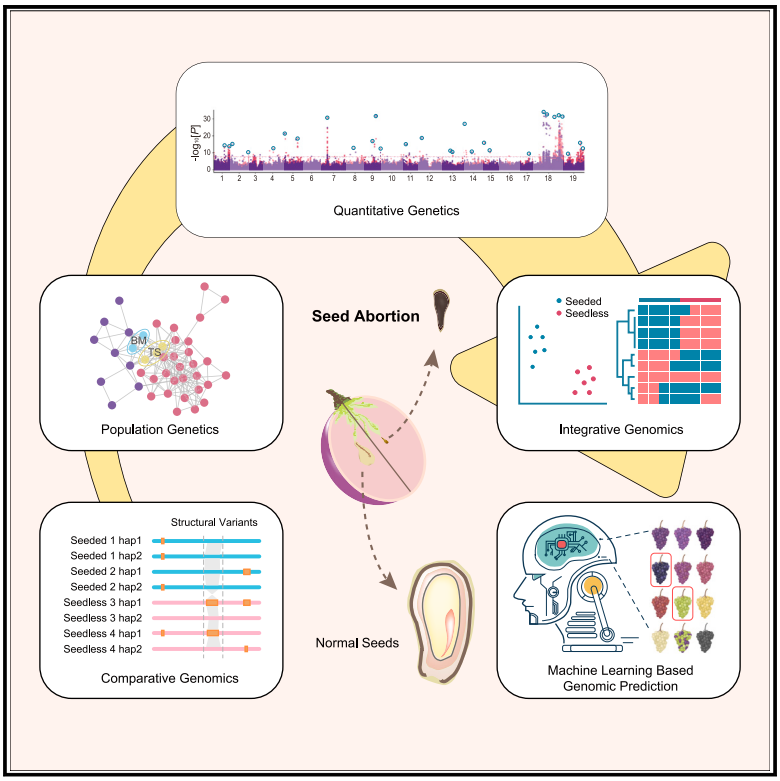
Article

Current Biology

Integrative genomics reveals the polygenic basis of seedlessness in grapevine

Graphical abstract



Highlights

- Integrative genomics investigates the polygenic basis of grapevine seedlessness
- Comparative genomics identified SVs specific to seedless grapes
- Introgression rather than convergent evolution driving evolution of seed abortion
- Machine learning-based GS predicts grape seedlessness with ~97% accuracy

Authors

Xu Wang, Zhongjie Liu, Fan Zhang, ..., Yiwen Wang, Yanling Peng, Yongfeng Zhou

Correspondence

zhouyongfeng@caas.cn

In brief

Wang et al. investigate the importance of structural variation, evolutionary origin, and polygenic basis of seedlessness in grapevine using integrative genomics. Their machine learning-based genomic prediction achieves a prediction accuracy of ~97% for seedlessness.

Wang et al., 2024, Current Biology *34*, 3763–3777
August 19, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cub.2024.07.022

CellPress

Current Biology

CellPress
OPEN ACCESS

Article

Integrative genomics reveals the polygenic basis of seedlessness in grapevine

Xu Wang,[1,2,6] Zhongjie Liu,[1,6] Fan Zhang,[1] Hua Xiao,[1] Shuo Cao,[1] Hui Xue,[1] Wenwen Liu,[1] Ying Su,[1] Zhenya Liu,[1] Haixia Zhong,[3] Fuchun Zhang,[3] Bilal Ahmad,[1] Qiming Long,[1] Yingchun Zhang,[1] Yuting Liu,[1] Yu Gan,[1] Ting Hou,[1] Zhongxin Jin,[1] Xinyu Wu,[3] Guotian Liu,[5] Yiwen Wang,[1] Yanling Peng,[1] and Yongfeng Zhou[1,4,7,8,*]

[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, 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

[5]State Key Laboratory of Crop Stress Resistance and High-Efficiency Production, College of Horticulture, Northwest A&F University, Yangling, Shaanxi 712100, China

[6]These authors contributed equally

[7]X (formerly Twitter): @yongfeng_zhou

[8]Lead contact

*Correspondence: zhouyongfeng@caas.cn

https://doi.org/10.1016/j.cub.2024.07.022

SUMMARY

Seedlessness is a crucial quality trait in table grape (*Vitis vinifera* L.) breeding. However, the development of seeds involved intricate regulations, and the polygenic basis of seed abortion remains unclear. Here, we combine comparative genomics, population genetics, quantitative genetics, and integrative genomics to unravel the evolution and polygenic basis of seedlessness in grapes. We generated the haplotype-resolved genomes for two seedless grape cultivars, "Thompson Seedless" (TS, syn. "Sultania") and "Black Monukka" (BM). Comparative genomics identified a ~4.25 Mb hemizygous inversion on Chr10 specific in seedless cultivars, with seedless-associated genes *VvTT16* and *VvSUS2* located at breakpoints. Population genomic analyses of 548 grapevine accessions revealed two distinct clusters of seedless cultivars, and the identity-by-descent (IBD) results indicated that the origin of the seedlessness trait could be traced back to "Sultania." Introgression, rather than convergent selection, shaped the evolutionary history of seedlessness in grape improvement. Genome-wide association study (GWAS) analysis identified 110 quantitative trait loci (QTLs) associated with 634 candidate genes, including previously unidentified candidate genes, such as three 11*S GLOBULIN SEED STORAGE PROTEIN* and two *CYTOCHROME P450* genes, and well-known genes like *VviAGL11*. Integrative genomic analyses resulted in 339 core candidate genes categorized into 13 functional categories related to seed development. Machine learning-based genomic selection achieved a remarkable prediction accuracy of 97% for seedlessness in grapevines. Our findings highlight the polygenic nature of seedlessness and provide candidate genes for molecular genetics and an effective prediction for seedlessness in grape genomic breeding.

INTRODUCTION

The production of seedless fruits leads to tremendous success in the global fruit market,[1] such as bananas,[2,3] citrus,[4,5] watermelons,[6] and table grapes.[7] Seed abortion in table grape has been a major focus of breeding efforts for decades, as seedless grapes are highly preferred by consumers owing to improved tastes and convenience. There are two primary methods employed to obtain seedless grapes. One involving the application of phytohormones, applying gibberellin acid (GA) and cytokinin analogs before the full bloom stage, can effectively induce seed abortion in seeded grapes.[8–10] Although this process has already become a common practice for producing seedless grapes, it raises concerns about food safety and labor costs.[11] Another one is based on genetic breeding of seedless grape cultivars. Breeders have explored diploid and triploid breeding approaches and the embryo rescue strategy to obtain new seedless varieties in recent decades.[12–14]

The development of various seed tissues in grapes involves intricate genomic regulations. Previous studies have identified

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

specific genes associated with different tissues of seed development.[15] For instance, the formation of the seed coat (integument) has been affected by genes like *VviAGL11*,[16–19] *VvMADS28*,[20] *VviINO*,[21,22] and *VvHB63*.[23–25] Nutrient storage in the endosperm is controlled by genes such as those encoding 7S and 11S globulin-like seed storage proteins,[26–28] while normal embryo growth relies on several GA related genes.[29] Moreover, the growth of ovules (young seeds) is influenced by genes such as *VvMADS39*,[30] *VvMADS45*,[31] *VviABCG20*,[32–34] *VvFUS3*,[35] *VvNAC26*,[36] *VvβVPE*,[37,38] *VviASN1*.[39] In general, multiple genes involved in tissue development of grapevine seeds. Seed abortion in grapes can occur when any of the seed tissues fail to develop properly. However, the polygenic basis of seedlessness in grapes remains unclear.

Previous studies have mapped multiple quantitative trait loci (QTLs) in different progenies in grapevine. In the linkage map of ''Dominga'' and ''Autumn Seedless,'' three QTLs for seed number (SN) and six QTLs for seed fresh weight (SFW) were detected.[40] Similarly, in the ''Muscat of Alexandria'' and ''Crimson Seedless'' progeny, six QTLs for SN and ten QTLs for SFW were detected.[41] 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 lineages.[42] However, the restricted genetic background in progenies and limited population samples hinder the investigation of the polygenic basis of seedlessness in grapes.

In this study, we first generated haplotype-resolved 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, was 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 seeded cultivars? Can we detect big SVs related to seedless/seeded cultivars?[43] Second, what evolutionary factors have 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 (GS) to enhance prediction precision in grape breeding? Overall, our work contributes to improving the understanding of the polygenic basis of seedlessness and facilitates genomic breeding of grapevine.

RESULTS

Comparative genomics between seeded and seedless grapes

To study the genetic basis of seedlessness, we generated haplotype-resolved assemblies for two seedless cultivars: TS and BM (Figures 1B and 1C), utilizing high-depth PacBio HiFi sequencing (~120× coverage) and Hi-C sequencing (~116× coverage; Figure S2C). The evaluation of K-mer heterozygosity in the TS and BM genomes, based on HiFi data, measured 1.51% and 1.41%, respectively (Data S1A), with all centromeres regions and mostly telomeres regions marked (Figure 1A; Data S1A and S1C). Statistical analysis of variants revealed that TS and BM, between their two haplotype genomes, harbor 5.35 and 5.04 Mb of SNPs, 5.01 and 4.54 Mb of insertions and deletions (InDels, < 50 bp), and 33.42 and 31.87 Mb of SVs (≥50 bp), respectively (Data S1D). 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 (Figure S3B), 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 (Data S1E). The differential chromosomes between the two haplotype genomes explain the polymorphism of alleles and the variation in the number of genes.[44,45]

To further investigate the SVs associated with seedless and seeded grapes, we aligned a total of eight varieties, including six varieties with 12 haplotype assemblies and two varieties with unphased assemblies, to the PN_T2T (Figure S3A). We detected a heterozygous inversion (PN_T2T, Chr10: 23.8–25.4 Mb) in seedless grape varieties (Figure 1D). The authenticity of these inversions was confirmed by Hi-C heatmaps and Integrative Genomics Viewer (IGV)[46] (Figure 1E; Data S2). A total of 210 genes (Chr10: 21.75–26.00 Mb) and 237 genes (Chr10:23.00-27.50 Mb) were identified in the inversion regions of TS hap1 and BM hap1, respectively (Figure 1F). Three seed development-related genes, *TRANSPARENT TESTA 16/ARABIDOPSIS BSISTER* (*TT16/ABS*) and two *SUCROSE SYNTHASE 2* (*SUS2*) genes, were discovered near the breakpoints of inversion region of the haplotype genomes (Data S1F and S1G). *TT16/ABS* controls the formation of the maternal-derived endothelial cells by interacting with *AGL11/SEEDSTICK* (*STK*) in the previous studies.[47–50] *VvTT16* was found to be present in both TS and BM haplotype genomes, while the two *VvSUS2* tandem duplication genes were hemizygous, present only in the hap1 genome of TS and BM (Data S1F and S1G). These findings suggest that the power of comparative genomics of haplotype-resolved genomes in uncovering overlooked new candidate genes underlying crucial agronomic traits.[44,45]

Introgression rather than convergent evolution underlying the evolution of seedlessness in grapevine

To explore the evolutionary history of seedlessness in grape improvement, we used WGS data from 548 grapevine accessions, including 46 seedless grapes, for population genetic analysis (Data S1H). A total of 4,462,797 SNPs, 443,812 InDels,

CellPress
OPEN ACCESS

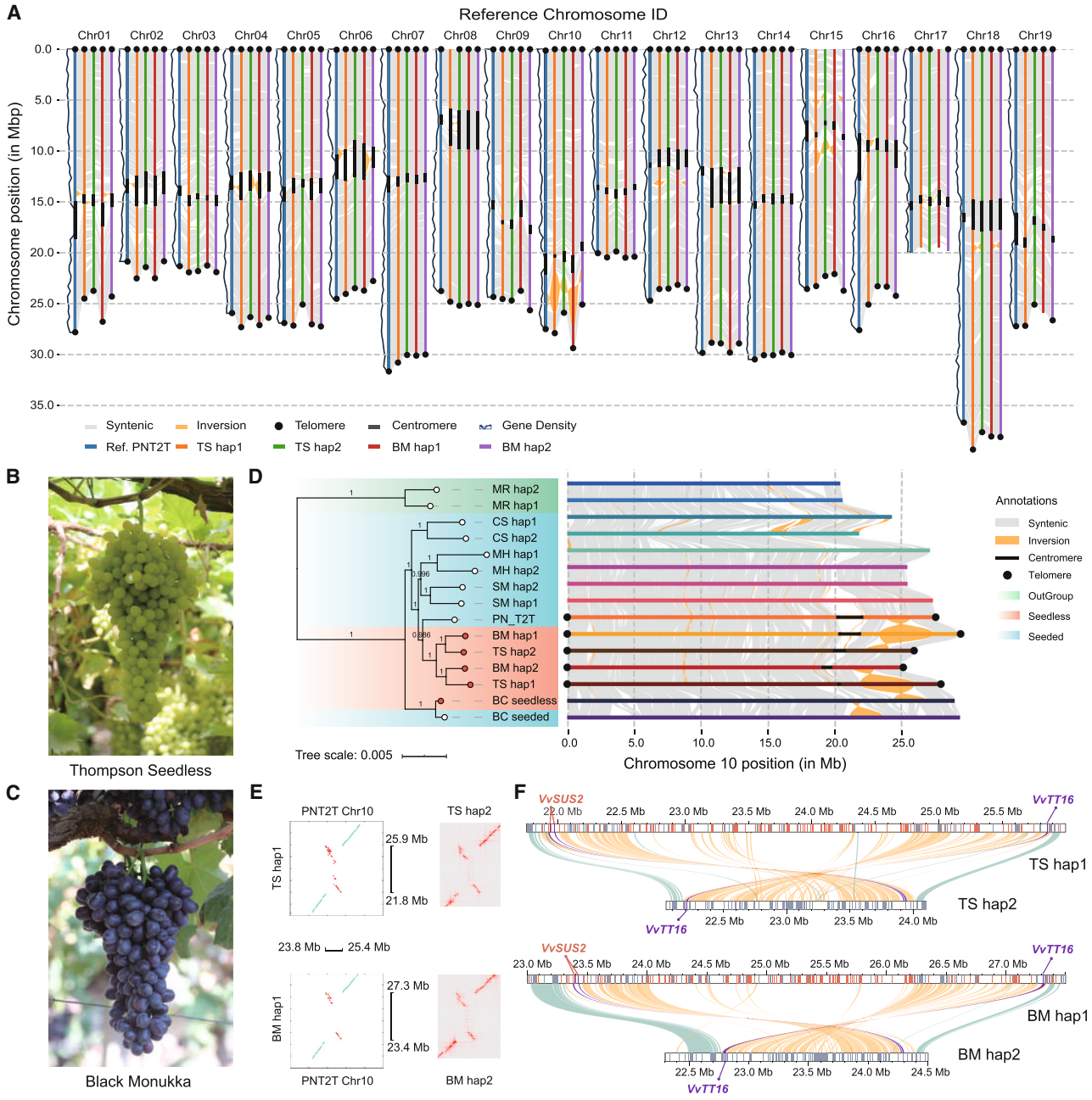


Figure 1. Comparative genomics between seeded and seedless grape cultivars

(A) Visualization of the haplotype-resolved genomes aligned with the PN_T2T, with discernible inversions marked in yellow.

(B and C) Morphological features of TS and BM.

(D) Comparative genomic analysis of 15 grape genomes prioritized by the phylogenetic tree constructed with single-copy gene features. Included genomes: MR, Muscadinia rotundifolia; CS, Cabernet Sauvignon; MH, Muscat Hamburg; SH, Shine Muscat; PN, PN40024; BC, Black Corinth.

(E) Detailed validation of a prominent Chr10 segment inversion, with its distinctive features illustrated in the Hi-C heatmap.

(F) Gene loss in hap2 genomes within the Chr10 inversion region. *VvSUS2* and *VvTT16* are located near the inversion boundaries. Red blocks represent lost genes, while gray blocks indicate genes shared between the two haplotype genomes.

See also Figures S2 and S3 and Data S1 and S2.

