNA fragments (gTUBEs, Covaris,
 445 USA) underwent treatment with the SMRTbell Enzyme Cleanup Kit (Pacific
 446 Biosciences, CA, USA) and purification using AMPure PB Beads. The resulting
 447 libraries were employed for HiFi sequencing on a PacBio Sequel II instrument (CCS
 448 mode) with Sequencing Primer V2 and Sequel II Binding Kit 2.0 in Grandomics,
 449 yielding 59.78 Gbp and 60.35 Gbp of sequencing data for TS and BM, respectively.

450 For Hi-C library preparation, the fresh leaves were cut into 2 cm pieces and vacuum
 451 infiltrated with nuclei isolation buffer supplemented with 2% formaldehyde. The
 452 isolated nuclei were then digested with 100 units of restriction enzyme DpnII. The
 453 resulting Hi-C sequencing data amounted to 59.39 Gbp (TS) and 57.67 Gbp (BM) via
 454 the Illumina Novaseq/MGI-2000 platform. Additionally, the genomic DNA of 29
 455 grape samples was extracted from fresh leaves, and the high-depth WGS sequencing
 456 was carried out using the PE150 mode of the Illumina Navaseq 6000 platform. For
 457 genome annotation, 1 µg total RNA was extracted from mixed tissues, such as roots,
 458 buds, and leaves. cDNA library was used TruSeq RNA Library Preparation Kit
 459 (Illumina, USA) and sequenced with 150 bp pair-end reads on the Illumina Navaseq
 460 6000 platform.

461 **Genome assembly**

462 The detailed workflow of haplotype-resolved genome assembly and annotation is
 463 described in **Extended Data Fig. 1**. The full pipeline for genome assembly and gap
 464 filling can be found on our lab GitHub@zhouyflab (see Code availability). In brief,
 465 we utilized the Hi-C and HiFi data integrated assembly algorithm to generate contig
 466 reads by Hifiasm (v. 0.16.1-r375)⁸⁵. The contig reads were oriented and ordered to
 467 scaffold level using RagTag (v. 2.1.0)⁸⁶ with default parameters. Hi-C reads were also
 468 employed to anchor scaffolds onto chromosomes by Juicer 2.0⁸⁷ and Juicebox (v.
 469 1.11.08)⁸⁸. The adjusted genome at the chromosome level was generated using
 470 3D-DNA (v. 201008)⁸⁹. The detailed workflow of haplotype-resolved genome
 471 assembly and annotation is described in **Extended Data Fig. 1**. The full pipeline for
 472 genome assembly and gap filling can be found on our lab GitHub@zhouyflab (see
 473 Code availability). The chloroplast genome was de novo assembly using GetOrganelle
 474 toolkit (v1.7.7.0)⁹⁰ with K-mer parameters set to 21, 65, 105, and 127. For the
 475 high-quality mitochondrial genome, de novo assembly was performed using Flye
 476 (v2.9.2)⁹¹ with HiFi long reads. Furthermore, using PN_T2T genome⁴⁶, we utilized
 477 RagTag to assemble three chromosomes-level genomes based on scaffold level:
 478 ‘Cabernet Sauvignon’ (CS)⁵², ‘Black Corinth Seedless’ (BC seedless)⁹², and ‘Black
 479 Corinth Seeded’ (BC seeded)⁹².

480 **Assembly assessment**

481 HiFi data from TS and BM were performed to assess genome heterozygosity based on
 482 k-mer using GenomeScope2.0⁹³. Meanwhile, basic genome statistics were calculated
 483 using seqkit (v. 2.2.0)⁹⁴, which included genome length, N50, and GC content. The
 484 completeness of the haplotype genomes was evaluated using BUSCOs (v. 5.3.0)⁹⁵
 485 with the embryophyta_odb10 database. Merquy (v. 1.3, best_k=19)⁹⁶ was employed
 486 to evaluate the quality value (QV) and completeness of the haplotype genomes based
 487 on whole-genome sequencing data. The Hi-C interactive signals on grape genome

488 were visualized using Juicebox.

489 **Genome annotation**

490 The genome-wide annotation pipeline, identification of telomeres and centromeres,
491 and annotation of transposable elements (TEs) were referred from pervious study^{46,97}.
492 The statistic results of TEs classification through Pan-genome TE annotation⁴⁶ can be
493 found in **Supplementary Table 2**. Additional details on telomere regions, telomere
494 copy numbers, and centromere regions for each chromosome are provided in
495 **Supplementary Table 3**. The CS genome was annotated based on sequence similarity
496 using Liftoff (v. 1.6.3)⁹⁸ and PN_T2T annotation files.

497 **Comparative genomics**

498 For genome level variant calling, we selected a total of 15 grape genomes, including
499 ten seeded accessions: ‘PN40024’ (PN_T2T), ‘Cabernet Sauvignon’ (CS, hap1 and
500 hap2)⁹², ‘Muscat Hamburg’ (MH, hap1 and hap2), ‘Shine Muscat’ (SM, hap1 and
501 hap2), ‘Muscadinia rotundifolia’ (MR, hap1 and hap2)⁹⁹, BC Seeded⁹², as well as five
502 seedless: TS (hap1 and hap2), BM (hap1 and hap2), BC Seedless⁹². Among them, TS
503 and BM were newly sequenced in this study. Four recently synchronized
504 haplotype-resolved genomes of MH and SM will be published in another study.

505 All chromosome-level genomes were aligned with PN_T2T genome using Mummer4
506 (v. 4.0.0rc1)¹⁰⁰, and the results were visualized using Plotsr (v. 0.5.4)¹⁰¹ and Linux
507 based Gunplot. R script was used for visualizing the gene density (line) of the
508 reference genome (see Code availability). To validate the authenticity of SVs, the raw
509 reads of HiFi and Hi-C were mapped to their haplotype genomes using Minimap2.
510 The corresponding BAM files were extracted using SAMtools (v. 1.13)¹⁰² and then
511 inputted into the IGV software for inversions validation. In addition, the genome
512 consensus phylogenetic tree was constructed using OrthoFinder (v. 2.5.4)¹⁰³ based on
513 the single-copy orthologous genes from the whole genomes.

514 **WGS variation detection**

515 To genotype SNPs, InDels and SVs in 548 accessions, low-quality resequencing reads
 516 were removed using Fastp (v. 0.23.2)¹⁰⁴ with default parameters. The filtered data
 517 were then mapped to the CS reference genome using BWA (v. 0.7.17-r1188)¹⁰⁵.
 518 Non-uniquely mapped and duplicated reads were excluded using SAMtools and
 519 GATK (v.4.2.3.0)¹⁰⁶. Subsequently, SNPs and InDels calling were performed using
 520 GTX (v. 2.1.11, <http://www.gtxlab.com/product/cat>), followed by the merging of
 521 genotyping of the gVCF files into a VCF file. Delly (v. 1.1.6)¹⁰⁷ was used to SVs
 522 calling with default parameters. Basic filtering of the VCF file was performed using
 523 VCFtools (v. 0.1.16)¹⁰⁸ with the following parameters: --max-missing 0.8, --minGQ
 524 20, --min-alleles 2, --max-alleles 2, --minDP 4, --maxDP 1000, and --maf 0.0005.
 525 PLINK (v. 1.90b6.21)¹⁰⁹ was utilized to reduce the size of the VCF file and improve
 526 computational efficiency. Finally, we obtained a total of 4,462,797 SNPs, 443,812
 527 InDels, and 487,204 SVs from 548 grape accessions in the nuclear genome, 38,267
 528 variation sites (SNPs and InDels) from 314 grape accessions in the MT genome, and
 529 2,247 variation sites (SNPs and InDels) from 314 grape accessions in the Pltd
 530 genome.

531 **Population genetics analysis**

532 Phylogenetic analysis was conducted using 250,821 high-quality SNPs filtered for LD
 533 by PLINK, with parameters: --indep-pairwise 20 5 0.2 --geno 0.1. The phylogenetic
 534 tree was inferred by Iqtree (v. 2.1.4-beta)¹¹⁰ with parameters: -m GTR+I+G -bb 1000
 535 -bnni -alrt 1000 -st DNA, and the result were visualized using iTOLs¹¹¹. Similarly, the
 536 phylogenetic analysis of the MT and PT genomes yielded 9,647 and 758 high-quality
 537 SNPs using PLINK (--geno 0.2 --maf 0.001), respectively. The PCA and IBD analysis
 538 were preformed using PLINK and visualized using R scripts and Cytoscape (v.
 539 3.9.1)¹¹², respectively. VCFtools was employed to calculate the F_{ST} and π statistics at
 540 whole genome level with 20 kb window size. Population introgression analysis was

541 calculated using the "ABBABABAwindows.py" script from the "general_genomics"
542 tool (https://github.com/simonhmartin/genomics_general) with 20 kb window size.
543 The population branch statistic (PBS) assesses differentiation in the same branch of
544 the phylogenetic tree with 50 SNP per window between seeded and seedless grapes in
545 both VV and VV×VL populations¹¹³. PBScan was utilized to estimate population
546 differentiation using D_{xy} (-div 2).

547 **Genome-wide association study**

548 We selected three populations for GWAS analysis, including VV population (n = 352),
549 VV×VL population (n = 92), and the admixed population (n = 444). High-quality
550 variants were obtained using PLINK with MAF ≥ 0.05 and missing < 0.2, resulting in
551 2,086,600 variants (1,881,457 SNPs, 173,547 InDels, and 31,596 SVs), 1,419,982
552 variants (1,274,130 SNPs, 124,054 InDels, and 21,798 SVs), and 2,292,404 variants
553 (2,065,307 SNPs, 192,536 InDels, and 34,561 SVs), respectively. GWAS analysis
554 utilized the mixed linear model in GEMMA (v. 0.98.3)¹¹⁴ with the first three PCAs as
555 a random effect matrix. Variant positions and p-wald test statistics (P-value) were
556 extracted to generate Manhattan plots and Q-Q plots by R scripts. A Python script was
557 utilized to identify the associated genes within 5 kb windows of the significant
558 variants, which were above the significance threshold of $-\log_{10}(0.05/\text{Variant}$
559 $\text{Numbers})$.

560 GO enrichment analysis was employed using the web tool DAVID¹¹⁵, and the results
561 were visualized with R script. The SDI locus was identified through molecular
562 markers SNP-25.24 and SNP-26.93²⁰. The LD linkage heatmap was visualized using
563 LDBlockshow (v. 1.40)¹¹⁶. Moreover, DIAMOND (v. 2.0.15)¹¹⁷ was employed for
564 sequence comparisons of whole-genome homologous protein from 14 genomes
565 (Seeded: Seedless: Outgroup = 8:5:1). Sequence alignments of candidate proteins
566 were conducted using ClustalW in MEGA11¹¹⁸, and sequence visualization was
567 accomplished using GeneDoc (v. gd322700)¹¹⁹. The visualization of FPKM values,

gene structures, and mutation ratios was achieved using R scripts (see Code availability).

Multi-transcriptome analysis

Three independent transcriptome datasets were downloaded from the NCBI database (see **Supplementary Table 15**). The 76 samples encompassed six grape varieties: PN and TS, which underwent repeated testing from 20 DAF to 50 DAF over a two-year period; 'Italia' and HS, continuously evaluated from 7 to 42 DAF within a single year; and Himrod (VV×VL population) and Jinzao (VV population), continuously monitored from full flowering stage to grape maturity within a single year. These raw fastq data underwent quality control and data cleaning using Fastp with default parameters. Using the PN_T2T genome as a reference, transcriptomic assembly was conducted on the processed data using STAR (v. 1.5.2)¹²⁰. The R package DESeq2 (v.1.36.0)¹²¹ was utilized for PAC analysis and pairwise comparisons between seeded and seedless grape samples at each development time point. The thresholds for differential expression genes were $|\text{Log2FoldChange}| \geq 2$ and $P\text{-adjust} < 0.05$.

Integrative genomic analysis

Homologous protein alignment identified 14,650 genes within the 34 seed development-related GO terms (EMBL-EBI, QuickGO database, 2023-01-03), and 6,529 genes with high sequence similarity were screened (identity $\geq 50\%$, e-value $\leq 1e-5$). Additionally, the primer sequences collected from previous studies in the past decade, including 451 genes and 7 molecular markers associated with seed abortion processes. All of them were then mapped to the PN_T2T genome based on sequence alignment using BALST in TBtools (v. 1.113)¹²². The results were visualized using Venn diagrams and whole-genome density plots using the R package ggvenn (v0.1.10) and RIdeogram¹²³, respectively. The integrated results in this study, including GWAS results, RNA-seq results, GO enrichment results, and previously reported family genes, were overlapped and filtered manually. The core candidate genes were

595 screened based on an average expression value across all time points (AVE FPKM $\geq$
596 100) and Fold-Changes (Fold = Seeded AVE FPKM / Seedless AVE FPKM, Fold $\leq$
597 0.5 or Fold $\geq$ 2.0). Data visualization was performed using an R script, which
598 included UpSet plot analysis and gene expression heatmaps (see Code availability).

599 **Genome selection based on machine learning**

600 For genome selection, we utilized the admixed population (n = 444) as training set,
601 and additional 39 samples as testing set for phenotype prediction, as pipeline showed
602 in **Extended Data Fig. 14**. Beagle (v. 5.2_21Apr21.304.jar)¹²⁴ was used to impute the
603 VCF file. Subsequently, the 794 variants, above the significance threshold of
604 $-\log_{10}(0.05/\text{Variant Numbers})$ (~ 7.66), were extracted from imputed VCF using
605 BCFtools (v. 1.13)¹²⁵, including 717 SNPs and 77 InDels. For model selection, we
606 utilized the Python package ‘sklearn’ (<https://scikit-learn.org/stable/install.html>) and
607 selected nine classical models: Cross-validated Elastic Net model (ElasticNetCV),
608 Kernel Ridge Regression (KernelRidge), Lasso Regression (Lasso), Linear
609 Regression (Linear), Logistic Regression (Logistic), PLS Regression, Linear Ridge
610 Regression (Ridge), Linear Support Vector Regression (SVR-Linear), and Polynomial
611 Support Vector Regression (SVR-poly). After 100 rounds of random cross-validation
612 with the training set, we chose the models with best performance, SVR-poly and
613 ElasticNetCV, for the prediction of testing set. Moreover, the Iqtree was employed to
614 construct an unrooted tree, while iTOLs was used for the visualization of the
615 phylogenetic tree. The data visualization used R scripts.

616 **Figure Legends**

617 **Fig. 1 | Comparative genomic between seeded and seedless grape cultivars. a,**
618 Visualization of the haplotype-resolved genomes aligned with the PN_T2T, with
619 discernible inversions marked in yellow. **b** and **c**, Morphological features of TS and
620 BM. **d**, Comparative genomic analysis of 15 grape genomes prioritized by the
621 phylogenetic tree constructed with single-copy gene features. Included genomes: MR

622 (Muscadinia rotundifolia), CS (Cabernet Sauvignon), MH (Muscat Hamburg), SH
623 (Shine Muscat), PN (PN40024), BC (Black Corinth). **e**, Detailed validation of a
624 prominent Chr10 segment inversion, with its distinctive features illustrated in the
625 Hi-C heatmap. **f**, Gene loss in hap2 genomes within the Chr10 inversion region.
626 *VvSUS2* and *VvTT16* are located near the inversion boundaries. Red blocks represent
627 lost genes, while gray blocks indicate genes shared between the two haplotype
628 genomes.

629 **Fig. 2 | Evolutionary genomics of seedlessness in grapevine.** **a**, Comparison of
630 grape clusters and cross-sectional views of berry from three cultivated grape varieties:
631 TS, BM, and CS. The scale bar indicates 5 cm for grape clusters and 1 cm for
632 individual berries. **b**, PCA analysis on the whole-genome sequencing (WGS) data,
633 except the outgroup, resulting in obviously separated five populations. Solid dots
634 denote seedless grapes, while transparent dots represent seeded grapes. CS genome
635 serves as the reference genome for variants calling. **c**, Phylogenetic tree analysis of
636 the six populations. The light-blue blocks represent seedless grapes and star symbols
637 indicate TS and BM (see Extended Data Fig. 6 for detailed information of the
638 phylogenetic tree). **d**, IBD analysis of 46 seedless grapes, filtering the results with a
639 threshold of 0.40. Purple points represent the *V. vinifera* × *V. labrusca* population,
640 while red points represent the *V. vinifera* population. **e**, The phylogenetic tree of 46
641 seedless grapes.

642 **Fig. 3 | GWAS mapping of the polygenic basis of seedlessness.** **a** and **b**,
643 Seedless-associated genomic loci and genes across three populations: the VV
644 population (in red, n=352), the VL population (in purple, n=92), and the admixed
645 population (in yellow and dark-blue, n = 444). The horizontal dashed lines denote the
646 Bonferroni thresholds ($-\log_{10}[0.05/\text{Variant Numbers}]$): 7.62 for VV, 7.45 for VV × VL,
647 and 7.66 for admixed population, respectively. Points represent SNPs, while triangles
648 represent SVs (and InDels). **c**, LD correlation analysis of a highly linked SDI2 locus
649 in Chr07, and genes associated with seed abortion are highlighted in red. **d**, Admixed

650 population analysis of a highly linked region in Chr18. The SDI locus (Chr18:
651 29.83-31.34 Mb) is identified based on SNP markers, and the redefined SDI locus
652 (SDInew, Chr18: 30.36-31.86 Mb) is defined based on GWAS results, depicted in the
653 grey block. The significant threshold: 7.61 for SNPs, and 6.66 for SVs (and InDels).
654 The dashed line for fixation indices (F_{ST}) indicates an average value of 0.126, with
655 genetic diversity (π) measured 0.0019 for VV×VL and 0.0016 for VV. The top 5th
656 percentile of f_d statistics is 0.277, with PBS statistics measuring 0.356 in VV and
657 0.415 in VV×VL. **e**, Gene flow pattern based on ABBA-BABA statistics, where SD
658 represents Seeded, SL represents Seedless.