487,204 SVs were identified by aligning WGS data to "Cabernet Sauvignon" (CS) genome.[51] The phylogenetic tree split into six primary population branches: European wild grapes (*V. vinifera* ssp. *sylvestris* EU population [EU], *n* = 69), Middle East and Caucasus region wild grapes (*V. vinifera* ssp. *sylvestris* ME population [ME] *n* = 23), domesticated grapes (*V. vinifera* ssp. *vinifera* [VV], *n* = 352), American fox grapes (*V. labrusca* [VL], *n* = 5), hybrid of VV and VL grapes (VV × VL, *V. vinifera* × *Vitis labrusca*,

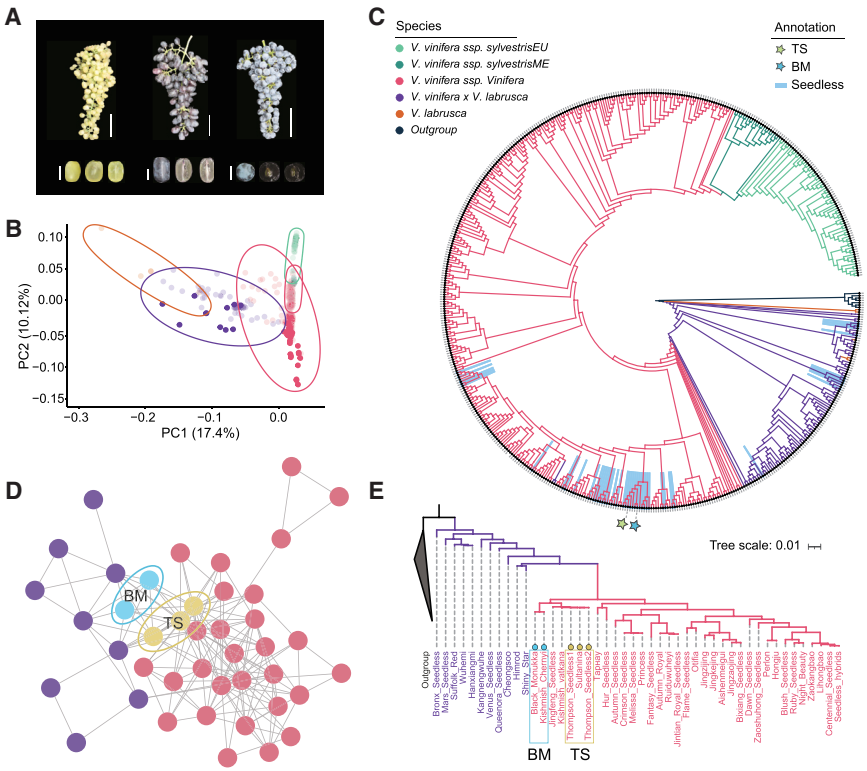


Figure 2. Evolutionary genomics of seedlessness in grapevine

(A) Comparison of grape clusters and cross-sectional views of berry from three cultivated grape varieties: TS, BM, and CS. The scale bar indicates 5 cm for grape clusters and 1 cm for individual berries.

(B) PCA on the whole-genome sequencing (WGS) data, except the outgroup, resulting in obviously separated five populations. Solid dots denote seedless grapes, while transparent dots represent seeded grapes. CS genome serves as the reference genome for variants calling.

(C) Phylogenetic tree analysis of the six populations. The light-blue blocks represent seedless grapes, and star symbols indicate TS and BM (see Figure S4 for detailed information of the phylogenetic tree).

(D) IBD analysis of 46 seedless grapes, filtering the results with a threshold of 0.40 (see Data S1I). Purple points represent the *V. vinifera* × *V. labrusca* population, while red points represent the *V. vinifera* population.

(E) The phylogenetic tree of 46 seedless grapes.

$n = 92$), and outgroup grapes (OG, $n = 7$; Figures 2C and S4). The results showed two independent lineages of seedless grapes nested in the VV × VL and VV branches, respectively, which is also supported by principal-component analysis (PCA) (Figures 2B and 2C). The seedless grapes nested in two branches, which could be driven by convergent artificial selection or introgression during grape improvement.

To distinguish convergent selection and introgression in generating seedless traits in different grapevine lineages, we preformed the introgression analyses, throughout the whole genome using f_d statistics,[52] following previous studies.[53,54] Interestingly, we detected significant genomic signals of introgression at seedless-associated locus (see the genome-wide association study section) between VV and VV × VL seedless grapes (the upper 5th percentile, $f_d = 0.266$, $p = 1.28e{-}39$), including the redefined *SEED DEVELOPMENT INHIBITOR* (SDI-new, 30.36–31.86 Mb) locus[18] and the newly detected QTL on Chr07 (SDI2, 8.85–8.86 Mb; Figures 3C and 3D). These results suggested that introgression rather than convergent artificial selection has driven the evolutionary history of seedlessness in grapes.

To validate the genetic relationship among seedless varieties, a network was constructed comprising 46 seedless grapes (35 VV and 11 VV × VL) based on the results of IBD analysis (Data S1H). The result revealed that TS group and BM group serve as bridges for gene flow between VV and VV × VL clusters (Figures 2D and 2E). In fact, BM grape traces its ancestry back to "Sultania" (TS) and "Ichkimar" (https://www.vivc.de/), with an IBD score of 0.50 supporting this observation (Data S1I). Additionally, we identified three grapes belonging to the "Sultania" somatic variants or synonym group ("Jingfeng seedless," TS1, and TS2, IBD > 0.95), seven grapes classified as parent-offspring relationship (IBD > 0.50), and 32 grapes (0.50 > IBD > 0.09) that were closely connected to "Sultania" variety (Data S1I). These findings provide evidence that "Sultania" had been extensively employed in crossbreeding with local grape varieties to enhance quality and develop new seedless grape cultivars.[7,55] Furthermore, we have ruled out the possibility of cytoplasmic male sterility (CMS) inheritance.[56,57] Phylogenetic tree and GWAS analysis of 312 grapevine accessions have proved that nuclear inheritance, rather than CMS inheritance, played a crucial role in controlling seed abortion (Figure S5). These findings corroborated IBD results that frequent introgression has facilitated the formation of seedlessness the VV and VV × VL populations (Figure 3E). As a result, the origin of the seedlessness trait could be traced back to "Sultania", and continuous introgression, rather than convergent evolution,[58] led to seed abortion in seedless grape varieties.

GWAS for seed abortion trait

To detect the QTLs and genes associated with seed abortion, we used three population for GWAS analysis, considering the effects of population structure in GWAS analyses: VV (35 seedless and 317 seeded grapes), VV × VL (11 seedless and 81 seeded grapes), and an admixed population (46 seedless and 398 seeded grapes; Data S1H). We identified a total of 110 QTLs (634 genes), including 20 QTLs (126 genes) specific to the VV population, 18 QTLs (106 genes) specific to the VV × VL population, and 72 consensus QTLs (402 genes) in admixed population, respectively (Data S1J, S1K, and S3). GO analysis revealed that genes specific to VV × VL were enriched in defense response and lignin catabolic process ($p < 0.05$), while genes specific to VV population were

CellPress
OPEN ACCESS

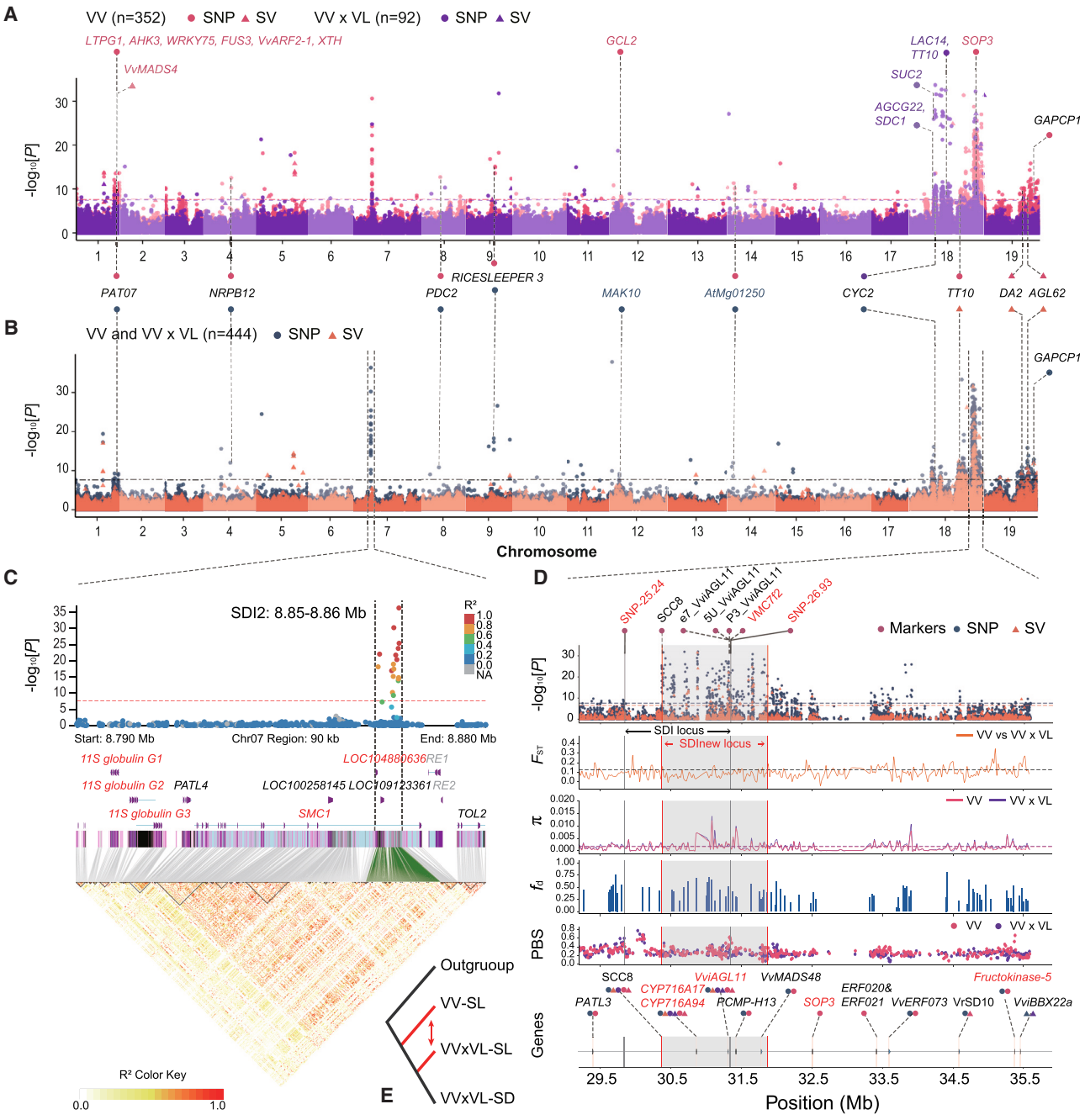


Figure 3. GWAS mapping of the polygenic basis of seedlessness

(A and B) Seedless-associated genomic loci and genes across three populations: the VV population (in red, $n = 352$), the VL population (in purple, $n = 92$), and the admixed population (in yellow and dark-blue, $n = 444$). The horizontal dashed lines denote the Bonferroni thresholds ($-\log_{10}$[0.05/variant numbers]): 7.62 for VV, 7.45 for VV × VL, and 7.66 for admixed population, respectively. Points represent SNPs, while triangles represent SVs (and InDels).

(C) LD correlation analysis of a highly linked SDI2 locus in Chr07, and genes associated with seed abortion are highlighted in red.

(D) Admixed population analysis of a highly linked region in Chr18. The SDI locus (Chr18: 29.83–31.34 Mb) is identified based on SNP markers, and the redefined SDI locus (SDInew, Chr18: 30.36–31.86 Mb) is defined based on GWAS results, depicted in the gray block. The significant threshold: 7.61 for SNPs and 6.66 for SVs (and InDels). The dashed line for fixation indices (F_{ST}) indicates an average value of 0.126, with genetic diversity (π) measured 0.0019 for VV × VL and 0.0016 for VV. The top 5th percentile of f_d statistics is 0.277, with PBS statistics measuring 0.356 in VV and 0.415 in VV × VL.

(E) Gene flow pattern based on ABBA-BABA statistics, where SD represents seeded, SL represents seedless.

See also Figures S6 and S7 and Data S1 and S3.

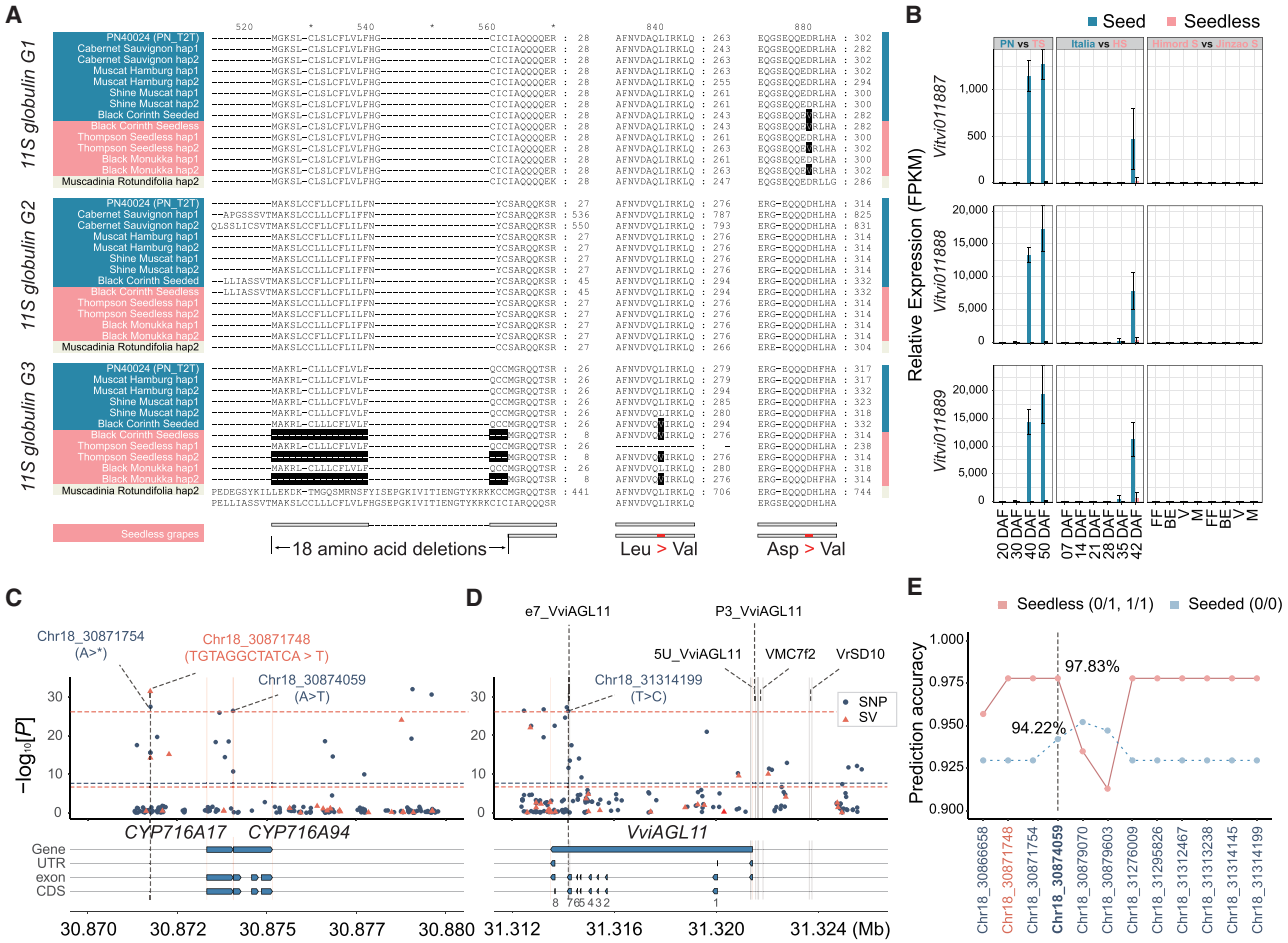


Figure 4. Deep mining of key loci associated with seed abortion

(A) Sequence alignment of three 11S GLOBULIN SEED STORAGE PROTEIN genes in Chr07 across 14 grape assemblies. Black blocks represent deletion and nonsynonymous mutations.

(B) Relative expression values of these three genes at different time points across six grape varieties. Error bars represent the standard deviation of three biological replicates at each time point. Abbreviations: DAF, Day after flowering; FF, Full flowering; BE, Berry expansion; V, Veraison; M, Maturity.