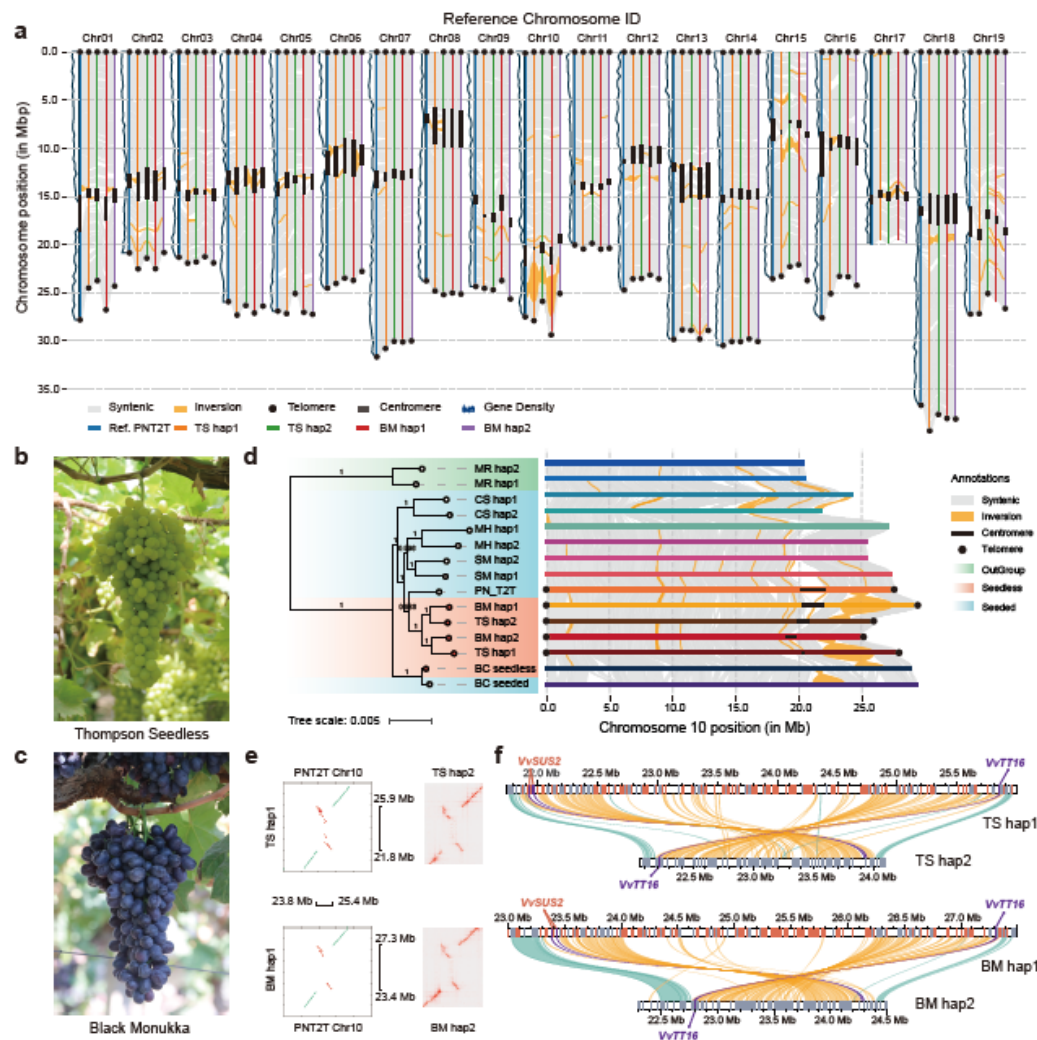
659 **Fig. 4 | Deep mining of key loci associated with seed abortion.** **a**, Sequence
660 alignment of three *11S GLOBULIN SEED STORAGE PROTEIN* genes in Chr07
661 across 14 grape varieties. Black blocks represent deletion and nonsynonymous
662 mutations. **b**, Relative expression values of these three genes at different time points
663 across six grape varieties. Abbreviations: DAF (Day after flowering), FF (Full
664 flowering), BE (Berry expansion), V (Veraison), M (Maturity). **c** and **d**, Visualization
665 of crucial candidate genes associated with seed abortion within the SDInew locus
666 region. GWAS significance thresholds ($-\log_{10}[0.05/\text{Variant Numbers}]$): 7.61 for SNPs,
667 and 6.66 for SVs (and InDels). The top ten percentiles of significant variants are
668 Chr18_31295826 ($y = 26.39$). **e**, Genotyping percentage of highly significant variants
669 is within the region Chr18: 30.70-31.32 Mb. ‘Seedless’ includes cases with genotypes
670 0/1 and 1/1, ‘Seeded’ includes the case with genotype 0/0.

671 **Fig. 5 | Integrative genomic analyses for grapevine seed abortion.** **a**, Results from
672 integrative genomic analyses: GWAS, transcriptomic analysis, reported genes and
673 markers mapping, and GO homologous genes overlapping. Blue bars represent the
674 number of genes uniquely and overlappedly through this approach. **b** and **c**, Heatmap
675 of $\text{Log}_2(\text{FPKM})$ for candidate genes related to embryo, endosperm, and seed coat
676 development, resulting from integrative genomic analyses. The genes are indicated in
677 relation to their expression in specific tissues and time points on schematic

678 representation transverse profiles of seeds. **d**, Expression patterns of genes involved in
 679 lipid synthesis, degradation and transportation, which also pinpointed on the
 680 schematic representation of oil body formation. Abbreviations: DAF (Day after
 681 flowering), FF (Full flowering), BE (Berry expansion), V (Veraison), M (Maturity).

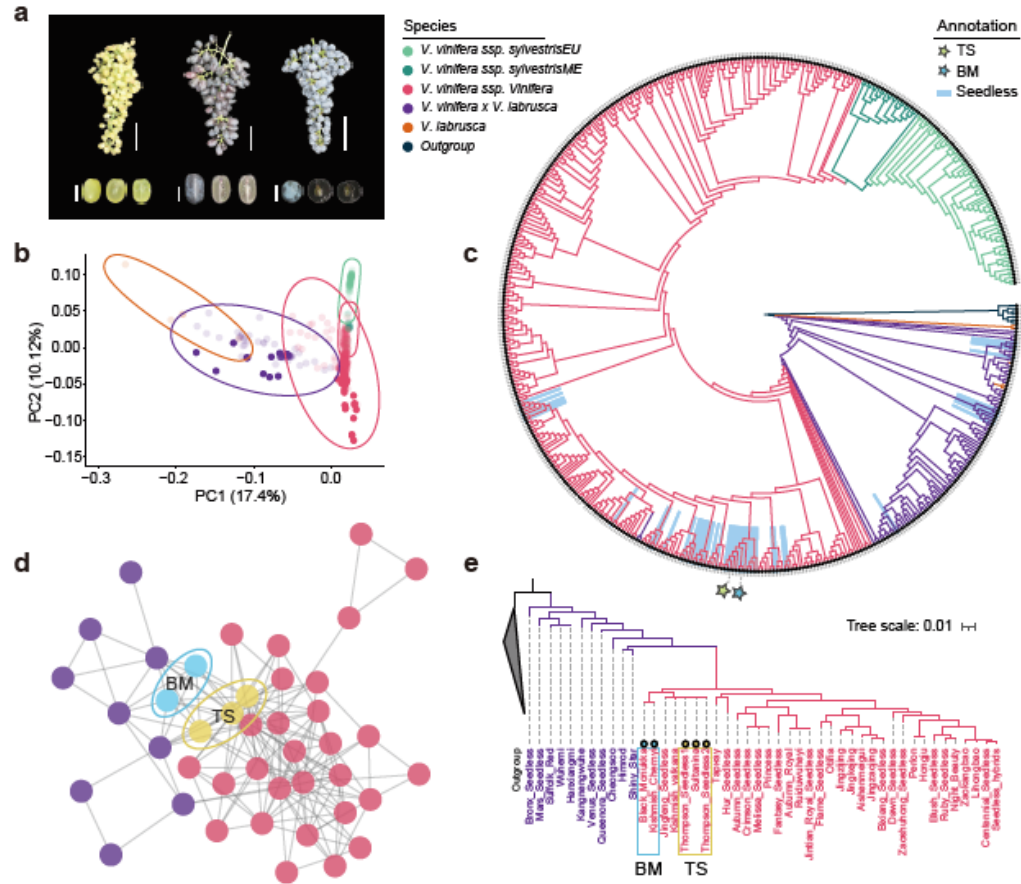
682 **Fig. 6 | Machine learning based genomic selection on seedlessness in grapevine**
 683 **breeding. a**, Phylogenetic clustering based on 794 significant variants (77 InDels and
 684 717 SNPs) derived from the GWAS analysis. **b**, Comparison of seedlessness
 685 prediction accuracy across nine classical models for genome selection. The training
 686 set comprises of 444 grape samples and their phenotypes. **c**, Prediction results of the
 687 two best-performing models. The testing set includes 39 samples, distinct from the
 688 444 samples used for model training. **d**, Genotyping visualization of the 794 variants,
 689 sorted based on prediction results of the SVR-ploy model.

690



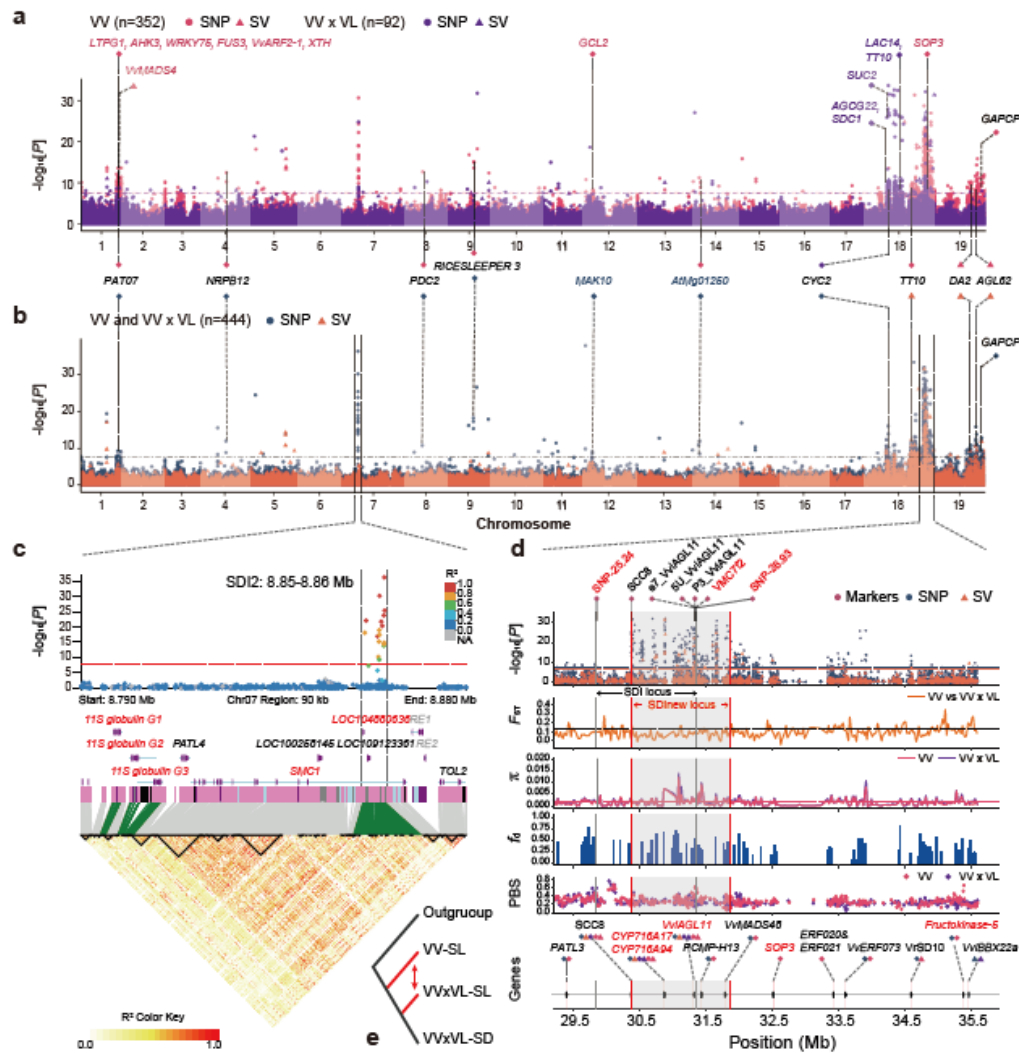
691

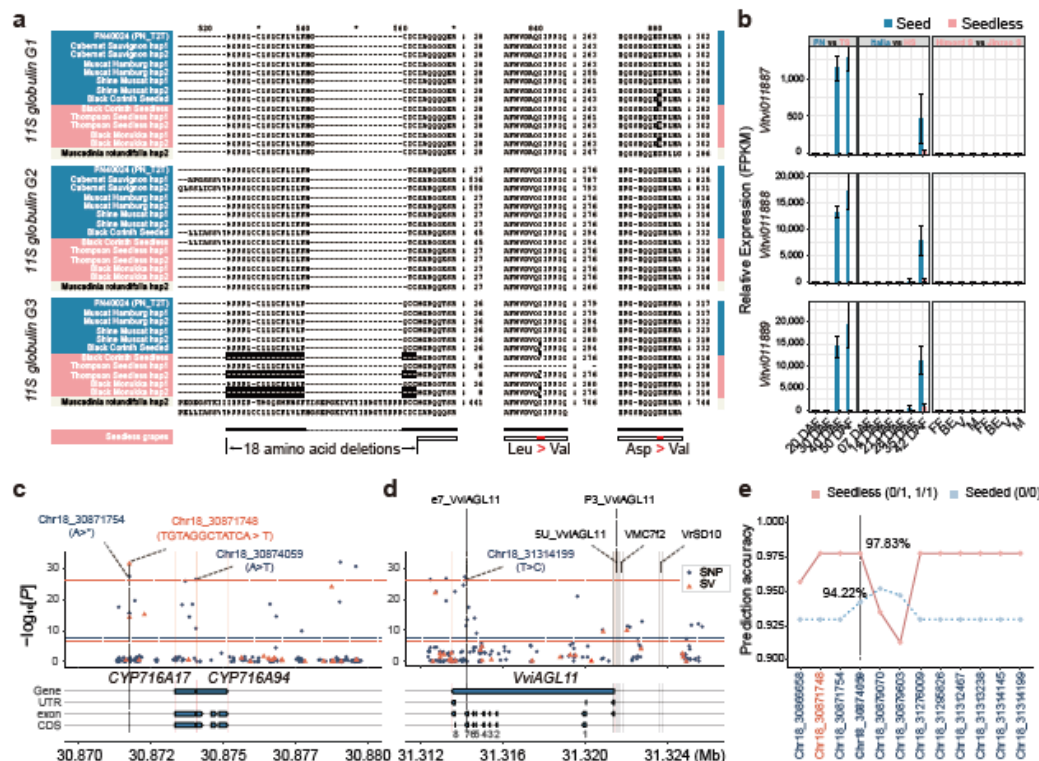
692



693

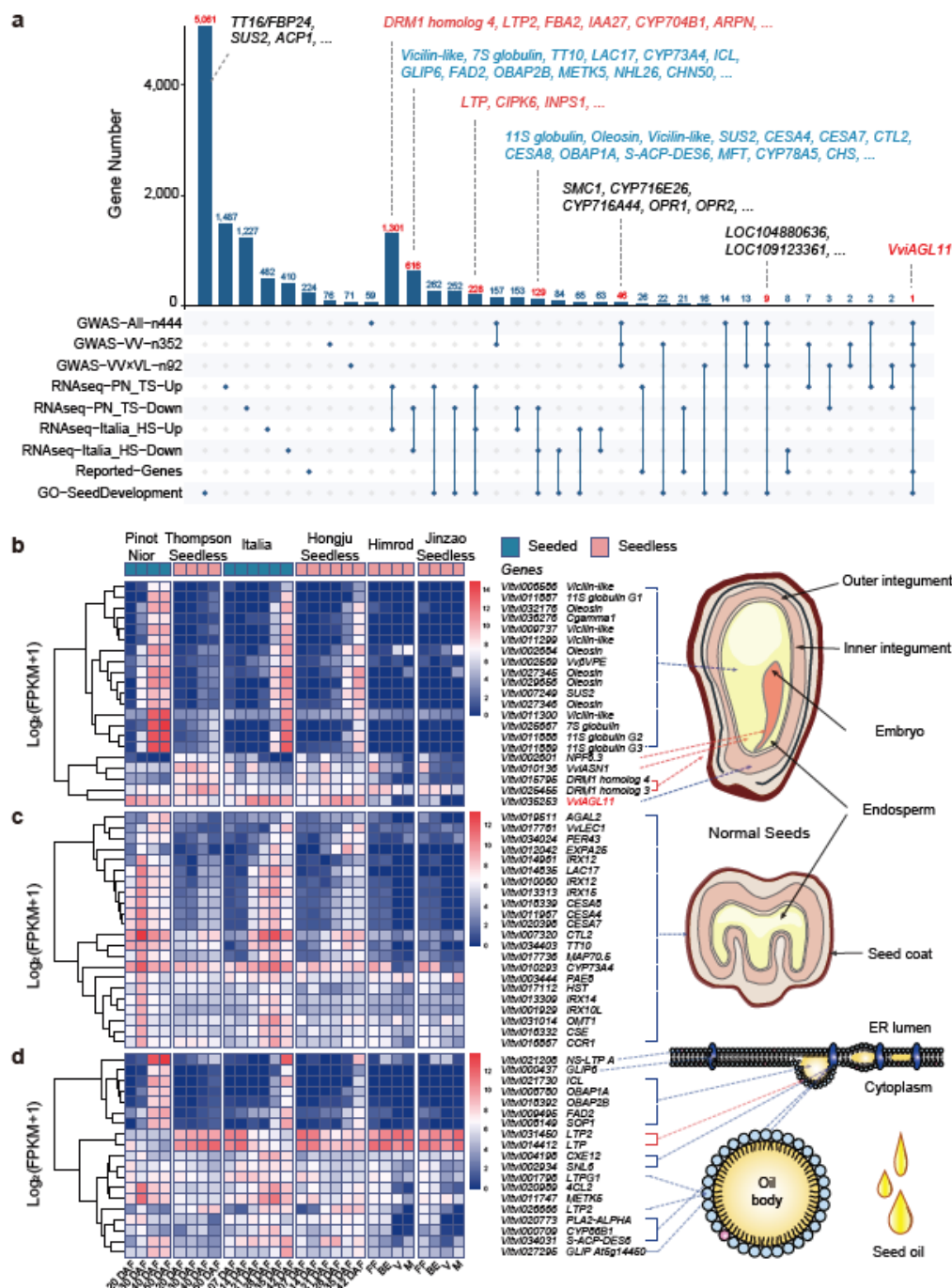
694

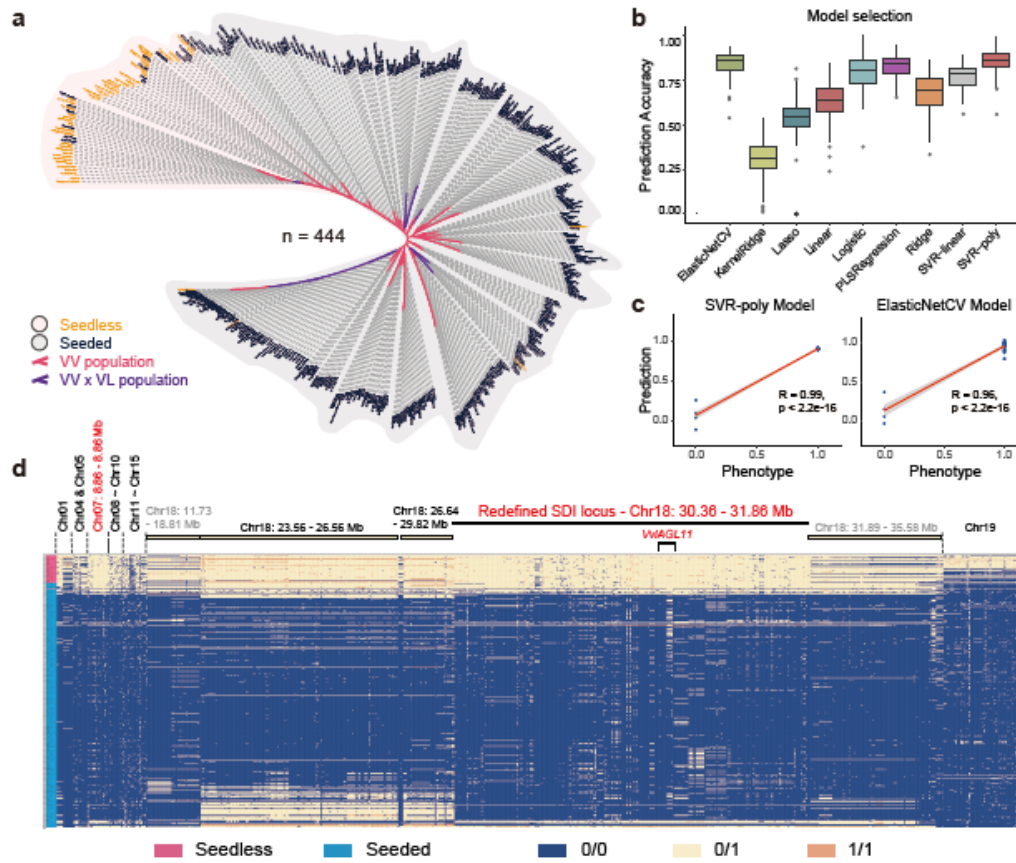




696

697





703 **Supplementary Data**

704 **Extended Data Fig. 1 | Complete workflow for haplotype-resolved genome**
 705 **assembly and annotation.** Additional details can be found on our lab website
 706 GitHub@zhouyflab.