(C and D) Visualization of crucial candidate genes associated with seed abortion within the SDInew locus region. GWAS significance thresholds ($-\log_{10}$[0.05/variant numbers]): 7.61 for SNPs and 6.66 for SVs (and InDels). The top ten percentiles of significant variants are Chr18_ 31295826 (y = 26.39).

(E) Genotyping percentage of highly significant variants is within the region Chr18: 30.70–31.32 Mb. "Seedless" includes cases with genotypes 0/1 and 1/1, and "seeded" includes the case with genotype 0/0.

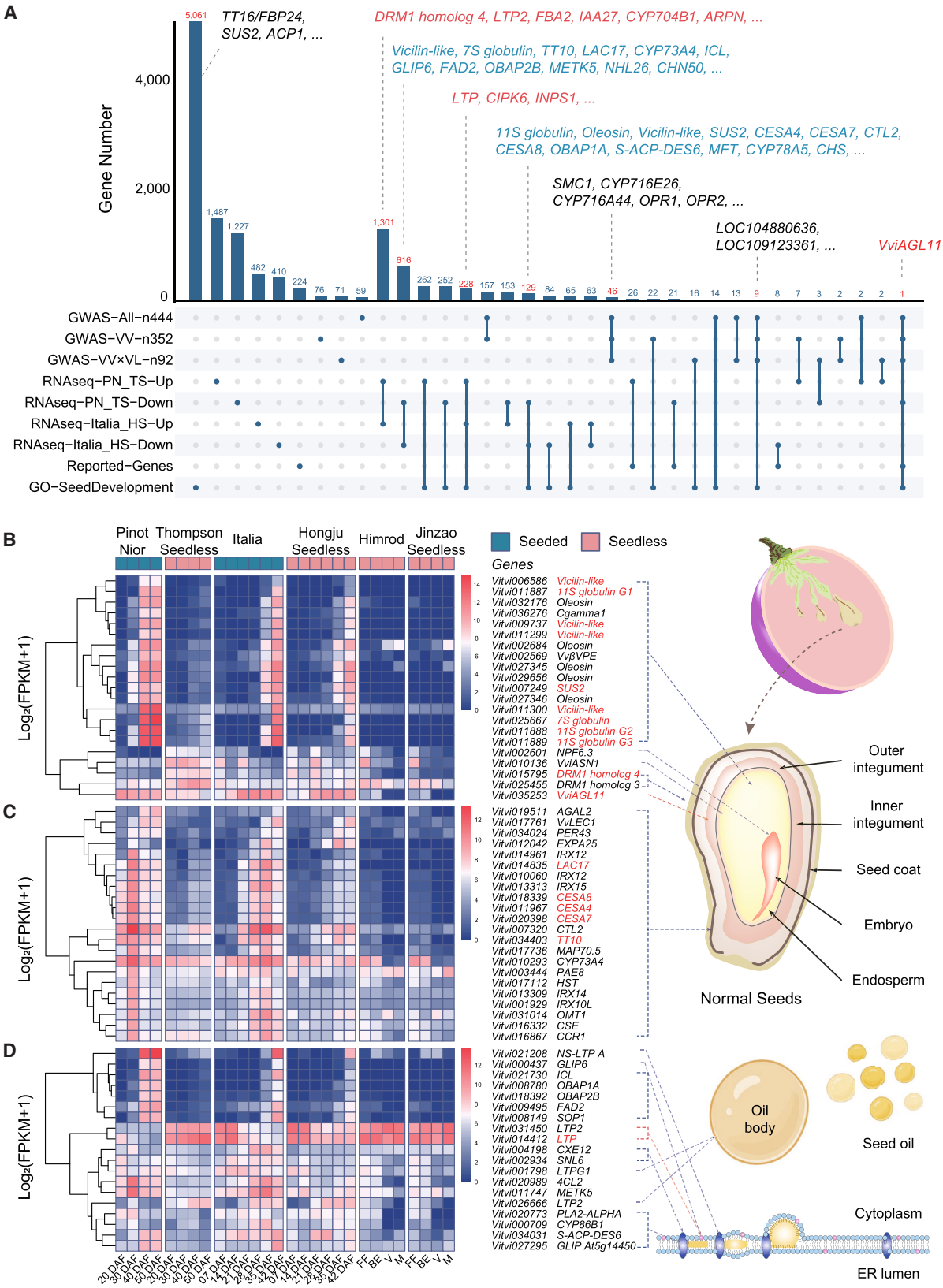
See also Data S1 and S4.

enriched in embryo development ending in seed dormancy and xylan metabolic process ($p < 0.05$; Figure S6; Data S1L).

Remarkably, two consensus regions exhibited high consistence in three populations: Chr07: 8.85–8.86 Mb and Chr18: 29.40–35.54 Mb (Figures 3A and 3B). In Chr07 locus (SDI2), two genes, *REVERSE TRANSCRIPTASE ZINC-BINDING DOMAIN-CONTAINING PROTEIN* (*LOC104880636*, *Vitvi011893*) and *STRUCTURAL MAINTENANCE OF CHROMOSOMES PROTEIN* (*SMC1*, *Vitvi011891*), were positioned within a tightly linked region with high linkage disequilibrium (LD) values (Figure 3C). The *LOC104880636* gene is annotated as the regulation of seed growth on UniProt, and the *smc1* mutants produced arrested early embryo development and blocked cellularization of the endosperm[59] (Data S1K). 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*[26–28,60] (Figure 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 assemblies (Figure 4A; Data S4). Among them, both the heterozygous Asp-to-Val and Leu-to-Val mutations were specific in seedless grapes, as well as the somatic mutations of "Black Corinth Seeded" grapes. However, the heterozygous deletion of the 18 amino acids (signal peptide) in *11S globulin G3* was only detected in seedless grapes (Figure 4A). The relative expression values of these genes showed a significant correlation with seed phenotypes, especially from 40 to 50 days after flowering (DAF; Figure 4B).

Another consensus region is located on the Chr18: 29.40–35.54 Mb, encompassing 17 QTLs and the reported SDI locus

CellPress
OPEN ACCESS



(legend on next page)

(Chr18: 29.83–31.34 Mb; Figure 3D). Due to the narrow genetic background of the samples used in the previous study,[18] we redefined the SDI locus (SDInew, Chr18: 30.36–31.86 Mb) through GWAS analysis, revealing a total of eight QTLs (Figure 3D; Data S3). 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; Data S1M). This finding was also supported by genetic diversity (π) statistics. The two seedless populations showed similar genetic diversity on the SDInew locus (Figure 3D). The f_d statistics 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. (Figure 3D). Notably, several SNPs and InDels were highly associated with these candidate genes, especially in the promoter and coding sequence region (Figures 4C, 4D, and S7), while several published molecular markers for seedlessness prediction, including e7_VviAGL11,[41] 5U_VviAGL11,[41] P3_VviAGL11,[16] and VMC7f2,[61] showed low predictive accuracy due to the absence of high quality variations (Figure 4D; Data S1N). We selected the top 10% of associated sites based on $-\log_{10}[p]$ values within the SDInew locus (Figure 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 (Data S3). 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 *VvMADS4*[62] and *VvARF2-1*[63] (Figure 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*,[64] *SUS2*, *LACCASE-14* (*LAC14*), and *TT10/LAC15* were highly associated with seed development (Figure 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 candidate genes to identify the core candidate genes associated with seed abortion (see STAR Methods). Among these, three transcriptomic groups, including 76 samples and 14 time points, were employed to detect differentially expressed genes (DEGs) between seeded and seedless grapes at each development stage (Data S1O). For the "Italia" and "Hongju Seedless" (HS) groups,[65] we identified a total of 2,680 significantly upregulated genes and 1,835 significantly downregulated genes from the six time points (Figure S9C). Similarly, in the "Pinot Noir" (PN) and TS groups,[66] we identified a total of 3,969 significantly upregulated genes and 2,695 significantly downregulated genes (Figure S9C). "Himrod Seedless" (Himrod) and "Jinzao Wuhe" (Jinzao) were used serve as a control for cross-validation during four fruit development stage.[67] Interestingly, we found *VviAGL11* was exclusively in the downregulated DEGs in the PN and TS comparison, but not in the Italia and HS comparison (Figures S8D and S9B). In addition, 1,301 upregulated genes and 616 downregulated genes were only identified in transcriptomic analyses using integrative genomic analysis, such as *DORMANCY-ASSOCIATED PROTEIN HOMOLOG 4* (*DRM1 homolog 4*), *NON-SPECIFIC LIPID-TRANSFER PROTEIN 2* (*LTP2*), *VICILIN-LIKE SEED STORAGE PROTEIN*, *7S SEED STORAGE PROTEIN* (*7S GLOBULIN*), *TT10*, *LAC17*, etc. (Figure 5A; Data S1K).

To include more important genes involved in seed development, we performed a protein sequence similarity alignment for all genes associated with 34 GO terms related to seed development against the PN_T2T reference genome (Figures S8E and S8F). This yielded 6,529 homologous genes (see STAR Methods), with 5,061 genes only present in the GO pathway, including the *TT16/FBP24* gene identified in the comparative genomics (Figure 5A). Notably, the intersection between the RNA sequencing (RNA-seq) related genes and the GO homologous genes revealed 163 downregulated genes and 294 upregulated genes (Figure S9B), such as *LTP*, *11S GLOBULIN*, *OLEOSIN*, *CELLULOSE SYNTHASE A CATALYTIC SUBUNIT* (*CESA4*, *CESA7*, and *CESA8*), etc. (Data S1K). The consensus GWAS genes and GO homologous genes exhibited nine candidate genes, such as previously mentioned *LOC104880636* (Figures 3C and 5A). Furthermore, we integrated a total of 451 family genes and seven molecular markers related to seed abortion from previous studies (Data S1P). All these elements were aligned against the PN_T2T reference genome (e < 0.1) and exhibited overlap with candidate

Figure 5. Integrative genomic analyses for grapevine seed abortion

(A) Results from integrative genomic analyses: GWAS, transcriptomic analysis, reported genes and markers mapping, and GO homologous genes overlapping. Blue bars represent the number of genes uniquely identified and overlapping through this approach.

(B and C) Heatmap of $\log_2$(FPKM + 1) for candidate genes related to embryo, endosperm, and seed coat development, resulting from integrative genomic analyses. FPKM (fragments per kilobase of transcript per million mapped reads) is a normalized measure of gene expression that accounts for sequencing depth and gene length. The genes are indicated in relation to their expression in specific tissues and time points on schematic representation transverse profiles of seeds.

(D) Expression patterns of genes involved in lipid synthesis, degradation, and transportation, which also pinpointed on the schematic representation of oil body formation. Abbreviations: DAF, Day after flowering; FF, Full flowering; BE, Berry expansion; V, Veraison; M, Maturity.

See also Figures S8 and S9 and Data S1.

genes in other analyses, including previously reported genes such as *VvβVPE*,[37] *HD-ZIP PROTEINS ATHB-1/HAT5*, *ATHB-12/VvHB56* and *ATHB40/VvHB18*,[23] *VviASN1*,[39] *VvMJE1*,[68] *VvLEC1*,[69] *VvMADS2/VvSEP1*,[70] etc. (Figure S9D; Data S1K).

Through integrative genomic analysis, we screened 339 core candidate genes, categorized into 13 groups based on their functions, by condition-based filtering that exhibited significant differential expression between seedless and seeded grape cultivars (see STAR Methods and Data S1K). Among them, 77 genes were directly associated with seed development-related GO homologous genes, and three groups deserve our attention: Firstly, the differential expression of candidate genes in the endosperm development impacts nutrient storage in seedless grapes (Figure 5B). Nutrient deficiency could be primary factor leading to seed abortion in later embryo development[71]; secondly, genes involved in the regulation of lignin and cellulose synthesis/degradation in seed coat exhibit higher activity levels in seeded grapes (Figure 5B); and thirdly, candidate genes manage the synthesis and transport of oil bodies, ensuring efficient lipid accumulation and utilization during seed development (Figure 5D). Overall, our findings indicate that multiple genes have an accumulative effect in the process of seed abortion, resulting varying degrees of seedlessness.[65] This complexity highlights the challenge of accurately distinguish seed abortion using a single gene or variant site.

Machine learning-based genomic selection for seedless grape breeding

Given the polygenic nature of the seedlessness trait, genomic prediction could greatly improve the speed and accuracy for seedless grape breeding. We extracted the information of all 794 high-quality variant sites that exceeded significant threshold (7.66) from the GWAS analysis, including 77 InDels and 717 SNPs (Data S1R). Using these variations, an unrooted phylogenetic tree was constructed based on admixed populations (n = 444), revealing that the majority of seedless individuals clustered together (Figure 6A). However, some seeded samples, like seeded hybrid progeny[18] and Rizamat[41] were mixed in seedless grapes, as well as seedless samples, such as "Dawn Seedless," "Bronx Seedless," "Cheongsoo," "Ruby Seedless," and "Jingkejing," were mixed in seeded grapes. Interestingly, the mutations in the top two QTLs were heterozygous: Chr07: 8.85–8.86 Mb (SDI2 locus) and Chr18: 30.36–31.86 Mb (SDInew locus; Figure 6D). These results suggest the complexity of seedless in grapes.

Therefore, to address this problem, we employed genomic selection based on machine learning to enhance predictive accuracy (Figure S10A). We utilized 794 variant sites along with phenotypic data from 444 admixed samples as the training set and assessed the predictive accuracy of nine classical models for the seedlessness trait, using 100 repetitions of random five-fold cross-validations (Figure 6B). To enhance GS efficiency, we filtered and reduced the dimensions of these features based on LD scores, removing one of two variants whenever the LD score exceeded 0.8, 0.5, or 0.2. This process yielded 137, 54, and 45 variant sites, respectively, for which we evaluated model prediction accuracy (Figure 6B; Data S1S). Among these models, machine learning-based methods, including three regression models—ElasticNetCV, PLSRegression, and SVR_poly—and the binary classification model LogisticRegression, achieved prediction accuracies of approximately 97%, standing out from the rest. Consequently, we applied these four models to predict phenotypic values on the testing set data (39 samples not included in the training set) (Data S1H). Even with fewer variant sites, the LogisticRegression and ElasticNetCV models yielded 100% accuracy and exhibited robust stability and high resolution, with correlation coefficients R values of 0.96 ($p < 2.2e{-}16$) and 1 ($p < 2.2e{-}16$), respectively (Figures 6C and S10B). Our findings demonstrated the effectiveness of machine learning-based genomic selection for seedlessness in grapes.

DISCUSSION

Seedlessness is an important quality trait in table grape breeding. Previous genetic studies the functions of multiple genes,[41,56] including *VvAGL11*, however, quantitative genetic analyses of natural population about seedlessness were not conducted in grapes. In this study, we conducted integrative genomic analyses to investigate the polygenic basis of seedlessness in grapes: (1) comparative analysis of 15 genomes allowed us to discover a heterozygous inversion in Chr10 associated with the seedless trait; (2) the evolutionary genomic analyses showed that seedless grapes were closely related with an origin from the well-known seedless grape Sultania (TS). Introgression rather than convergent evolution was associated with the evolution of seedlessness in grapes; (3) a total of 110 QTLs associated with 634 candidate genes were identified through GWAS analysis within a large natural population, including four significant linkage regions such as Chr01: 18.49–19.96 Mb (specific to VV population), Chr07: 8.85–8.86 Mb (shared, SDI2 locus), Chr18: 11.7–20.0 Mb (specific to VV × VL population), and Chr18: 23.9–35.5 Mb (shared, SDInew locus); (4) a total of 339 core genes associated with seedlessness was detected through integrative genomics analyses; (5) machine learning-based genome selection was built to accurately predict the seedless phenotypes. Importantly, these findings could efficiently save the cost and time in table grape breeding.