707 **Extended Data Fig. 2 | Evaluation of four haplotype-resolved genomes. a,**
 708 BUSCO assessment of genome completeness using the embryophyta_odb10 database.
 709 **b,** Evaluation of quality value (QV) and haplotype completeness using Merqury based
 710 on k-mer. **c,** Visualization of the diploid genome via Hi-C heatmap using Juicebox.
 711 Most single contig (green rectangle) are directly composed of chromosomes (blue
 712 rectangles).

713 **Extended Data Fig. 3 | Comparative genomics results. a,** Sequence alignment of 15
 714 grape genomes with the PN_T2T genome. Red blocks represent seedless samples,
 715 green blocks indicate seeded samples, and gray blocks denote outgroup samples. Red
 716 stars highlight inversions associated with seed abortion. The Chr07 of '*Muscadinia*
 717 *rotundifolia*' is composed of Chr07 and Chr20. **b,** Sequence alignment results of
 718 Chr15 for the 15 genomes, as well as the inversion Hi-C heatmap in inversion
 719 boundary. The phylogenetic tree was constructed using single-copy genes from the
 720 whole genome proteins.

721 **Extended Data Fig. 4 | Reads mapping at the inversion breakpoints in seedless**
 722 **haplotype genome. a, c, and e,** represent the start points of the inversions, while **b, d,**
 723 and **f,** represent the end points of the inversions. Coverage depth is halved before and
 724 after the breakpoint junctions, revealing transitions between heterozygous and
 725 homozygous states of reads sequences are observed.

726 **Extended Data Fig. 5 | Detailed phylogenetic tree of 548 grapevine accessions.**
 727 This figure provides a full zoomed-in version of **Fig. 3c**, which includes the six
 728 populations. Light-blue blocks represent seed abortion samples and black star

729 symbols indicate TS and BM.

730 **Extended Data Fig. 6 | Phylogenetic tree of mitochondrial and chloroplast**
 731 **genomes in 314 grapevine accessions.** a-c, represent the consensus phylogenies
 732 constructed based on the mitochondrial genomes, nuclear genomes, and chloroplast
 733 genomes, respectively. d-e, depict the complete phylogenetic trees of the
 734 mitochondrial and chloroplast, encompassing six populations.

735 **Extended Data Fig. 7 | Visualization of whole-genome analyses.** From top to
 736 bottom, these results included QTL peaks (red: VV, purple: VV×VL, black:
 737 consensus peaks), GWAS analyses within three populations (red: VV, purple: VV×
 738 VL, dark-blue and yellow: admixed population; points: SNPs, triangles: InDels and
 739 SVs), population analyses of fixation indices (F_{ST}), nucleotide diversity (π),
 740 introgression (f_d), and divergent selection (PBS) (refer to Supplementary Table 13),
 741 and 339 core candidate genes (refer to Supplementary Table 11).

742 **Extended Data Fig. 8 | Quantile-Quantile (Q-Q) plot and Gene Ontology (GO)**
 743 **enrichment analysis.** Genome-Wide Association Study (GWAS) Q-Q plot for the
 744 three populations, along with GO enrichment analysis for biological processes in
 745 genes specific to VV×VL and VV, as well as those consensus gene for admixed
 746 populations (refer to Supplementary Table 12).

747 **Extended Data Fig. 9 | Sequence alignment of 11S globulin G1-G3 homologous**
 748 **genes from 14 grape genomes:** includes ‘PN_T2T’ (PN40024), ‘Cabernet Sauvignon’
 749 (CS), ‘Muscat Hamburg’ (MH), ‘Shine Muscat’ (SM), ‘Black Corinth Seeded’
 750 (BCsd), ‘Black Corinth Seedless’ (BCsl), ‘Thompson Seedless’ (TS), ‘Black
 751 Monukka’ (BM), and ‘Muscadinia Rotundifolia’.

752 **Extended Data Fig. 10 | Three candidate genes related to seedlessness in Chr18.**
 753 a-c, Candidate genes were identified through GWAS analysis using the admixed
 754 population, with a significant threshold of 7.61 for SNPs and 6.66 for SVs (and

755 InDels).

756 **Extended Data Fig. 11 | Principal Component Analysis (PCA) analysis of the**
 757 **three transcriptomic datasets, and relative expression values (FPKM) for**
 758 **VviAGL11.**

759 **Extended Data Fig. 12 | Enrichment analysis of GO homologous genes related to**
 760 **grape seed development. a**, Enrichment analysis of 14,650 seed development genes
 761 from the GO database. **b**, Enrichment analysis of 6,529 GO homologous genes
 762 associated with grape seed development.

763 **Extended Data Fig. 13 | Visualization of multiple seed development-related**
 764 **datasets. a** and **b**, Results from two independent transcriptomic analyses, ‘Italia’ vs
 765 ‘Hongju Seedless’ (HS) and ‘Pinot Noir’ (PN) vs ‘Thompson Seedless’ (TS),
 766 overlapping with GO homologous gene. Red indicates up-regulated DEGs, blue
 767 represents down-regulated DEGs, and yellow denotes GO homologous genes
 768 associated with seedlessness. **c**, Results from integrative genomic analyses: GWAS,
 769 transcriptomics, reported genes mapping, and GO homologous genes. Green bars
 770 represent the total number of genes identified through this approach. **d**, Visualization
 771 of three datasets: GO homologous genes, reported gene families, and reported
 772 molecular markers.

773 **Extended Data Fig. 14 | Genome selection workflow.** The 794 significant variants
 774 extracted from GWAS results, including 77 InDels and 717 SNPs. More detail code
 775 can be found on our lab GitHub@zhouyflab.

776

777 **Supplementary Table 1. Assessment of genome quality.** A comparison between
 778 four T2T haplotype-resolved genomes and the completed reference genome PN_T2T.

779 **Supplementary Table 2. Pan-genome TE annotation results in the four**
 780 **haplotype-resolved genomes.**

781 **Supplementary Table 3. Centromere and telomere regions in the four**
 782 **haplotype-resolved genomes.** This information includes the start position, end
 783 position, copy numbers, and TRF ID.

784 **Supplementary Table 4. Comparative genomic statistics.** Genome alignment was
 785 conducted using the SyRI (v. 1.5.4) with input from the output file generated by
 786 Mummer4 (v. 4.0.0rc1).

787 **Supplementary Table 5. Summary of genome annotation in the TS hap2 genome**
 788 **region Chr15: 8.72-9.90 Mb.** This region contains a total of 111 genes, including 73
 789 shared genes and 38 genes that are gained in the TS hap1 genome.

790 **Supplementary Table 6. Summary of genome annotation in the TS hap1 genome**
 791 **region Chr10: 21.75-26.00 Mb.** This region contains a total of 210 genes, including
 792 79 shared genes and 131 genes that are exclusively lost in the TS hap2 genome.

793 **Supplementary Table 7. Summary of genome annotation in the BM hap1**
 794 **genome region Chr10: 23.00-27.50 Mb.** This region contains a total of 237 genes,
 795 including 69 shared genes and 168 genes that are exclusively lost in the BM hap2
 796 genome.

797 **Supplementary Table 8. Population information for grape resequencing.** This
 798 table provides details, such as NCBI accessions, the full name of sample, species,
 799 seed conditions, population used in GWAS, samples used in phylogenetic trees,
 800 samples for population analyses (F_{ST} , π , f_d , and PBS), the training set and testing set
 801 for genome selection, as well as prediction results of best-performance models.

802 **Supplementary Table 9. Identity by Descent (IBD) matrix for 46 seedless**
 803 **individuals.** In this matrix, the VV population is denoted by purple, while VV×VL
 804 population is indicated by red. Detailed information for all samples used can be found
 805 in Supplementary Table 8.

806 **Supplementary Table 10. GWAS results of three populations.** This table includes
807 the variant positions, allele changes, -Log10 (*P* value), and associated genes (within ±
808 5 kb of the variant sites).

809 **Supplementary Table 11. Integrative genomic analysis across all study results.**
810 This table consolidates GWAS analysis results in different populations, differentially
811 expressed genes from three transcriptomic analyses, grape seed development
812 associated GO homologous genes, reported gene families and molecular markers, as
813 well as 339 core candidate genes identified through integrative genomic analysis. The
814 reference genome utilized ‘Cabernet Sauvignon’ (CS), and the homologous proteins
815 was aligned with the UniProt database and PN40024 12X (GCF_000003745.3),
816 respectively.