The polygenic nature of seedlessness in grapes

The occurrence of seedlessness phenotypes from a cumulative polygenic effect associated with different tissues and development stages.[15] The limited observations employing single methods, such as transcriptomic analysis, are insufficient. As a result, numerous important candidate genes were ignored by previous studies. For example, *TT16/ABS*, found at the inversion boundary through comparative genomics (Figure 1D), is one of promising candidate genes. Due to the redundant function between *STK* and *SHATTERPROOF* (*SHP1* and *SHP2*), the double *abs stk* mutants and triple *tt16 shp1 shp2* mutants all induced fewer seeds and exhibited defects in seed coat formation.[15,49] Candidate genes mapped by GWAS, such as *SMC1*, *11S globulin G1/2/3*, *LTG1*, *SUS2*, *LAC14*, *TT10/LAC15*, *MANNAN ENDO-1,4-BETA-MANNOSIDASE 5* (*MAN5*), *PROBABLE FRUCTOKINASE-5*, *E3 UBIQUITIN-PROTEIN LIGASE* (*DA2*), *AGL62*, etc., and well-known gene *VviAGL11*, play crucial roles in pollen, endosperm and seed coat development (Figures 3A and 3B; Data S1K). Additionally, three transcriptomic analyses, GO homologous genes, and previously reported

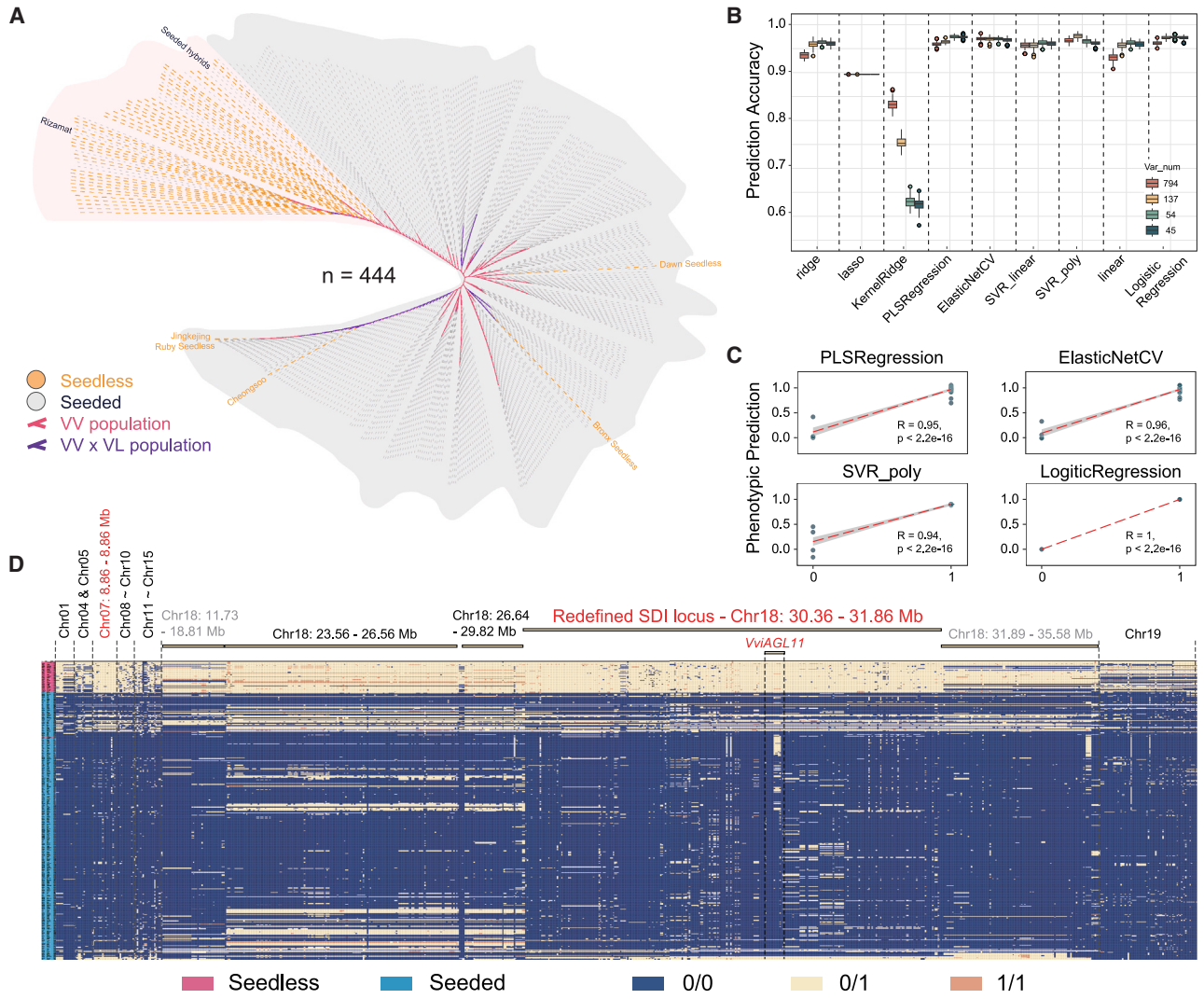


Figure 6. Machine learning-based genomic selection on seedlessness in grapevine breeding

(A) Phylogenetic clustering based on 794 significant variants (77 InDels and 717 SNPs) derived from the GWAS analysis.

(B) Comparison of seedlessness prediction accuracy among nine classical models for genome selection. Dimensionality reduction was achieved by removing variants with strong linkage disequilibrium (LD). The 794 represents all variant sites without LD score filtering, while 137 variant sites were filtered with LD < 0.8, 54 with LD < 0.5, and 45 with LD < 0.2. The training dataset consists of genotyping data and phenotypes from 444 grape samples.

(C) Phenotypic prediction results of the four best-performing models based on 45 variant sites (LD < 0.2). The testing set encompasses 39 samples, using models built from the 444-sample training set. Phenotypic values: 0 for seedless and 1 for seeded grapes. The R and *p* values were calculated with the Pearson correlation method. Prediction results for models using 794, 137, and 54 sites are presented in Figure S10.

(D) Genotyping visualization of the 794 variants, sorted based on prediction results of the SVR-ploy model.

See also Data S1.

family genes provide us with numerous significant candidate genes (Figure 5A; Data S1K).

Importantly, to define the interconnections among multiple genomic analyses, integrative genomic analysis was applied in this case, screened out 339 core genes with high significance from thousands of candidate genes (Figure 5A; Data S1K). Most of these genes are associated with three main tissues related to seed coat, endosperm, and embryo, as shown in Figures 5B–5D, and ten other development processes (Data S1K). For example, we identified numerous candidate genes involving in seed hormone regulation, such as *MOTHER OF FT AND TFL1* (*MFT*) in abscisic acid (ABA) and GA pathways, *VvABI3-1*[72] in ABA pathway, *VvGH3.9*[73] in auxin pathways, and *VvMJE1*[68] in jasmonate pathway. Additionally, genes controlling the development of floral organs, especially pollen and stigma, significantly influence the subsequent ovule development. This includes candidate genes like *CYP78A5*, *VvMADS27*, and metal ion transport genes *METALLOTHIONEIN-LIKE PROTEINS MT1* and *MT3*.[74] Except for lipid accumulation (Figure 5D), the processes of sugar and amino acid synthesis also contribute to seed development, such as *BGLU15*, *VviGAPDH*,[75] glycine-rich protein (*Vitvi021557*, *Vitvi036421*), and 36.4 kDa

proline-rich protein (*Vitvi002605*). Overall, these findings not only emphasize the polygenic nature of seedlessness but also provide novel candidate genes for functional genetics, highlighting the complexity of seed abortion regulation.

The implications of genomic breeding of seedless table grapes

Both individual markers for marker-assisted selection (MAS)[41,76–78] and marker set for genomic selection generated in this study could be efficiently used in table grape breeding (Figure 6D). Interestingly, in the variation map around *VviAGL11*, we observed that all the previously developed molecular markers based on lineages with a narrow genetic background for marker-assisted in table grape breeding were filtered away, due to either the low genotyping quality or a high missing rate, such as SCF27,[79] e7_VviAGL11,[41] 5U_VviAGL11,[41] P3_VviAGL11,[16] VMC7f2,[61] and VrSD10[80] (Figure 4D). Luckily, we designed a set of 12 markers, based on our GWAS analyses of natural population with species-wide genetic diversity. These markers could accurately delimit seeded and seedless grapes, achieving a precision rate >90% in nature populations (Figure 4E).

Quantitative genetics analysis revealed 110 high-quality QTLs associated with seedlessness in grapevines, including 634 candidate genes (Data S1K). By extracting GWAS significant variants from the admixed population, we applied all 794 significant variant sites for GS. To enhance prediction efficiency and balance GS costs, we reduced dimensions by filtering based on LD scores to retain high-quality sites. Our results demonstrate that GS, using machine learning algorithms such as LogisticRegression and ElasticNetCV, achieved an impressive predicted accuracy of 97% within a few seconds using only 45 variants (LD < 0.2). This approach could facilitate early genomic selection of natural germplasms and hybrid progeny. Many crops, such as tomato,[81] potato,[82,83] cereal,[84] rice,[85,86] wheat, and maize,[87,88] have successfully applied genomic selection and prediction in their breeding programs. However, applying GS to grape breeding was a pioneering effort. In the future, numerous agronomic traits such as fruit aroma, disease resistance, pest resistance,[89] and soluble solids content could be integrated into a single GS chip, offering a powerful genomic tool for genomic design of grapevine breeding.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Grape genome sequencing
  - Genome assembly
  - Genome annotation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Assembly assessment
  - Comparative genomics
  - WGS variation detection
  - Population genetics analysis
  - Genome-wide association study
  - Multi-transcriptome analysis
  - Integrative genomic analysis
  - Genome selection based on machine learning

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.07.022.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (no. 2023YFF1000100; 2023YFD2200700), the Science Fund Program for Distinguished Young Scholars of the National Natural Science Foundation of China (Overseas) to Yongfeng Zhou, the National Natural Science Foundation of China (no. 32372662), the Project of Fund for Stable Support to Agricultural Sci-Tech Renovation (xjnkywdzc-2022009), Xinjiang Uygur Autonomous Region Tianchi Talent – Special Expert Project (whole genome design and breeding of grapes), and the central government guides local science and technology development special fund projects (2022)-Germplasm Innovation and Breeding Ability Improvement of Characteristic Fruit Trees. We thank members of the Zhou lab at Agricultural Genomics Institute at Shenzhen (AGIS) for discussions and comments on the project. We thank Lin Cheng, Hongbo Li, Zhigui Bao, and Qing Zhang at AGIS for giving guidance and scripts to haplotype genome assembly. We express our gratitude to Zhiwu Zhang of Washington State University, USA, for providing invaluable suggestions regarding genome selection and prediction, and to Pádraig Cunningham of University College Dublin, Ireland, for his exceptional module on machine learning theory.

AUTHOR CONTRIBUTIONS

Yongfeng Zhou conceived and designed the study. H.Z., Fuchun Zhang, and Xinyu Wu collected the plant materials. H.Z. and Zhongjie Liu performed experiments and genome sequencing. Xu Wang, S.C., Y.S., T.H., and Yingchun Zhang assembled and annotated the haplotype-resolved genomes. Xu Wang, Zhongjie Liu, Fan Zhang, Hui Xue, W.L., Zhenya Liu, and Y.G. performed data analysis. Yongfeng Zhou, Xu Wang, Zhongjie Liu, Fan Zhang, and Hua Xiao wrote the original draft of the manuscript. All authors provided critical feedback and revised the manuscript.

DECLARATION OF INTERESTS

The authors have a patent application related to this work. The patent application number is CN202311654558.7. This declaration is made to ensure full transparency regarding potential conflicts of interest.

Received: March 3, 2024
Revised: May 4, 2024
Accepted: July 3, 2024
Published: August 1, 2024

REFERENCES

1. Varoquaux, F., Blanvillain, R., Delseny, M., and Gallois, P. (2000). Less is better: new approaches for seedless fruit production. Trends Biotechnol. *18*, 233–242.
2. Sardos, J., Rouard, M., Hueber, Y., Cenci, A., Hyma, K.E., Van Den Houwe, I., Hribova, E., Courtois, B., and Roux, N. (2016). A genome-wide association study on the seedless phenotype in banana (Musa spp.) reveals the potential of a selected panel to detect candidate genes in a vegetatively propagated crop. PLoS One *11*, e0154448.
3. Sidhu, J.S., and Zafar, T.A. (2018). Bioactive compounds in banana fruits and their health benefits. Food Qual. Saf. *2*, 183–188.

4. Ye, W., Qin, Y., Ye, Z., da Silva, J.A.T., Zhang, L., Wu, X., Lin, S., and Hu, G. (2009). Seedless mechanism of a new mandarin cultivar 'Wuzishatangju'(Citrus reticulata Blanco). Plant Sci. *177*, 19–27.
5. Zhang, S., Shi, Q., Albrecht, U., Shatters, R.G., Stange, R., McCollum, G., Zhang, S., Fan, C., and Stover, E. (2017). Comparative transcriptome analysis during early fruit development between three seedy citrus genotypes and their seedless mutants. Hortic. Res. *4*, 17041.
6. Wijesinghe, S.A.E.C., Evans, L.J., Kirkland, L., and Rader, R. (2020). A global review of watermelon pollination biology and ecology: the increasing importance of seedless cultivars. Sci. Hortic. *271*, 109493.
7. Akkurt, M., Tahmaz, H., and Veziroğlu, S. (2019). Recent developments in seedless grapevine breeding. S. Afr. J. Enol. Vitic. *40*, 1.
8. Reynolds, A.G., Wardle, D.A., Zurowski, C., and Looney, N.E. (1992). Phenylureas CPPU and thidiazuron affect yield components, fruit composition, and storage potential of four seedless grape selections. J. Am. Soc. Hortic. Sci. *117*, 85–89.
9. Cheng, C., Xu, X., Singer, S.D., Li, J., Zhang, H., Gao, M., Wang, L., Song, J., and Wang, X. (2013). Effect of GA3 treatment on seed development and seed-related gene expression in grape. PLoS One *8*, e80044.
10. Tyagi, K., Maoz, I., Kochanek, B., Sela, N., Lerno, L., Ebeler, S.E., and Lichter, A. (2021). Cytokinin but not gibberellin application had major impact on the phenylpropanoid pathway in grape. Hortic. Res. *8*, 51.
11. Park, Y.-S., Lee, J.-C., Jeong, H.-N., Um, N.-Y., and Heo, J.-Y. (2022). A red triploid seedless grape 'Red Dream'. Hortscience *57*, 741–742.
12. Park, S.M., Hiramatsu, M., and Wakana, A. (1999). Aneuploid plants derived from crosses with triploid grapes through immature seed culture and subsequent embryo culture. Plant Cell Tissue Organ Cult. *59*, 125–133.
13. Ji, W., Li, Z.Q., Zhou, Q., Yao, W.K., and Wang, Y.J. (2013). Breeding new seedless grape by means of in vitro embryo rescue. Genet. Mol. Res. *12*, 859–869.
14. Li, J., Wang, X., Wang, X., and Wang, Y. (2015). Embryo rescue technique and its applications for seedless breeding in grape. Plant Cell Tissue Organ Cult. PCTOC *120*, 861–880.
15. Ehlers, K., Bhide, A.S., Tekleyohans, D.G., Wittkop, B., Snowdon, R.J., and Becker, A. (2016). The MADS box genes ABS, SHP1, and SHP2 are essential for the coordination of cell divisions in ovule and seed coat development and for endosperm formation in Arabidopsis thaliana. PLoS One *11*, e0165075.
16. Mejia, N., Soto, B., Guerrero, M., Casanueva, X., Houel, C., de los Angeles Miccono, M., Ramos, R., Le Cunff, L., Boursiquot, J.-M., and Hinrichsen, P. (2011). Molecular, genetic and transcriptional evidence for a role of VvAGL11 in stenospermocarpic seedlessness in grapevine. BMC Plant Biol. *11*, 1–19.
17. Malabarba, J., Buffon, V., Mariath, J.E.A., Gaeta, M.L., Dornelas, M.C., Margis-Pinheiro, M., Pasquali, G., and Revers, L.F. (2017). The MADS-box gene Agamous-like 11 is essential for seed morphogenesis in grapevine. J. Exp. Bot. *68*, 1493–1506.
18. Royo, C., Torres-Pérez, R., Mauri, N., Diestro, N., Cabezas, J.A., Marchal, C., Lacombe, T., Ibáñez, J., Tornel, M., Carreño, J., et al. (2018). The major origin of seedless grapes is associated with a missense mutation in the MADS-box gene VviAGL11. Plant Physiol. *177*, 1234–1253.
19. Amato, A., Cardone, M.F., Ocarez, N., Alagna, F., Ruperti, B., Fattorini, C., Velasco, R., Mejía, N., Zenoni, S., and Bergamini, C. (2022). VviAGL11 self-regulates and targets hormone-and secondary metabolism-related genes during seed development. Hortic. Res. *9*, uhac133.
20. Zhang, S., Wang, L., Yao, J., Wu, N., Ahmad, B., Nocker, S., Wu, J., Abudureheman, R., Li, Z., and Wang, X. (2023). Control of ovule development in Vitis vinifera by VvMADS28 and interacting genes. Hortic. Res. *10*, uhad070.
21. Lora, J., Hormaza, J.I., Herrero, M., and Gasser, C.S. (2011). Seedless fruits and the disruption of a conserved genetic pathway in angiosperm ovule development. Proc. Natl. Acad. Sci. USA *108*, 5461–5465.
22. di Rienzo, V., Imanifard, Z., Mascio, I., Gasser, C.S., Skinner, D.J., Pierri, C.L., Marini, M., Fanelli, V., Sabetta, W., Montemurro, C., et al. (2021). Functional conservation of the grapevine candidate gene INNER NO OUTER for ovule development and seed formation. Hortic. Res. *8*, 29.
23. Li, Y., Zhu, Y., Yao, J., Zhang, S., Wang, L., Guo, C., Van Nocker, S., and Wang, X. (2017). Genome-wide identification and expression analyses of the homeobox transcription factor family during ovule development in seedless and seeded grapes. Sci. Rep. *7*, 12638.
24. Li, Y., Zhang, S., Dong, R., Wang, L., Yao, J., van Nocker, S., and Wang, X. (2019). The grapevine homeobox gene VvHB58 influences seed and fruit development through multiple hormonal signaling pathways. BMC Plant Biol. *19*, 523.
25. Yao, J., Zhang, S., Wu, N., Li, X., Ahmad, B., Wu, J., Guo, R., and Wang, X. (2023). KNOX transcription factor VvHB63 affects grape seed development by interacting with protein VvHB06. Plant Sci. *330*, 111665.
26. Gazzola, D., Vincenzi, S., Gastaldon, L., Tolin, S., Pasini, G., and Curioni, A. (2014). The proteins of the grape (Vitis vinifera L.) seed endosperm: fractionation and identification of the major components. Food Chem. *155*, 132–139.
27. Chamizo-González, F., Heredia, F.J., Rodríguez-Pulido, F.J., González-Miret, M.L., and Gordillo, B. (2022). Proteomic and computational characterisation of 11S globulins from grape seed flour by-product and its interaction with malvidin 3-glucoside by molecular docking. Food Chem. *386*, 132842.
28. Chamizo-González, F., Estévez, I.G., Gordillo, B., Manjón, E., Escribano-Bailón, M.T., Heredia, F.J., and González-Miret, M.L. (2023). First insights into the binding mechanism and colour effect of the interaction of grape seed 11S globulin with malvidin 3-O-glucoside by fluorescence spectroscopy, differential colorimetry and molecular modelling. Food Chem. *413*, 135591.
29. He, H., Liang, G., Lu, S., Wang, P., Liu, T., Ma, Z., Zuo, C., Sun, X., Chen, B., and Mao, J. (2019). Genome-wide identification and expression analysis of GA2ox, GA3ox, and GA20ox are related to gibberellin oxidase genes in grape (Vitis vinifera L.). Genes *10*, 680.
30. Zhang, S., Yao, J., Wang, L., Wu, N., van Nocker, S., Li, Z., Gao, M., and Wang, X. (2022). Role of grapevine SEPALLATA-related MADS-box gene VvMADS39 in flower and ovule development. Plant J. *111*, 1565–1579.
31. Sun, X., Zhang, S., Li, X., Zhang, X., Wang, X., Wang, L., Li, Z., and Wang, X. (2020). A MADS-box transcription factor from grapevine, VvMADS45, influences seed development. Plant Cell Tissue Organ Cult. PCTOC *141*, 105–118.
32. Tang, Y., Liu, B., Li, Y., Van Nocker, S., Wang, Y., and Zhang, C. (2020). Differential expression of the seed-specific gene ABCG20 between seedless and seeded grapes and its roles in tomato seed development. S. Afr. J. Bot. *131*, 428–436.
33. Wang, L., Dai, W., Shi, Y., Wang, Y., and Zhang, C. (2022). Cloning and activity analysis of the highly expressed gene VviABCG20 promoter in seed and its activity is negatively regulated by the transcription factor VviDof14. Plant Sci. *315*, 111152.
34. Wang, L., Yao, W., Zhang, X., Tang, Y., Van Nocker, S., Wang, Y., and Zhang, C. (2023). The putative ABCG transporter VviABCG20 from grapevine (Vitis vinifera) is strongly expressed in the seed coat of developing seeds and may participate in suberin biosynthesis. Physiol. Mol. Biol. Plants *29*, 23–34.
35. Ahmad, B., Zhang, S., Yao, J., Chai, S., Yadav, V., Athar, H.R., Rahman, M.U., Wang, L., and Wang, X. (2022). Ectopic expression of VvFUS3, B3-domain transcription factor, in tomato influences seed development via affecting endoreduplication and hormones. Hortic. Plant J. *8*, 351–360.
36. Zhang, S., Dong, R., Wang, Y., Li, X., Ji, M., and Wang, X. (2021). NAC domain gene VvNAC26 interacts with VvMADS9 and influences seed and fruit development. Plant Physiol. Biochem. *164*, 63–72.
37. Tang, Y., Wang, R., Gong, P., Li, S., Wang, Y., and Zhang, C. (2016). Gene cloning, expression and enzyme activity of Vitis vinifera vacuolar processing enzymes (VvVPEs). PLoS One *11*, e0160945.

38. Gong, P., Wei, R., Li, Y., Wang, R., Tang, Y., Wang, L., Zhu, H., Wang, Y., and Zhang, C. (2019). Molecular cloning and functional characterization of a seed-specific VvβVPE gene promoter from Vitis vinifera. Planta *250*, 657–665.

39. Wang, L., Yan, J., Zhang, X., Qu, Z., Yao, W., Tang, Y., Wang, R., Van Nocker, S., Wang, Y., Li, Y., et al. (2022). Genome-wide identification and comprehensive expression analysis of VviASN and VviGS gene families during seed development/abortion in grapevine. Sci. Hortic. *292*, 110625.

40. Cabezas, J.A., Cervera, M.T., Ruiz-García, L., Carreño, J., and Martínez-Zapater, J.M. (2006). A genetic analysis of seed and berry weight in grapevine. Genome *49*, 1572–1585.

41. Ocarez, N., Jiménez, N., Núñez, R., Perniola, R., Marsico, A.D., Cardone, M.F., Bergamini, C., and Mejía, N. (2020). Unraveling the deep genetic architecture for seedlessness in grapevine and the development and validation of a new set of markers for VviAGL11-based gene-assisted selection. Genes *11*, 151.

42. Kim, M.-S., Hur, Y.Y., Kim, J.H., and Jeong, S.-C. (2020). Genome resequencing, improvement of variant calling, and population genomic analyses provide insights into the seedlessness in the genus vitis. G3 (Bethesda) *10*, 3365–3377.

43. Zhang, T., Peng, W., Xiao, H., Cao, S., Chen, Z., Su, X., Luo, Y., Liu, Z., Peng, Y., Yang, X., et al. (2024). Population genomics highlights structural variations in local adaptation to saline coastal environments in woolly grape. J. Integr. Plant Biol. https://doi.org/10.1111/jipb.13653.

44. Long, Q., Cao, S., Huang, G., Wang, X., Liu, Z., Liu, W., Wang, Y., Xiao, H., Peng, Y., and Zhou, Y. (2024). Population comparative genomics discovers gene gain and loss during grapevine domestication. Plant Physiol. *195*, 1401–1413.

45. Peng, Y., Wang, Y., Liu, Y., Fang, X., Cheng, L., Long, Q., Xu, Q., Wang, N., Zhang, F., and Liu, Z. (2024). The genomic and epigenomic dynamics of hemizygous genes across crops with contrasting mating systems. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.24.595664.

46. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

47. Nesi, N., Debeaujon, I., Jond, C., Stewart, A.J., Jenkins, G.I., Caboche, M., and Lepiniec, L.c. (2002). The TRANSPARENT TESTA16 locus encodes the Arabidopsis BSISTER MADS domain protein and is required for proper development and pigmentation of the seed coat. Plant Cell *14*, 2463–2479.

48. De Folter, S., Shchennikova, A.V., Franken, J., Busscher, M., Baskar, R., Grossniklaus, U., Angenent, G.C., and Immink, R.G.H. (2006). A Bsister MADS-box gene involved in ovule and seed development in petunia and Arabidopsis. Plant J. *47*, 934–946.

49. Mizzotti, C., Mendes, M.A., Caporali, E., Schnittger, A., Kater, M.M., Battaglia, R., and Colombo, L. (2012). The MADS box genes seedstick and Arabidopsis Bsister play a maternal role in fertilization and seed development. Plant J. *70*, 409–420.

50. Gupta, M., Dey, S., Marbaniang, D., Pal, P., Ray, S., and Mazumder, B. (2020). Grape seed extract: having a potential health benefits. J. Food Sci. Technol. *57*, 1205–1215.

51. Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., and Gaut, B.S. (2019). The population genetics of structural variants in grapevine domestication. Nat. Plants *5*, 965–979.

52. Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol. Biol. Evol. *32*, 244–257.

53. Wang, N., Cao, S., Liu, Z., Xiao, H., Hu, J., Xu, X., Chen, P., Ma, Z., Ye, J., Chai, L., et al. (2023). Genomic conservation of crop wild relatives: A case study of citrus. PLoS Genet. *19*, e1010811.

54. Xiao, H., Liu, Z., Wang, N., Long, Q., Cao, S., Huang, G., Liu, W., Peng, Y., Riaz, S., Walker, A.M., et al. (2023). Adaptive and maladaptive introgression in grapevine domestication. Proc. Natl. Acad. Sci. USA *120*, e2222041120.

55. Ledbetter, C., and Ramming, D. (1989). Seedlessness in grapes. Hortic. Rev. *11*, 159–184.

56. Wang, L., Zhang, S., Jiao, C., Li, Z., Liu, C., and Wang, X. (2022). QTL-seq analysis of the seed size trait in grape provides new molecular insights on seedlessness. J. Integr. Agric. *21*, 2910–2925.

57. Wang, N., Li, C., Kuang, L., Wu, X., Xie, K., Zhu, A., Xu, Q., Larkin, R.M., Zhou, Y., Deng, X., et al. (2022). Pan-mitogenomics reveals the genetic basis of cytonuclear conflicts in citrus hybridization, domestication, and diversification. Proc. Natl. Acad. Sci. USA *119*, e2206076119.

58. Zhou, Y., Duvaux, L., Ren, G., Zhang, L., Savolainen, O., and Liu, J. (2017). Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. Heredity *118*, 211–220.

59. Liu Cm, C.M., McElver, J., Tzafrir, I., Joosen, R., Wittich, P., Patton, D., Van Lammeren, A.A.M., and Meinke, D. (2002). Condensin and cohesin knockouts in Arabidopsis exhibit a titan seed phenotype. Plant J. *29*, 405–415.

60. Liu, C., Wang, H., Cui, Z., He, X., Wang, X., Zeng, X., and Ma, H. (2007). Optimization of extraction and isolation for 11S and 7S globulins of soybean seed storage protein. Food Chem. *102*, 1310–1316.

61. Thomas, M., Edwards, K., and Pellerone, F. (2001). Grapevine microsatellite repeats: isolation, characterisation and use ofr genotyping of grape germplasm from Southern Italy. Vitis J. Grapevine Res. *40*, 179–186.

62. Wang, L., Yin, X., Cheng, C., Wang, H., Guo, R., Xu, X., Zhao, J., Zheng, Y., and Wang, X. (2015). Evolutionary and expression analysis of a MADS-box gene superfamily involved in ovule development of seeded and seedless grapevines. Mol. Genet. Genomics *290*, 825–846.

63. Sheng, Z., Xuan, X., Wang, F., Sadeghnezhad, E., Gong, P., Xiao, Y., Dong, T., Zhang, P., Wang, X., Fang, J., et al. (2022). Identification and characterization of AUXIN response factor gene family reveals their regulatory network to respond the multi-hormones crosstalk during GA-induced grape parthenocarpic berry. Int. J. Mol. Sci. *23*, 11108.

64. Tang, Y., Wang, Q., Xue, J., Li, Y., Li, R., Van Nocker, S., Wang, Y., and Zhang, C. (2018). Gene cloning and expression analyses of WBC genes in the developing grapevine seeds. J. Integr. Agric. *17*, 1348–1359.

65. Yim, B., Mun, J.H., Park, S.H., Shim, H.R., Kim, J.H., Hur, Y.Y., and Yu, H.J. (2020). Anatomical, biochemical and transcriptome analyses of Vitis vinifera cv.'Hongju'reveal novel information regarding the seed hardness of stenospermocarpic soft-seed grapes. Plant Breed. *139*, 672–683.

66. Li, Z., Jiao, Y., Zhang, C., Dou, M., Weng, K., Wang, Y., and Xu, Y. (2021). VvHDZ28 positively regulate salicylic acid biosynthesis during seed abortion in Thompson Seedless. Plant Biotechnol. J. *19*, 1824–1838.

67. Huang, J., Zhang, G., Li, Y., Lyu, M., Zhang, H., Zhang, N., and Chen, R. (2023). Integrative genomic and transcriptomic analyses of a bud sport mutant 'Jinzao Wuhe'with the phenotype of large berries in grapevines. PeerJ *11*, e14617.

68. Zhao, N., Lin, H., Lan, S., Jia, Q., Chen, X., Guo, H., and Chen, F. (2016). VvMJE1 of the grapevine (Vitis vinifera) VvMES methylesterase family encodes for methyl jasmonate esterase and has a role in stress response. Plant Physiol. Biochem. *102*, 125–132.

69. Cui, M., Wang, W., Guo, F., Fan, X., Guan, L., Zheng, T., Zhu, X., Jia, H., Fang, J., Wang, C., et al. (2020). Characterization and temporal–spatial expression analysis of LEC1 gene in the development of seedless berries in grape induced by gibberellin. Plant Growth Regul. *90*, 585–596.

70. Wang, Y., Liu, Z., Wu, J., Hong, L., Liang, J., Ren, Y., Guan, P., and Hu, J. (2021). MADS-box protein complex VvAG2, VvSEP3 and VvAGL11 regulates the formation of ovules in Vitis vinifera L. cv.'Xiangfei'. Genes *12*, 647.

71. Shewry, P.R., Napier, J.A., and Tatham, A.S. (1995). Seed storage proteins: structures and biosynthesis. Plant Cell *7*, 945–956.

72. Ahmad, B., Zhang, S., Yao, J., Rahman, M.U., Hanif, M., Zhu, Y., and Wang, X. (2019). Genomic organization of the B3-domain transcription factor family in grapevine (Vitis vinifera L.) and expression during seed development in seedless and seeded cultivars. Int. J. Mol. Sci. *20*, 4553.

73. Khan, S., and Stone, J.M. (2007). Arabidopsis thaliana GH3. 9 in auxin and jasmonate cross talk. Plant Signal. Behav. *2*, 483–485.

74. Brkljačić, J.M., Samardžić, J.T., Timotijević, G.S., and Maksimović, V.R. (2004). Expression analysis of buckwheat (Fagopyrum esculentum Moench) metallothionein-like gene (MT3) under different stress and physiological conditions. J. Plant Physiol. *161*, 741–746.

75. Tang, Y., Huang, C., Li, Y., Wang, Y., and Zhang, C. (2021). Genome-wide identification, phylogenetic analysis, and expression profiling of glycine-rich RNA-binding protein (GRPs) genes in seeded and seedless grapes (Vitis vinifera). Physiol. Mol. Biol. Plants *27*, 2231–2243.

76. Akkurt, M., Çakir, A., Shidfar, M., Çelikkol, B.P., and Söylemezoğlu, G. (2012). Using SCC8, SCF27 and VMC7f2 markers in grapevine breeding for seedlessness via marker assisted selection. Genet. Mol. Res. *11*, 2288–2294.

77. Dong, Z., Liu, W., Li, X., Tan, W., Zhao, Q., Wang, M., Ren, R., Ma, X., and Tang, X. (2018). Genetic relationships of 34 grapevine varieties and construction of molecular fingerprints by SSR markers. Biotechnol. Biotechnol. Equip. *32*, 942–950.

78. Wang, Y., Li, Y., Sun, F., Wu, G., Sulaiman, A., Bai, S., Hua, Z., and Guo, P. (2021). Embryo rescue and moleclar marker-assisted selection of hybrid seedless grape. https://doi.org/10.21203/rs.3.rs-1185496/v1.