817 **Supplementary Table 12. Distinct and consensus genes, and GO enrichment**
818 **analyses in three populations.** This table presents candidate genes specific to the VV
819 population, those specific to VV×VL population, and those consensus in the admixed
820 population as identified through GWAS. GO enrichment analysis was conducted
821 using the online toolkit DAVID (<https://david.ncifcrf.gov/tools.jsp>).

822 **Supplementary Table 13. Population analyses results.** This table includes the
823 genome-wide fixation indices (F_{ST}) analysis for 11 VV and 35 VV×VL seedless
824 samples, genetic diversity (π) analysis for two population (11 VV and 35 VV×VL
825 seedless samples), f_d statistics analysis for gene introgression, and population branch
826 statistic (PBS) analysis for VV and VV×VL populations. Samples were selected from
827 a branch in the phylogenetic tree, with wild grapes (ME) as the outgroup. The
828 statistical window size is 20 kb for F_{ST} , π and f_d , while 50 SNPs per window size for
829 PBS analysis. Population information can be found in Supplementary Table 8.

830 **Supplementary Table 14. Mutation ratio statistics of significant variants**
831 **identified from GWAS analysis within Chr18:30.70-31.32 Mb in the admixed**
832 **population.**

833 **Supplementary Table 15. Information on three transcriptomic groups, including**
834 **NCBI accessions, full name of samples, time points, species and so on.**

835 **Supplementary Table 16. Primer sequences and data sources for 451 family**
836 **genes and 7 molecular markers, mapping proteins with the PN_T2T genome.**

837 **Supplementary Table 17. Overlapped genes statistics used for upset plot.** Further
838 gene details can be found in Supplementary Table 11.

839 **Supplementary Table 18. Genotyping information of 794 high-quality variants**
840 **(77 InDels and 717 SNPs).** This table also includes prediction results using genome
841 selection based on the SVR-poly model. ‘0’ represents 0/0, ‘1’ represents 0/1, and ‘2’
842 represents 1/1.

843 **Acknowledgements**

844 We express our gratitude to Zhiwu Zhang of Washington State University, USA, for
845 providing invaluable suggestions regarding genome selection and prediction. We
846 thank members of the Zhou lab at Agricultural Genomics Institute at Shenzhen (AGIS)
847 for discussions and comments on the project. We thank Lin Cheng, Hongbo Li, Zhigui
848 Bao, and Qing Zhang at AGIS for giving guidance and scripts to haplotype genome
849 assembly.

850 **Author contributions**

851 Y. Z. conceived and designed the study. H. Z., F. Z. and X. W. collected the plant
852 materials. H. Z. and Z. L. performed experiments and genome sequencing. X. W., S.
853 C., Y. S., T. H., and Y. Z assembled and annotated the haplotype-resolved T2T
854 genomes. X. W., Z. L., F. Z., X. H., W. L., Z. L., and Y. G. performed data analysis. Y.
855 Z., X. W., Z. L., F. Z., H. X. wrote the original draft of the manuscript. All authors
856 provided critical feedback and revised the manuscript. X. W. and Z. L. contributed
857 equally to this work.

858 **Competing interests**

859 The authors declare no competing interests.

860 **Data availability**

861 The raw sequencing data, comprising PacBio HiFi long-reads, Illumina Hi-C reads,
862 RNA-seq reads, and 29 WGS grape accessions, is accessible on NCBI under
863 BioProject ID PRJNA1021353 and on the National Genomics Data Center (NGDC)
864 under BioProject ID PRJCA022010. The genome assembly and their annotations have
865 been deposited into in Zenodo: <https://doi.org/10.5281/zenodo.8278185>.

866 **Code availability**

867 All scripts performed in this study are available on GitHub:

868 https://github.com/zhouyflab/Polygenetic_Basis_Seedless_Grapes

869 **Funding**

870 This work was supported by the National Key Research and Development Program of
871 China (No. 2023YFF1000100; 2023YFD2200700), the National Natural Science
872 Fund for Excellent Young Scientists Fund Program (Overseas) to Yongfeng Zhou, the
873 National Natural Science Fund (No. 32372662), the Project of Fund for Stable
874 Support to Agricultural Sci-Tech Renovation (xjnkywdzc-2022009), Xinjiang Uygur
875 Autonomous Region Tianchi Talent - Special Expert Project (Whole genome design
876 and breeding of grapes), and the central government guides local science and
877 technology development special fund projects (2022)-Germplasm Innovation and
878 Breeding Ability Improvement of Characteristic Fruit Trees.

879

880

881 **Reference**

- 882 1. Varoquaux, F., Blanvillain, R., Delseny, M. & Gallois, P. Less is better: new
883 approaches for seedless fruit production. *Trends in biotechnology* **18**, 233-242 (2000).
- 884 2. Sardos, J. *et al.* A genome-wide association study on the seedless phenotype in
885 banana (*Musa* spp.) reveals the potential of a selected panel to detect candidate
886 genes in a vegetatively propagated crop. *PLoS One* **11**, e0154448 (2016).
- 887 3. Sidhu, J.S. & Zafar, T.A. Bioactive compounds in banana fruits and their health
888 benefits. *Food Quality and Safety* **2**, 183-188 (2018).
- 889 4. Ye, W. *et al.* Seedless mechanism of a new mandarin cultivar 'Wuzishatangju'(Citrus
890 reticulata Blanco). *Plant Science* **177**, 19-27 (2009).
- 891 5. Zhang, S. *et al.* Comparative transcriptome analysis during early fruit development
892 between three seedy citrus genotypes and their seedless mutants. *Horticulture*
893 *research* **4**(2017).
- 894 6. Andrus, C.F., Seshadri, V. & Grimball, P.C. *Production of seedless watermelons*, (US
895 Agricultural Research Service, 1971).
- 896 7. Wijesinghe, S., Evans, L., Kirkland, L. & Rader, R. A global review of watermelon
897 pollination biology and ecology: The increasing importance of seedless cultivars.
898 *Scientia Horticulturae* **271**, 109493 (2020).
- 899 8. Akkurt, M., Tahmaz, H. & Veziroğlu, S. Recent developments in seedless grapevine
900 breeding. *South African Journal of Enology and Viticulture* **40**, 1-1 (2019).
- 901 9. Reynolds, A., Wardle, D., Zurowski, C. & Looney, N. Phenylureas CPPU and

902 thidiazuron affect yield components, fruit composition, and storage potential of four
 903 seedless grape selections. *Journal of the American Society for Horticultural Science*
 904 **117**, 85-89 (1992).

905 10. Cheng, C. *et al.* Effect of GA3 treatment on seed development and seed-related gene
 906 expression in grape. *PLoS One* **8**, e80044 (2013).

907 11. Tyagi, K. *et al.* Cytokinin but not gibberellin application had major impact on the
 908 phenylpropanoid pathway in grape. *Horticulture research* **8**(2021).

909 12. Park, Y.-S., Lee, J.-C., Jeong, H.-N., Um, N.-Y. & Heo, J.-Y. A red triploid seedless
 910 grape 'Red Dream'. *HortScience* **57**, 741-742 (2022).

911 13. Park, S., Hiramatsu, M. & Wakana, A. Aneuploid plants derived from crosses with
 912 triploid grapes through immature seed culture and subsequent embryo culture. *Plant*
 913 *cell, tissue and organ culture* **59**, 125-133 (1999).

914 14. Ji, W., Li, Z., Zhou, Q., Yao, W. & Wang, Y. Breeding new seedless grape by means of
 915 in vitro embryo rescue. *Genetics and Molecular Research* **12**, 859-869 (2013).

916 15. Li, J., Wang, X., Wang, X. & Wang, Y. Embryo rescue technique and its applications
 917 for seedless breeding in grape. *Plant Cell, Tissue and Organ Culture (PCTOC)* **120**,
 918 861-880 (2015).

919 16. Yamashita, H., Shigehara, I. & Haniuda, T. Production of triploid grapes by in ovulo
 920 embryo culture. *VITIS-Journal of Grapevine Research* **37**, 113 (2015).

921 17. Ehlers, K. *et al.* The MADS box genes ABS, SHP1, and SHP2 are essential for the
 922 coordination of cell divisions in ovule and seed coat development and for endosperm

923 formation in *Arabidopsis thaliana*. *PLoS one* **11**, e0165075 (2016).

924 18. Mejia, N. *et al.* Molecular, genetic and transcriptional evidence for a role of VvAGL11
925 in stenospermocarpic seedlessness in grapevine. *BMC plant biology* **11**, 1-19 (2011).

926 19. Malabarba, J. *et al.* The MADS-box gene Agamous-like 11 is essential for seed
927 morphogenesis in grapevine. *Journal of experimental botany* **68**, 1493-1506 (2017).

928 20. Royo, C. *et al.* The Major Origin of Seedless Grapes Is Associated with a Missense
929 Mutation in the MADS-Box Gene VviAGL11 *Plant Physiology* **177**, 1234-1253
930 (2018).

931 21. Amato, A. *et al.* VviAGL11 self-regulates and targets hormone-and secondary
932 metabolism-related genes during seed development. *Horticulture Research* **9**(2022).

933 22. Zhang, S. *et al.* Control of ovule development in *Vitis vinifera* by VvMADS28 and
934 interacting genes. *Horticulture Research*, uhad070 (2023).

935 23. Lora, J., Hormaza, J.I., Herrero, M. & Gasser, C.S. Seedless fruits and the disruption
936 of a conserved genetic pathway in angiosperm ovule development. *Proceedings of the*
937 *National Academy of Sciences* **108**, 5461-5465 (2011).