79. Mejía, N., and Hinrichsen, P. (2003). A new, highly assertive SCAR marker potentially useful to assist selection for seedlessness in table grape breeding. Acta Hortic. *603*, 559–564.

80. Ma, Y., Feng, J., Liu, C., Fan, X., Sun, H., Jiang, J., and Zhang, Y. (2018). Development and application of SSR new molecular marker for seedless traits in grape. Sci. Agric. Sinica *51*, 2622–2630.

81. Slater, A.T., Cogan, N.O.I., Forster, J.W., Hayes, B.J., and Daetwyler, H.D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. Plant Genome *9*, https://doi.org/10.3835/plantgenome2016.02.0021.

82. Cappetta, E., Andolfo, G., Di Matteo, A., Barone, A., Frusciante, L., and Ercolano, M.R. (2020). Accelerating tomato breeding by exploiting genomic selection approaches. Plants (Basel) *9*, 1236.

83. Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. Nature *606*, 527–534.

84. Robertsen, C.D., Hjortshøj, R.L., and Janss, L.L. (2019). Genomic selection in cereal breeding. Agronomy *9*, 95.

85. Grenier, C., Cao, T.-V., Ospina, Y., Quintero, C., Châtel, M.H., Tohme, J., Courtois, B., and Ahmadi, N. (2015). Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. PLoS One *10*, e0136594.

86. Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., Li, C., Li, P., Yang, Z., Xu, C., et al. (2021). Genomic selection: A breakthrough technology in rice breeding. Crop J. *9*, 669–677.

87. Crossa, J., Pérez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., and Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. J. Crop Improv. *25*, 239–261.

88. Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. *22*, 961–975.

89. Gan, Y., Liu, Z., Zhang, F., Xu, Q., Wang, X., Xue, H., Su, X., Ma, W., Long, Q., and Ma, A. (2024). Deep learning based genomic breeding of pest-resistant grapevine. Preprint at bioRxiv. https://doi.org/10.1101/2024.03.16.585323.

90. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175.

91. Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., and Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. *23*, 258.

92. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. *3*, 95–98.

93. Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. *3*, 99–101.

94. Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92–95.

95. Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., DePamphilis, C.W., Yi, T.-S., and Li, D.-Z. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. *21*, 241.

96. Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P.L., et al. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat. Methods *17*, 1103–1110.

97. Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. *11*, 1432.

98. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One *11*, e0163962.

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

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

101. Shi, X., Cao, S., Wang, X., Huang, S., Wang, Y., Liu, Z., Liu, W., Leng, X., Peng, Y., Wang, N., et al. (2023). The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding. Hortic. Res. *10*, uhad061.

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

103. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLoS Comp. Biol. *14*, e1005944.

104. Goel, M., and Schneeberger, K. (2022). plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics *38*, 2922–2926.

105. Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. *20*, 277.

106. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; Subgroup, G.P.D.P. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

107. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

108. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

109. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arXiv.1303.3997.

110. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al.

(2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

111. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.
112. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
113. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.
114. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534.
115. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296.
116. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.
117. Hämälä, T., and Savolainen, O. (2019). Genomic patterns of local adaptation under gene flow in Arabidopsis lyrata. Mol. Biol. Evol. *36*, 2557–2571.
118. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
119. Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T., and Chang, W. (2022). David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. *50*, W216–W221.
120. Dong, S.-S., He, W.-M., Ji, J.-J., Zhang, C., Guo, Y., and Yang, T.-L. (2021). LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief. Bioinform. *22*, bbaa227.
121. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.
122. Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. Mol. Biol. Evol. *38*, 3022–3027.
123. Nicholas, K.B., Nicholas, H.B., and Deerfield, D.W.I. (1997). Gene Doc: analysis and visualization of genetic variation. EMBnet news *4*, 1–4.
124. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
125. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
126. Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y., and Xia, R. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol. Plant *13*, 1194–1202.
127. Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., and Chen, J. (2020). RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput. Sci. *6*, e251.
128. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. *108*, 1880–1890.
129. Massonnet, M., Cochetel, N., Minio, A., Vondras, A.M., Lin, J., Muyle, A., Garcia, J.F., Zhou, Y., Delledonne, M., Riaz, S., et al. (2020). The genetic basis of sex determination in grapes. Nat. Commun. *11*, 2902.
130. Yue, J., Chen, Q., Wang, Y., Zhang, L., Ye, C., Wang, X., Cao, S., Lin, Y., Huang, W., Xian, H., et al. (2023). Telomere-to-telomere and gap-free reference genome assembly of the kiwifruit Actinidia chinensis. Hortic. Res. *10*, uhac264.
131. Cochetel, N., Minio, A., Massonnet, M., Vondras, A.M., Figueroa-Balderas, R., and Cantu, D. (2021). Diploid chromosome-scale assembly of the Muscadinia rotundifolia genome supports chromosome fusion and disease resistance gene expansion during Vitis and Muscadinia divergence. G3 (Bethesda) *11*, jkab033.
132. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Samples were used for genome sequencing from Xinjiang Academy of Agricultural Sciences in China	This study	N/A
Critical commercial assays		
QIAGEN Genomic kit	QIAGEN, Germany	CAT#13343
g-TUBE	Covaris, USA	CAT#520079
AMPure PB Beads	PacBio, USA	CAT#100-265-900
SMRTbell Enzyme Cleanup Kit	PacBio, USA	CAT#101-932-600
Sequel II Binding Kit 2.0	PacBio, USA	CAT#101-789-500
TruSeq RNA Library Preparation Kit	Illuminia, USA	CAT#RS-122-2001
Deposited data		
PacBio HiFi sequencing reads	This study	NCBI PRJNA1021353
Illumina Hi-C sequencing reads	This study	NCBI PRJNA1021353
Raw RNA sequencing reads	This study	NCBI PRJNA1021353
Raw RNA sequencing reads	Li et al.[66]	NCBI PRJNA391820
Raw RNA sequencing reads	Yim et al.[65]	NCBI PRJNA533855
Raw RNA sequencing reads	Huang et al.[67]	NCBI PRJNA795246
29 raw reads for whole genome sequences	This study	NCBI PRJNA1021353
Raw reads for whole genome sequences	NCBI	See Data S1H
Software and algorithms		
Hifiasm v. 0.16.1-r375	Cheng et al.[90]	https://github.com/chhylp123/hifiasm
RagTag v. 2.1.0	Alonge et al.[91]	https://github.com/malonge/RagTag
Juicer v. 2.0	Durand et al.[92]	https://github.com/aidenlab/juicer
Juicebox v. 1.11.08	Durand et al.[93]	https://github.com/aidenlab/Juicebox
3D-DNA v. 201008	Dudchenko et al.[94]	https://github.com/aidenlab/3d-dna/tree/201008
GetOrganelle v1.7.7.0	Jin et al.[95]	https://github.com/Kinggerm/GetOrganelle
Flye v2.9.2	Kolmogorov et al.[96]	https://github.com/fenderglass/Flye/
GenomeScope2.0	Ranallo-Benavidez et al.[97]	https://github.com/tbenavi1/genomescope2.0
seqkit v. 2.2.0	Shen et al.[98]	https://github.com/shenwei356/seqkit
BUSCOs v. 5.3.0	Simão et al.[99]	https://busco.ezlab.org/
Merqury v. 1.3	Rhie et al.[100]	https://github.com/marbl/merqury
Pan-genome TE annotation	Shi et al.[101]	https://github.com/zhouyflab/PN40024_T2T_Genome
Liftoff v. 1.6.3	Shumate et al.[102]	https://github.com/agshumate/Liftoff
Mummer4 v. 4.0.0rc1	Marçais et al.[103]	https://mummer4.github.io/
Plotsr v. 0.5.4	Goel and Schneeberger[104]	https://github.com/schneebergerlab/plotsr
SyRI (v. 1.5.4)	Goel et al.[105]	https://github.com/schneebergerlab/syri
SAMtools v. 1.13	Li et al.[106]	https://github.com/samtools/samtools

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
OrthoFinder v. 2.5.4	Emms et al.[107]	https://github.com/davidemms/OrthoFinder
Fastp v. 0.23.2	Chen et al.[108]	https://github.com/OpenGene/fastp
BWA v. 0.7.17-r1188	Li[109]	https://github.com/lh3/bwa
GATK v.4.2.3.0	McKenna et al.[110]	https://www.broadinstitute.org/gatk/
GTX v. 2.1.11	N/A	http://www.gtxlab.com/product/cat
Delly v. 1.1.6	Rausch et al.[111]	https://github.com/dellytools/delly
VCFtools v. 0.1.16	Danecek et al.[112]	https://vcftools.github.io/
PLINK v. 1.90b6.21	Chang et al.[113]	https://www.cog-genomics.org/plink2
Iqtree v. 2.1.4-beta	Minh et al.[114]	https://github.com/iqtree/iqtree2
iTOLs	Letunic and Bork[115]	https://itol.embl.de/
Cytoscape v. 3.9.1	Shannon et al.[116]	https://cytoscape.org/
general_genomics	N/A	https://github.com/simonhmartin/genomics_general
PBScan	Hämälä and Savolainen[117]	https://github.com/thamala/PBScan
GEMMA v. 0.98.3	Zhou et al.[118]	https://github.com/genetics-statistics/GEMMA
DAVID	Sherman et al.[119]	https://david.ncifcrf.gov/tools.jsp
LDBlockshow v. 1.40	Dong et al.[120]	https://github.com/BGI-shenzhen/LDBlockShow
DIAMOND v. 2.0.15	Buchfink et al.[121]	https://github.com/bbuchfink/diamond
MEGA11	Tamura et al.[122]	https://www.megasoftware.net/
GeneDoc v. gd322700	KB[123]	http://nrbsc.org/gfx/genedoc
STAR v. 1.5.2	Dobin et al.[124]	https://github.com/alexdobin/STAR
DESeq2 v.1.36.0	Love et al.[125]	https://github.com/thelovelab/DESeq2
TBtools v. 1_113	Chen et al.[126]	https://github.com/CJ-Chen/TBtools-II
ggvenn v0.1.10	N/A	https://github.com/yanlinlin82/ggvenn
RIdeogram	Hao et al.[127]	https://github.com/TickingClock1992/RIdeogram
Beagle v. 5.2_21Apr21.304.jar	Browning et al.[128]	https://anaconda.org/bioconda/beagle
sklearn	N/A	https://scikit-learn.org/stable/install.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Prof. Yongfeng Zhou (zhouyongfeng@caas.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The raw sequencing data, comprising PacBio HiFi long-reads, Illumina Hi-C reads, RNA-seq reads, and 29 WGS grape accessions, is accessible on NCBI under BioProject ID: PRJNA1021353 and on the National Genomics Data Center (NGDC) under BioProject ID: PRJCA022010.
- The genome assembly and their annotations have been deposited into in Zenodo: https://doi.org/10.5281/zenodo.8278185, and NCBI BioSample accessions: SAMN37625511 and SAMN37625512.
- All scripts and codes performed in this study are available on GitHub: https://github.com/zhouyflab/Polygenic_Basis_Seedless_Grapes.
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

To enrich the genetic diversity of seedless grape varieties, we collected fresh tissues of grape varieties, such as "Thompson Seedless" (TS) and "Black Monukka" (BM), from the Anningqu Experimental Station (87°28′00″E, 45°56′00″ N) at the Xinjiang Academy of Agricultural Sciences in China.

METHOD DETAILS

Grape genome sequencing

Genomic DNA was extracted from grape leaves using CTAB method, followed by purification with the QIAGEN Genomic kit. For each sample, a total of 15 μg DNA was utilized for HiFi (SMRTbell) library preparation. The sheared DNA fragments (g-TUBE) underwent treatment with the SMRTbell Enzyme Cleanup Kit and purification using AMPure PB Beads. The resulting libraries were employed for HiFi sequencing on a PacBio Sequel II instrument (CCS mode) with Sequencing Primer V2 and Sequel II Binding Kit 2.0, yielding 59.78 Gbp and 60.35 Gbp of sequencing data for TS and BM, respectively.

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

Genome assembly

The detailed workflow of haplotype-resolved genome assembly and annotation is described in Figure S1. In brief, we utilized the Hi-C and HiFi data integrated assembly algorithm to generate contig reads by Hifiasm (v. 0.16.1-r375).[90] The contig reads were oriented and ordered to scaffold level using RagTag (v. 2.1.0)[91] with default parameters. Hi-C reads were also employed to anchor scaffolds onto chromosomes by Juicer 2.0[92] and Juicebox (v. 1.11.08).[93] The adjusted genome at the chromosome level was generated using 3D-DNA (v. 201008).[94] The full pipeline for genome assembly and gap filling can be found on our lab GitHub@zhouyflab (see Code availability). The chloroplast genome was de novo assembly using GetOrganelle toolkit (v1.7.7.0)[95] with K-mer parameters set to 21, 65, 105, and 127. For the high-quality mitochondrial genome, de novo assembly was performed using Flye (v2.9.2)[96] with HiFi long reads. Furthermore, using PN_T2T genome as the reference genome,[101] we utilized RagTag to assemble three chromosomes-level genomes based on scaffold level: 'Cabernet Sauvignon' (CS),[51] 'Black Corinth Seedless' (BC seedless),[129] and 'Black Corinth Seeded' (BC seeded).[129]

Genome annotation

The genome-wide annotation pipeline, identification of telomeres and centromeres, and annotation of transposable elements (TEs) were referred from pervious study.[101,130] The statistic results of TEs classification through Pan-genome TE annotation[101] can be found in Data S1B. Additional details on telomere regions, telomere copy numbers, and centromere regions for each chromosome are provided in Data S1C. The CS genome was annotated based on sequence similarity using Liftoff (v. 1.6.3)[102] and PN_T2T annotation files.

QUANTIFICATION AND STATISTICAL ANALYSIS

Assembly assessment

HiFi data from TS and BM were performed to assess genome heterozygosity based on k-mer using GenomeScope2.0.[97] Meanwhile, basic genome statistics were calculated using seqkit (v. 2.2.0),[98] which included genome length, N50, and GC content. The completeness of the haplotype genomes was evaluated using BUSCOs (v. 5.3.0)[99] with the embryophyta_odb10 database. Merqury (v. 1.3, best_ k=19)[100] was employed to evaluate the quality value (QV) and completeness of the haplotype genomes based on whole-genome sequencing data. The calc_switchErr tool was used for estimation of switch error (https://github.com/tangerzhang/calc_switchErr/). The mapping rate of HiFi and Hi-C reads were determined using minimap2 (with parameters: -ax map-hifi; -ax sr) and samtools (flagstat) toolkits. The Hi-C interactive signals on grape genome were visualized using Juicebox.

Comparative genomics

For genome level variant calling, we selected a total of 15 grape assemblies, including ten seeded assemblies: 'PN40024' (PN_T2T), 'Cabernet Sauvignon' (CS, hap1 and hap2),[129] 'Muscat Hamburg' (MH, hap1 and hap2), 'Shine Muscat' (SM, hap1 and hap2), 'Muscadinia rotundifolia' (MR, hap1 and hap2),[131] BC Seeded,[129] as well as five seedless assemblies: TS (hap1 and hap2), BM (hap1 and hap2), BC Seedless.[129] Among them, TS and BM were newly sequenced in this study. Four recently synchronized haplotype-resolved genomes of MH and SM will be published in another study.

CellPress
OPEN ACCESS

All chromosome-level genomes were aligned with PN_T2T genome using (v. 4.0.0rc1),[103] and the results were visualized using Plotsr (v. 0.5.4)[104] and Linux based Gunplot. R script was used for visualizing the gene density (line) of the reference genome (see Code availability). To validate the authenticity of SVs, the raw reads of HiFi and Hi-C were mapped to their haplotype genomes using Minimap2. The corresponding BAM files were extracted using SAMtools (v. 1.13)[106] and then inputted into the IGV software for inversions validation. In addition, the genome consensus phylogenetic tree was constructed using OrthoFinder (v. 2.5.4)[107] based on the single-copy orthologous genes from the whole genomes.