938 24. di Rienzo, V. *et al.* Functional conservation of the grapevine candidate gene INNER
939 NO OUTER for ovule development and seed formation. *Horticulture research* **8**(2021).

940 25. Li, Y. *et al.* Genome-wide identification and expression analyses of the homeobox
941 transcription factor family during ovule development in seedless and seeded grapes.
942 *Scientific Reports* **7**, 12638 (2017).

943 26. Li, Y. *et al.* The grapevine homeobox gene VvHB58 influences seed and fruit

944 development through multiple hormonal signaling pathways. *BMC plant biology* **19**,
945 1-18 (2019).

946 27. Yao, J. *et al.* KNOX transcription factor VvHB63 affects grape seed development by
947 interacting with protein VvHB06. *Plant Science* **330**, 111665 (2023).

948 28. Gazzola, D. *et al.* The proteins of the grape (*Vitis vinifera* L.) seed endosperm:
949 Fractionation and identification of the major components. *Food Chemistry* **155**,
950 132-139 (2014).

951 29. Chamizo-González, F., Heredia, F.J., Rodríguez-Pulido, F.J., González-Miret, M.L. &
952 Gordillo, B. Proteomic and computational characterisation of 11S globulins from grape
953 seed flour by-product and its interaction with malvidin 3-glucoside by molecular
954 docking. *Food Chemistry* **386**, 132842 (2022).

955 30. Chamizo-González, F. *et al.* First insights into the binding mechanism and colour
956 effect of the interaction of grape seed 11S globulin with malvidin 3-O-glucoside by
957 fluorescence spectroscopy, differential colorimetry and molecular modelling. *Food*
958 *Chemistry* **413**, 135591 (2023).

959 31. He, H. *et al.* Genome-wide identification and expression analysis of GA2ox, GA3ox,
960 and GA20ox are related to gibberellin oxidase genes in grape (*Vitis Vinifera* L.).
961 *Genes* **10**, 680 (2019).

962 32. Bai, Y. *et al.* miR3633a-GA3ox2 Module Conducts Grape Seed-Embryo Abortion in
963 Response to Gibberellin. *International Journal of Molecular Sciences* **23**, 8767 (2022).

964 33. Zhang, S. *et al.* Role of grapevine SEPALLATA-related MADS-box gene VvMADS39

965 in flower and ovule development. *The Plant Journal* **111**, 1565-1579 (2022).

966 34. Sun, X. *et al.* A MADS-box transcription factor from grapevine, VvMADS45, influences
967 seed development. *Plant Cell, Tissue and Organ Culture (PCTOC)* **141**, 105-118
968 (2020).

969 35. Tang, Y. *et al.* Differential expression of the seed-specific gene ABCG20 between
970 seedless and seeded grapes and its roles in tomato seed development. *South African*
971 *journal of botany* **131**, 428-436 (2020).

972 36. Wang, L., Dai, W., Shi, Y., Wang, Y. & Zhang, C. Cloning and activity analysis of the
973 highly expressed gene VviABCG20 promoter in seed and its activity is negatively
974 regulated by the transcription factor VviDof14. *Plant Science* **315**, 111152 (2022).

975 37. Wang, L. *et al.* The putative ABCG transporter VviABCG20 from grapevine (*Vitis*
976 *vinifera*) is strongly expressed in the seed coat of developing seeds and may
977 participate in suberin biosynthesis. *Physiology and Molecular Biology of Plants*, 1-12
978 (2023).

979 38. Ahmad, B. *et al.* Ectopic expression of VvFUS3, B3-domain transcription factor, in
980 tomato influences seed development via affecting endoreduplication and hormones.
981 *Horticultural Plant Journal* **8**, 351-360 (2022).

982 39. Zhang, S. *et al.* NAC domain gene VvNAC26 interacts with VvMADS9 and influences
983 seed and fruit development. *Plant Physiology and Biochemistry* **164**, 63-72 (2021).

984 40. Tang, Y. *et al.* Gene cloning, expression and enzyme activity of *Vitis vinifera* vacuolar
985 processing enzymes (VvVPEs). *PloS one* **11**, e0160945 (2016).

986 41. Gong, P. *et al.* Molecular cloning and functional characterization of a seed-specific
987 VvβVPE gene promoter from *Vitis vinifera*. *Planta* **250**, 657-665 (2019).

988 42. Wang, L. *et al.* Genome-wide identification and comprehensive expression analysis of
989 VviASN and VviGS gene families during seed development/abortion in grapevine.
990 *Scientia Horticulturae* **292**, 110625 (2022).

991 43. Cabezas, J.A., Cervera, M.T., Ruiz-García, L., Carreno, J. & Martínez-Zapater, J.M. A
992 genetic analysis of seed and berry weight in grapevine. *Genome* **49**, 1572-1585
993 (2006).

994 44. Ocarez, N. *et al.* Unraveling the deep genetic architecture for seedlessness in
995 grapevine and the development and validation of a new set of markers for
996 VviAGL11-based gene-assisted selection. *Genes* **11**, 151 (2020).

997 45. Kim, M.-S., Hur, Y.Y., Kim, J.H. & Jeong, S.-C. Genome resequencing, improvement
998 of variant calling, and population genomic analyses provide insights into the
999 seedlessness in the genus *vitis*. *G3: Genes, Genomes, Genetics* **10**, 3365-3377
1000 (2020).

1001 46. Shi, X. *et al.* The complete reference genome for grapevine (*Vitis vinifera* L.) genetics
1002 and breeding. *Horticulture Research* **10**, uhad061 (2023).

1003 47. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV):
1004 high-performance genomics data visualization and exploration. *Briefings in*
1005 *bioinformatics* **14**, 178-192 (2013).

1006 48. Nesi, N. *et al.* The TRANSPARENT TESTA16 locus encodes the ARABIDOPSIS

1007 BSISTER MADS domain protein and is required for proper development and
1008 pigmentation of the seed coat. *The Plant Cell* **14**, 2463-2479 (2002).

1009 49. De Folter, S. *et al.* A Bsister MADS-box gene involved in ovule and seed development
1010 in petunia and Arabidopsis. *The Plant Journal* **47**, 934-946 (2006).

1011 50. Mizzotti, C. *et al.* The MADS box genes SEEDSTICK and ARABIDOPSIS Bsister play
1012 a maternal role in fertilization and seed development. *The Plant Journal* **70**, 409-420
1013 (2012).

1014 51. Gupta, M. *et al.* Grape seed extract: Having a potential health benefits. *Journal of food*
1015 *science and technology* **57**, 1205-1215 (2020).

1016 52. Zhou, Y. *et al.* The population genetics of structural variants in grapevine
1017 domestication. *Nature plants* **5**, 965-979 (2019).

1018 53. Martin, S.H., Davey, J.W. & Jiggins, C.D. Evaluating the Use of ABBA-BABA
1019 Statistics to Locate Introgressed Loci. *Molecular Biology and Evolution* **32**, 244-257
1020 (2014).

1021 54. Wang, N. *et al.* Genomic conservation of crop wild relatives: A case study of citrus.
1022 *PLoS genetics* **19**, e1010811 (2023).

1023 55. Xiao, H. *et al.* Adaptive and maladaptive introgression in grapevine domestication.
1024 *Proceedings of the National Academy of Sciences* **120**, e2222041120 (2023).

1025 56. Maul, E. *et al.* 30 Years VIVC-Vitis International Variety Catalogue (www.vivc.de). in
1026 *XI International Conference on Grapevine Breeding and Genetics, Yanqing, Beijing,*
1027 *China, July 28-August 2, 2014* (2014).

1028 57. Ledbetter, C. & Ramming, D. Seedlessness in grapes. *Horticultural Reviews* **11**,
1029 159-184 (1989).

1030 58. Wang, N. *et al.* Pan-mitogenomics reveals the genetic basis of cytonuclear conflicts in
1031 citrus hybridization, domestication, and diversification. *Proceedings of the National*
1032 *Academy of Sciences* **119**, e2206076119 (2022).

1033 59. Liu, C.m. *et al.* Condensin and cohesin knockouts in Arabidopsis exhibit a titan seed
1034 phenotype. *The Plant Journal* **29**, 405-415 (2002).

1035 60. Liu, C. *et al.* Optimization of extraction and isolation for 11S and 7S globulins of
1036 soybean seed storage protein. *Food chemistry* **102**, 1310-1316 (2007).

1037 61. Thomas, M., Edwards, K. & Pellerone, F. Grapevine microsatellite repeats: Isolation,
1038 characterisation and use of genotyping of grape germplasm from Southern Italy. *Vitis:*
1039 *Journal of Grapevine Research* **40**, 179-186 (2001).

1040 62. Wang, L. *et al.* Evolutionary and expression analysis of a MADS-box gene superfamily
1041 involved in ovule development of seeded and seedless grapevines. *Molecular*
1042 *Genetics and Genomics* **290**, 825-846 (2015).

1043 63. Sheng, Z. *et al.* Identification and Characterization of AUXIN Response Factor Gene
1044 Family Reveals Their Regulatory Network to Respond the Multi-Hormones Crosstalk
1045 during GA-Induced Grape Parthenocarpic Berry. *International Journal of Molecular*
1046 *Sciences* **23**, 11108 (2022).

1047 64. Zhao, N. *et al.* VvMJE1 of the grapevine (*Vitis vinifera*) VvMES methyl esterase family
1048 encodes for methyl jasmonate esterase and has a role in stress response. *Plant*

1049 *Physiology and Biochemistry* **102**, 125-132 (2016).

1050 65. Cui, M. *et al.* Characterization and temporal-spatial expression analysis of LEC1 gene
1051 in the development of seedless berries in grape induced by gibberellin. *Plant Growth*
1052 *Regulation* **90**, 585-596 (2020).