WGS variation detection

To genotype SNPs, InDels and SVs in 548 accessions, low-quality resequencing reads were removed using Fastp (v. 0.23.2)[108] with default parameters. The filtered data were then mapped to the CS reference genome using BWA(v. 0.7.17-r1188).[109] Non-uniquely mapped and duplicated reads were excluded using SAMtools and GATK (v.4.2.3.0).[110] Subsequently, SNPs and InDels calling were performed using GTX (v. 2.1.11, http://www.gtxlab.com/product/cat), followed by the merging of genotyping of the gVCF files into a VCF file. Delly (v. 1.1.6)[111] was used to SVs calling with default parameters. Basic filtering of the VCF file was performed using VCFtools (v. 0.1.16)[112] with the following parameters: --max-missing 0.8, --minGQ 20, --min-alleles 2, --max-alleles 2, --minDP 4, --maxDP 1000, and—maf 0.0005. PLINK (v. 1.90b6.21)[113] was utilized to reduce the size of the VCF file and improve computational efficiency. Finally, we obtained a total of 4,462,797 SNPs, 443,812 InDels, and 487,204 SVs from 548 grape accessions in the nuclear genome (NCBI GenBank accession number FM179380.1), 38,267 variation sites (SNPs and InDels) from 312 grape accessions in the mitochondrion genome, and 2,247 variation sites (SNPs and InDels) from 312 grape accessions in the chloroplast genome (NCBI GenBank accession number DQ424856.1).

Population genetics analysis

Phylogenetic analysis was conducted using 250,821 high-quality SNPs filtered for LD by PLINK, with parameters: --indep-pairwise 20 5 0.2 --geno 0.1. The phylogenetic tree was inferred by Iqtree (v. 2.1.4-beta)[114] with parameters: -m GTR+I+G -bb 1000 -bnni -alrt 1000 -st DNA, and the result were visualized using iTOLs.[115] Similarly, the phylogenetic analysis of the mitochondrial and chloroplast genomes yielded 9,647 and 758 high-quality SNPs using PLINK (--geno 0.2 --maf 0.001), respectively. The PCA and IBD analysis were preformed using PLINK and visualized using R scripts and Cytoscape (v. 3.9.1),[116] respectively. VCFtools was employed to calculate the F_{ST} and π statistics at whole genome level with 20 kb window size. Population introgression analysis was calculated using the "ABBABABAwindows.py" script from the "general_genomics" tool (https://github.com/simonhmartin/genomics_general) with 20 kb window size. The population branch statistic (PBS) assesses differentiation in the same branch of the phylogenetic tree with 50 SNP per window between seeded and seedless grapes in both VV and VV×VL populati ons.[117] PBScan was utilized to estimate population differentiation using D_{xy} (-div 2).

Genome-wide association study

We selected three populations for GWAS analysis, including VV population (n = 352), VV×VL population (n = 92), and the admixed population (n = 444). High-quality variants were obtained using PLINK with MAF $\geq$ 0.05, Hardy-Weinberg equilibrium < 1e-5 and missing < 0.2, resulting in 2,086,600 variants (1,881,457 SNPs, 173,547 InDels, and 31,596 SVs), 1,419,982 variants (1,274,130 SNPs, 124,054 InDels, and 21,798 SVs), and 2,292,404 variants (2,065,307 SNPs, 192,536 InDels, and 34,561 SVs), respectively. Additionally, we used the same filtering conditions to obtain 2,630 variants in mitochondrial genome, while none of variants met the quality control criteria for the chloroplast genome in Figure S5F. GWAS analysis utilized the mixed linear model in GEMMA (v. 0.98.3)[118] with the first three PCAs as a random effect matrix. Variant positions and p-wald test statistics (*P*-value) were extracted to generate Manhattan plots and Q-Q plots by R scripts. A Python script was utilized to identify the associated genes within 5 kb windows of the significant variants, which were above the significance threshold of $-\log_{10}(0.05/\text{Variant Numbers})$.

GO enrichment analysis was employed using the web tool DAVID,[119] and the results were visualized with R script. The SDI locus was identified through molecular markers SNP-25.24 and SNP-26.93.[18] The LD linkage heatmap was visualized using LDBlockshow (v. 1.40).[120] Moreover, DIAMOND (v. 2.0.15)[121] was employed for sequence comparisons of whole-genome homologous protein from 14 grape assemblies (Seeded: Seedless: Outgroup = 8:5:1). Sequence alignments of candidate proteins were conducted using ClustalW in MEGA11,[122] and sequence visualization was accomplished using GeneDoc (v. gd322700).[123] 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[65–67] (see Data S1O). 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 the reference, transcriptomic assembly was conducted on the processed data using STAR (v. 1.5.2).[124] The R package DESeq2 (v.1.36.0)[125] was utilized for PCA analysis and pairwise comparisons between seeded and seedless grape samples at each development time point (Figures S8A–S8C). The thresholds for differential expression genes were $|\text{Log2FoldChange}| \geq 2$ and P-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, Figures S8E and S8F). 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).[126] The results were visualized using Venn diagrams and whole-genome density plots using the R package ggvenn (v0.1.10) and RIdeogram,[127] respectively. The integrated results in this study, including GWAS results, RNA-seq results, GO enrichment results, and previously reported family genes, was performed for data visualization using an R script, which included UpSet plot analysis and gene expression heatmaps (see Code availability). The definition of 'core genes' is as follows: For example, in a set of RNA-seq data, PN and TS (Data S1K), we denote the average FPKM values across all time points as AVE. The ratio of AVE in PN to AVE in TS is defined as fold change (Fold). If AVE for PN is greater than AVE for TS (Fold $\geq$ 2.0 and PN AVE $\geq$ 100), then these genes were selected. Conversely, if AVE for PN is less than AVE for TS (Fold $\leq$ 0.5 and TS AVE $\geq$ 100), then these genes were also selected. Genes meeting these criteria are collectively referred to as 'core genes.'

Genome selection based on machine learning

For genome selection, we utilized the admixed population (n = 444) as training set, and additional 39 samples as testing set for phenotype prediction, as pipeline showed in Figure S10. Beagle (v. 5.2_21Apr21.304.jar)[128] was used to impute the VCF file. Subsequently, the 794 variants, above the significance threshold of $-\log_{10}(0.05/\text{Variant Numbers})$ ($\sim$ 7.66), were extracted from imputed VCF using BCFtools (v. 1.13),[132] including 717 SNPs and 77 InDels. To break the linkage among the 794 variants, we used PLINK –r2 to calculate LD metrics and manually removed one of two variants whenever the LD score exceeded 0.8, 0.5, or 0.2 among 794 variants. We subsequently obtained 137, 54, and 45 variant sites, respectively. For model selection, we utilized the Python package 'sklearn' (https://scikit-learn.org/stable/install.html) and selected nine classical models: Cross-validated Elastic Net model (ElasticNetCV), Kernel Ridge Regression (KernelRidge), Lasso Regression (Lasso), Linear Regression (Linear), Logistic Regression (Logistic), PLS Regression, Linear Ridge Regression (Ridge), Linear Support Vector Regression (SVR-Linear), and Polynomial Support Vector Regression (SVR-poly). After 100 repetitions of random five-fold cross-validation with the training set, we chose the models with best performance, for the prediction of testing set (Data S1H). Moreover, the Iqtree was employed to construct an unrooted tree, while iTOLs was used for the visualization of the phylogenetic tree. The data visualization used R scripts.

Current Biology, Volume *34*

Supplemental Information

Integrative genomics reveals the polygenic basis of seedlessness in grapevine

Xu Wang, Zhongjie Liu, Fan Zhang, Hua Xiao, Shuo Cao, Hui Xue, Wenwen Liu, Ying Su, Zhenya Liu, Haixia Zhong, Fuchun Zhang, Bilal Ahmad, Qiming Long, Yingchun Zhang, Yuting Liu, Yu Gan, Ting Hou, Zhongxin Jin, Xinyu Wu, Guotian Liu, Yiwen Wang, Yanling Peng, and Yongfeng Zhou

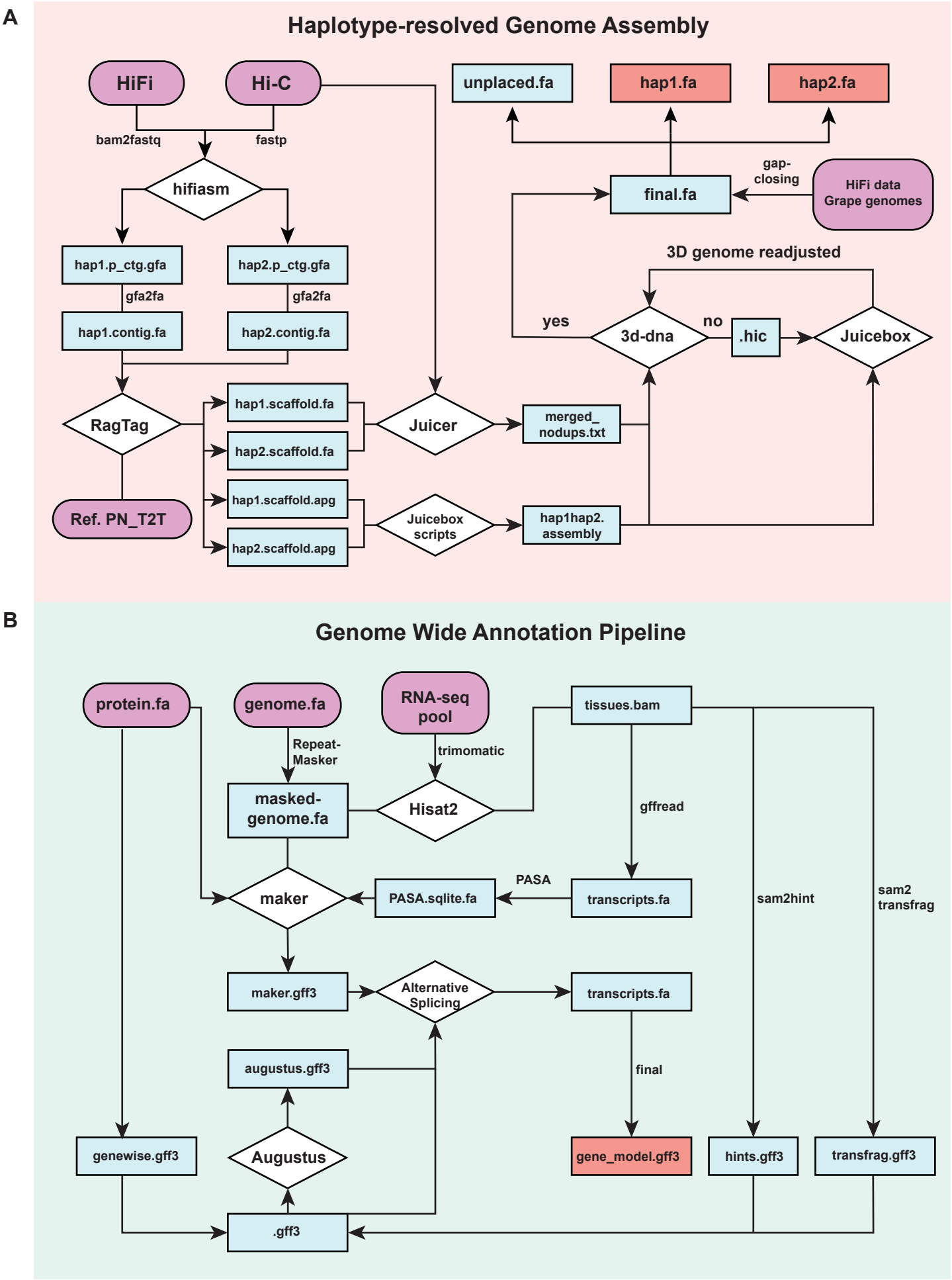


Figure S1. Complete workflow for haplotype-resolved genome assembly and annotation, Related to STAR Methods. Additional details can be found on our lab website GitHub@zhouyflab, related to STAR Methods for genome assembly and genome annotation.

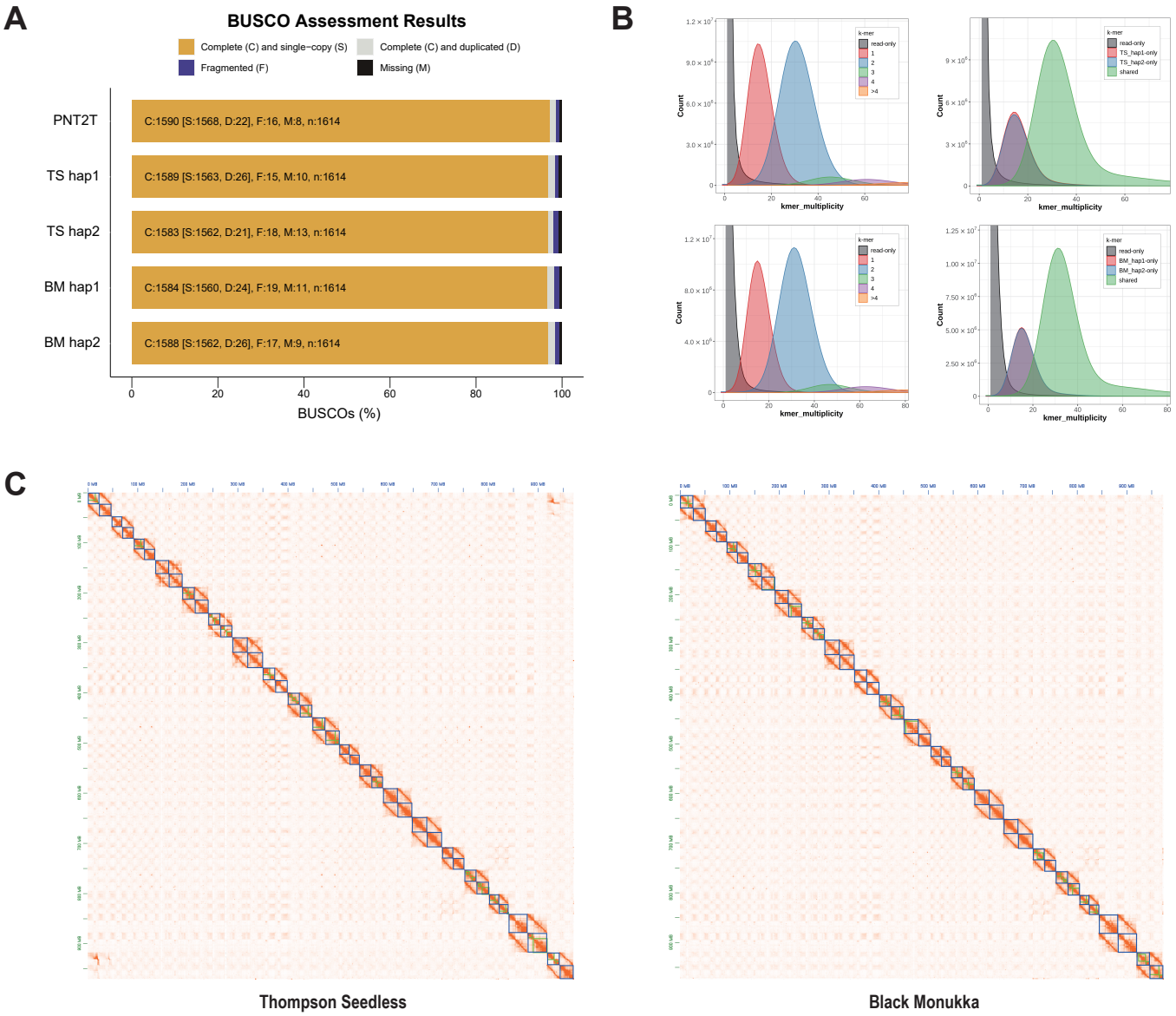


Figure S2. Evaluation of four haplotype-resolved genomes, Related to Figure 1. (A) BUSCO assessment of genome com-pleteness using the embryophyta_odb10 database. (B) Evaluation of quality value (QV) and haplotype completeness using Merqury based on k-mer. (C) Visualization of the diploid genome via Hi-C heatmap using Juicebox, related to Figure 1(E). Most single contig (green rectangle) are directly composed of chromosomes (blue rectan-gles).