1053 66. Wang, Y. *et al.* MADS-Box Protein Complex VvAG2, VvSEP3 and VvAGL11
1054 Regulates the Formation of Ovules in *Vitis vinifera* L. cv.'Xiangfei'. *Genes* **12**, 647
1055 (2021).

1056 67. Yim, B. *et al.* Anatomical, biochemical and transcriptome analyses of *Vitis vinifera*
1057 cv.'Hongju'reveal novel information regarding the seed hardness of
1058 stenospermocarpic soft-seed grapes. *Plant Breeding* **139**, 672-683 (2020).

1059 68. Ahmad, B. *et al.* Genomic organization of the B3-domain transcription factor family in
1060 grapevine (*Vitis vinifera* L.) and expression during seed development in seedless and
1061 seeded cultivars. *International Journal of Molecular Sciences* **20**, 4553 (2019).

1062 69. Khan, S. & Stone, J. Arabidopsis thaliana GH3. 9 in auxin and jasmonate cross talk.
1063 *Plant signaling & behavior* **2**, 483-485 (2007).

1064 70. Brkljačić, J.M., Samardžić, J.T., Timotijević, G.S. & Maksimović, V.R. Expression
1065 analysis of buckwheat (*Fagopyrum esculentum* Moench) metallothionein-like gene
1066 (MT3) under different stress and physiological conditions. *Journal of plant physiology*
1067 **161**, 741-746 (2004).

1068 71. Tang, Y., Huang, C., Li, Y., Wang, Y. & Zhang, C. Genome-wide identification,
1069 phylogenetic analysis, and expression profiling of glycine-rich RNA-binding protein

1070 (GRPs) genes in seeded and seedless grapes (*Vitis vinifera*). *Physiology and*
1071 *Molecular Biology of Plants* **27**, 2231-2243 (2021).

1072 72. Akkurt, M., Çakir, A., Shidfar, M., Çelikkol, B. & Soylemezoglu, G. Using SCC8,
1073 SCF27 and VMC7f2 markers in grapevine breeding for seedlessness via marker
1074 assisted selection. (2012).

1075 73. Dong, Z. *et al.* Genetic relationships of 34 grapevine varieties and construction of
1076 molecular fingerprints by SSR markers. *Biotechnology & Biotechnological Equipment*
1077 **32**, 942-950 (2018).

1078 74. Wang, Y. *et al.* Embryo Rescue and Molecular Marker-Assisted Selection of Hybrid
1079 Seedless Grape. (2021).

1080 75. Mejia, N. & Hinrichsen, P. A new, highly assertive SCAR marker potentially useful to
1081 assist selection for seedlessness in table grape breeding. in *VIII International*
1082 *Conference on Grape Genetics and Breeding* 603 559-564 (2002).

1083 76. Ma, Y. *et al.* Development and application of SSR new molecular marker for seedless
1084 traits in grape. *Scientia Agricultura Sinica* **51**, 2622-2630 (2018).

1085 77. Slater, A.T., Cogan, N.O., Forster, J.W., Hayes, B.J. & Daetwyler, H.D. Improving
1086 genetic gain with genomic selection in autotetraploid potato. *The plant genome* **9**,
1087 plantgenome2016.02.0021 (2016).

1088 78. Cappetta, E. *et al.* Accelerating tomato breeding by exploiting genomic selection
1089 approaches. *Plants* **9**, 1236 (2020).

1090 79. Zhou, Y. *et al.* Graph pangenome captures missing heritability and empowers tomato

1091 breeding. *Nature* **606**, 527-534 (2022).

1092 80. Robertsen, C.D., Hjortshøj, R.L. & Janss, L.L. Genomic selection in cereal breeding.
1093 *Agronomy* **9**, 95 (2019).

1094 81. Grenier, C. *et al.* Accuracy of genomic selection in a rice synthetic population
1095 developed for recurrent selection breeding. *PloS one* **10**, e0136594 (2015).

1096 82. Xu, Y. *et al.* Genomic selection: A breakthrough technology in rice breeding. *The Crop*
1097 *Journal* **9**, 669-677 (2021).

1098 83. Crossa, J. *et al.* Genomic selection and prediction in plant breeding. *Journal of Crop*
1099 *Improvement* **25**, 239-261 (2011).

1100 84. Crossa, J. *et al.* Genomic selection in plant breeding: methods, models, and
1101 perspectives. *Trends in plant science* **22**, 961-975 (2017).

1102 85. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de
1103 novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**,
1104 170-175 (2021).

1105 86. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new
1106 tomato system for high-throughput genome editing. *Genome biology* **23**, 1-19 (2022).

1107 87. Durand, N.C. *et al.* Juicer provides a one-click system for analyzing loop-resolution
1108 Hi-C experiments. *Cell systems* **3**, 95-98 (2016).

1109 88. Durand, N.C. *et al.* Juicebox provides a visualization system for Hi-C contact maps
1110 with unlimited zoom. *Cell systems* **3**, 99-101 (2016).

1111 89. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C

1112 yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).

1113 90. Jin, J.-J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly
1114 of organelle genomes. *Genome biology* **21**, 1-31 (2020).

1115 91. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using
1116 repeat graphs. *Nature Methods* **17**, 1103-1110 (2020).

1117 92. Massonnet, M. *et al.* The genetic basis of sex determination in grapes. *Nature*
1118 *communications* **11**, 2902 (2020).

1119 93. Ranallo-Benavidez, T.R., Jaron, K.S. & Schatz, M.C. GenomeScope 2.0 and
1120 Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications*
1121 **11**, 1432 (2020).

1122 94. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for
1123 FASTA/Q file manipulation. *PloS one* **11**, e0163962 (2016).

1124 95. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M.
1125 BUSCO: assessing genome assembly and annotation completeness with single-copy
1126 orthologs. *Bioinformatics* **31**, 3210-3212 (2015).

1127 96. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality,
1128 completeness, and phasing assessment for genome assemblies. *Genome biology* **21**,
1129 1-27 (2020).

1130 97. Yue, J. *et al.* Telomere-to-telomere and gap-free reference genome assembly of the
1131 kiwifruit *Actinidia chinensis*. *Horticulture Research* **10**(2022).

1132 98. Shumate, A. & Salzberg, S.L. Liftoff: accurate mapping of gene annotations.

1133 *Bioinformatics* **37**, 1639-1643 (2021).

1134 99. Cochetel, N. *et al.* Diploid chromosome-scale assembly of the *Muscadinia rotundifolia*
1135 genome supports chromosome fusion and disease resistance gene expansion during
1136 *Vitis* and *Muscadinia* divergence. *G3 Genes/Genomes/Genetics* **11**(2021).

1137 100. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS*
1138 *computational biology* **14**, e1005944 (2018).

1139 101. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and
1140 rearrangements between multiple genomes. *Bioinformatics* **38**, 2922-2926 (2022).

1141 102. Li, H. *et al.* The sequence alignment/map format and SAMtools. *bioinformatics* **25**,
1142 2078-2079 (2009).

1143 103. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
1144 genomics. *Genome biology* **20**, 1-14 (2019).

1145 104. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ
1146 preprocessor. *Bioinformatics* **34**, i884-i890 (2018).

1147 105. Li, H. Aligning sequence reads, clone sequences and assembly contigs with
1148 BWA-MEM. *arXiv preprint arXiv:1303.3997*(2013).

1149 106. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
1150 analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303
1151 (2010).

1152 107. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and
1153 split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).

1154 108. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158
1155 (2011).

1156 109. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and
1157 richer datasets. *Gigascience* **4**, s13742-015-0047-8 (2015).

1158 110. Minh, B.Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic
1159 inference in the genomic era. *Molecular biology and evolution* **37**, 1530-1534 (2020).

1160 111. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic
1161 tree display and annotation. *Nucleic acids research* **49**, W293-W296 (2021).

1162 112. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of
1163 biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).

1164 113. Härmälä, T. & Savolainen, O. Genomic patterns of local adaptation under gene flow in
1165 *Arabidopsis lyrata*. *Molecular Biology and Evolution* **36**, 2557-2571 (2019).

1166 114. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association
1167 studies. *Nature genetics* **44**, 821-824 (2012).

1168 115. Sherman, B.T. *et al.* DAVID: a web server for functional enrichment analysis and
1169 functional annotation of gene lists (2021 update). *Nucleic acids research* **50**,
1170 W216-W221 (2022).

1171 116. Dong, S.-S. *et al.* LDBlockShow: a fast and convenient tool for visualizing linkage
1172 disequilibrium and haplotype blocks based on variant call format files. *Briefings in*
1173 *Bioinformatics* **22**, bbaa227 (2021).

1174 117. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using

1175 DIAMOND. *Nature methods* **12**, 59-60 (2015).

1176 118. Tamura, K., Stecher, G. & Kumar, S. MEGA11: molecular evolutionary genetics
1177 analysis version 11. *Molecular biology and evolution* **38**, 3022-3027 (2021).

1178 119. KB, N. GeneDoc: analysis and visualization of genetic variation. *EMBnet news* **4**, 1-4
1179 (1997).

1180 120. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21
1181 (2013).

1182 121. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and
1183 dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).

1184 122. Chen, C. *et al.* TBtools: an integrative toolkit developed for interactive analyses of big
1185 biological data. *Molecular plant* **13**, 1194-1202 (2020).

1186 123. Hao, Z. *et al.* Rldeogram: drawing SVG graphics to visualize and map genome-wide
1187 data on the ideograms. *PeerJ Computer Science* **6**, e251 (2020).

1188 124. Browning, B.L., Tian, X., Zhou, Y. & Browning, S.R. Fast two-stage phasing of
1189 large-scale sequence data. *The American Journal of Human Genetics* **108**, 1880-1890
1190 (2021).

1191 125. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008
1192 (2021).

1193