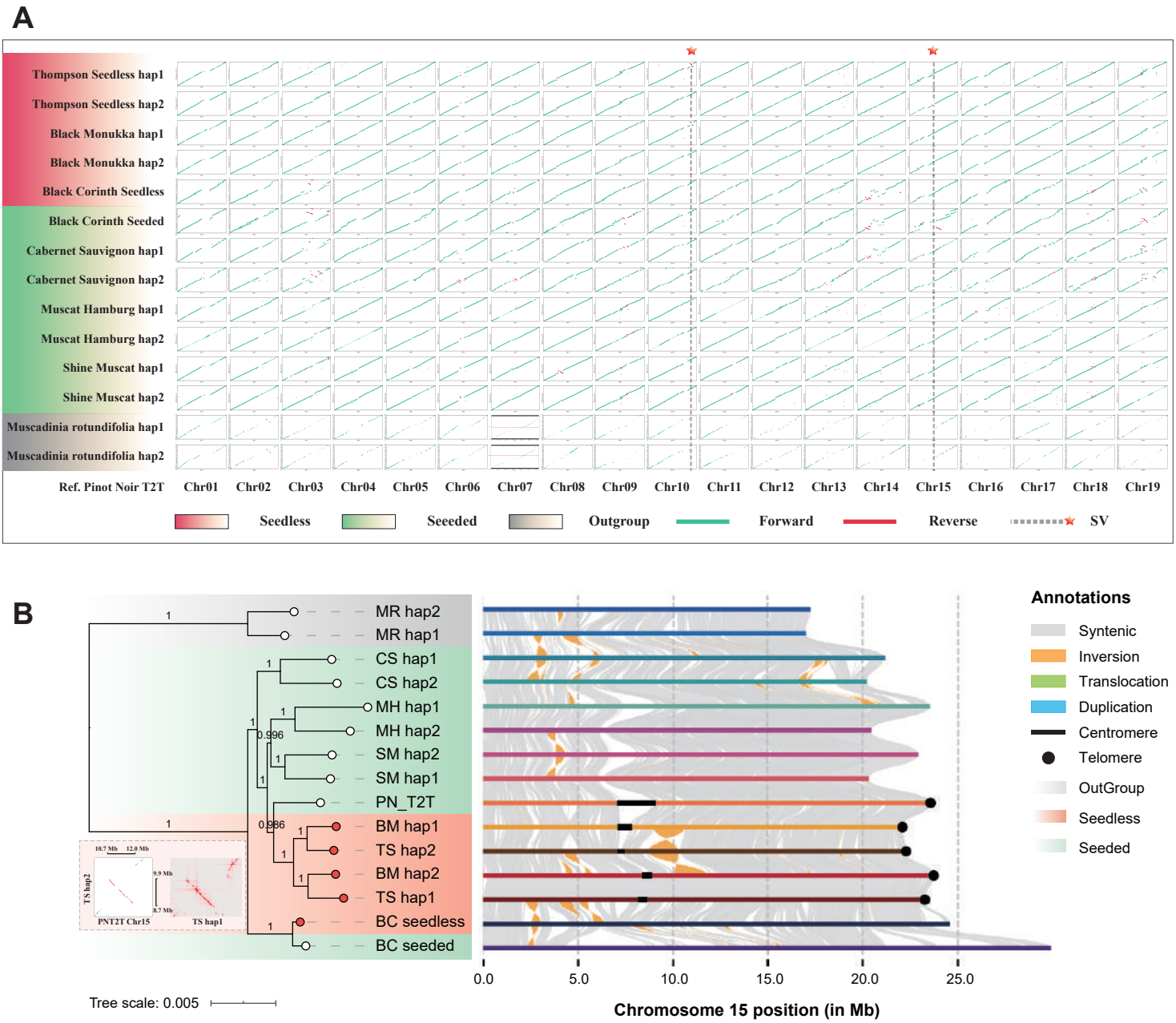


Figure S3. Comparative genomics results, Related to Figure 1. (A) Sequence alignment of 15 grape genomes with the PN_T2T genome, related to Figure 1(D). Red blocks represent seedless samples, green blocks indicate seeded samples, and gray blocks denote outgroup samples. Red stars highlight inversions associated with seed abortion. The Chr07 of 'Muscadinia rotundifolia' is composed of Chr07 and Chr20. (B) Sequence alignment results of Chr15 for the 15 genomes, as well as the inversion Hi-C heatmap in inversion boundary. The phylogenetic tree was constructed using single-copy genes from the whole genome proteins.

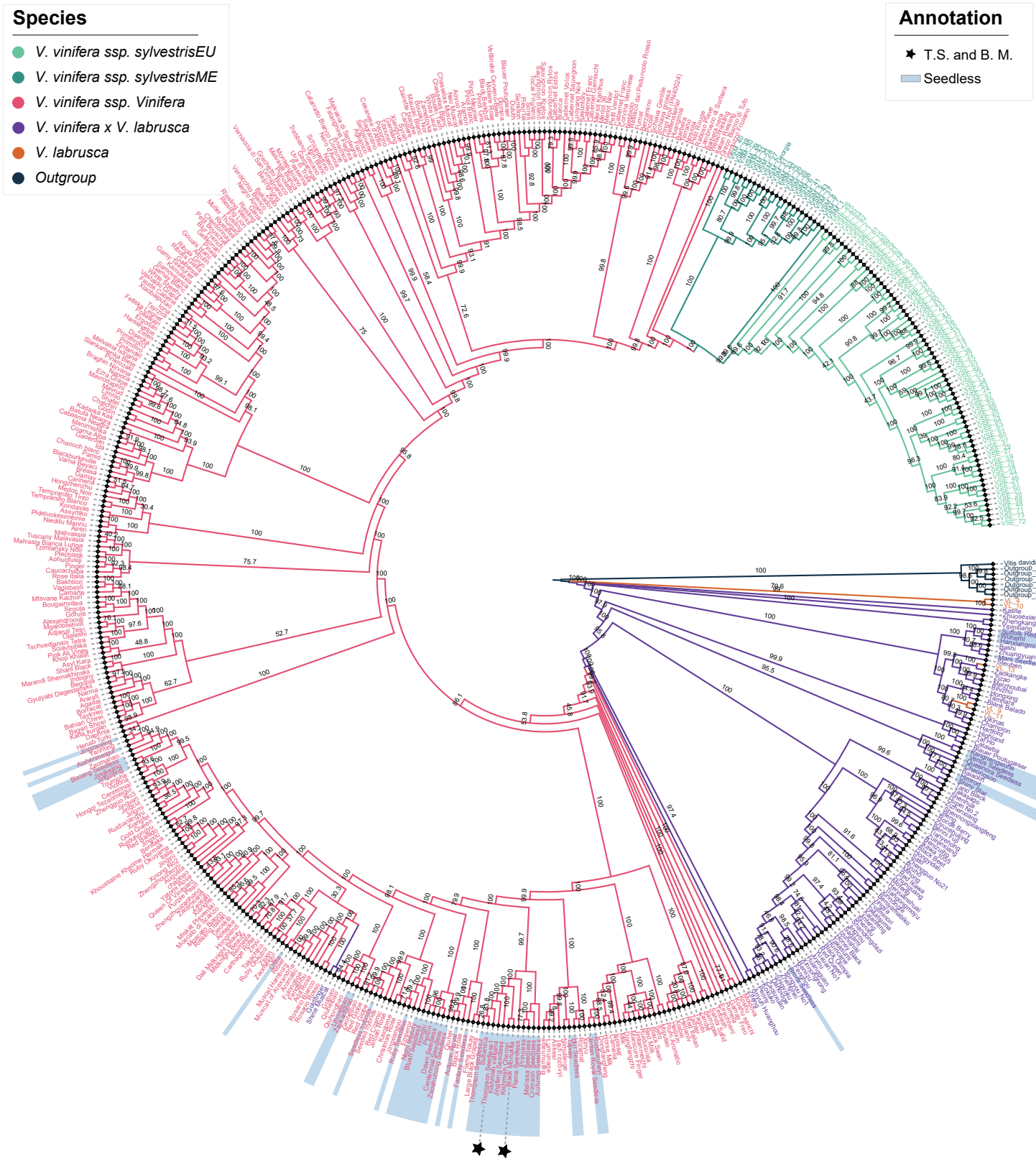


Figure S4. Detailed phylogenetic tree of 548 grapevine accessions, Related to Figure 2. This figure provides a full zoomed-in version of **Figure 2C**, which includes the six populations. Light-blue blocks represent seed abortion samples and black star symbols indicate TS and BM.

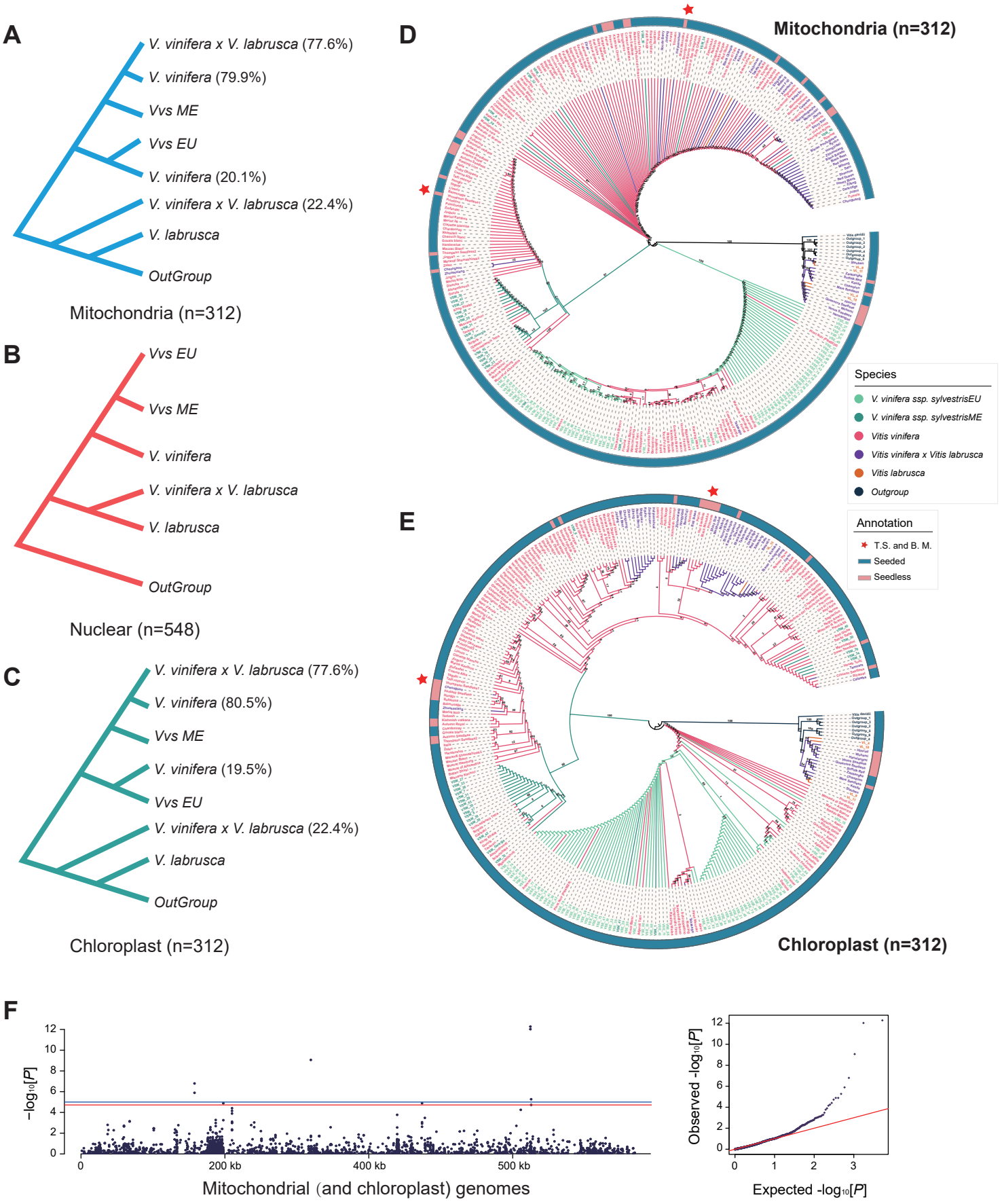


Figure S5. Phylogenetic tree of mitochondrial and chloroplast genomes in 312 grapevine accessions, Related STAR Methods. (A-C) represent the consensus phylogenies constructed based on the mitochondrial genomes, nuclear genomes, and chloroplast genomes, respectively. (D-E) depict the complete phylogenetic trees of the mitochondrial and chlo-roplast, encompassing six populations. (F) Genome-Wide Association Study (GWAS) and Q-Q plot for mito-chondrial and chloroplast genomes based on 312 grape accessions. All variants of chloroplast genomes were filtered out by quality control and referred to STAR Methods for genome-wide association study. The sample information is available in Data S1H.

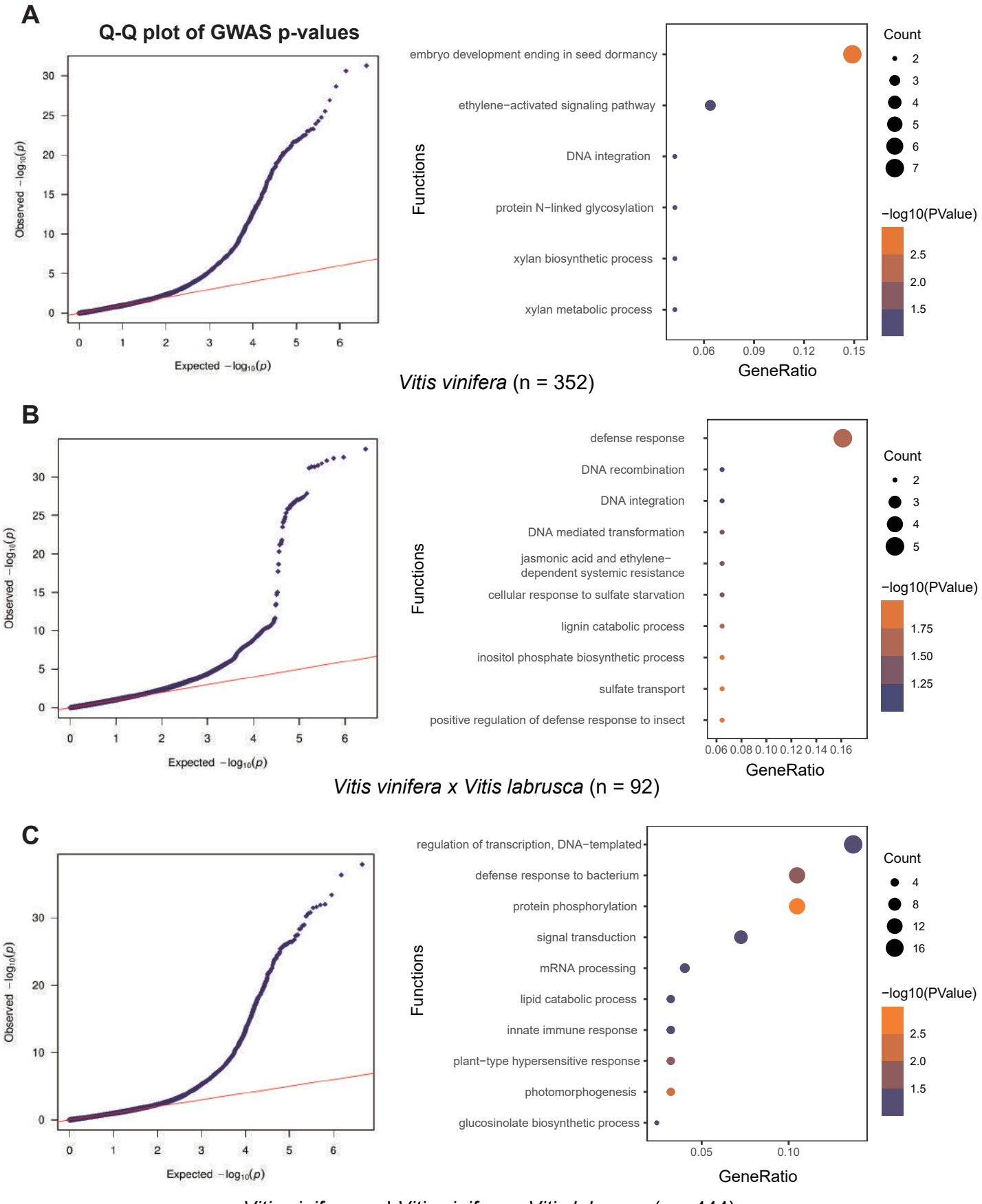


Figure S6. Quantile-Quantile (Q-Q) plot and Gene Ontology (GO) enrichment analysis, Related to Figure 3. Genome-Wide Association Study (GWAS) Q-Q plot for the three populations in Figure 3(A) and 3(B), along with GO enrichment analysis for biological processes in genes specific to VV×VL and VV, as well as those consensus gene for admixed populations in Data S1L.

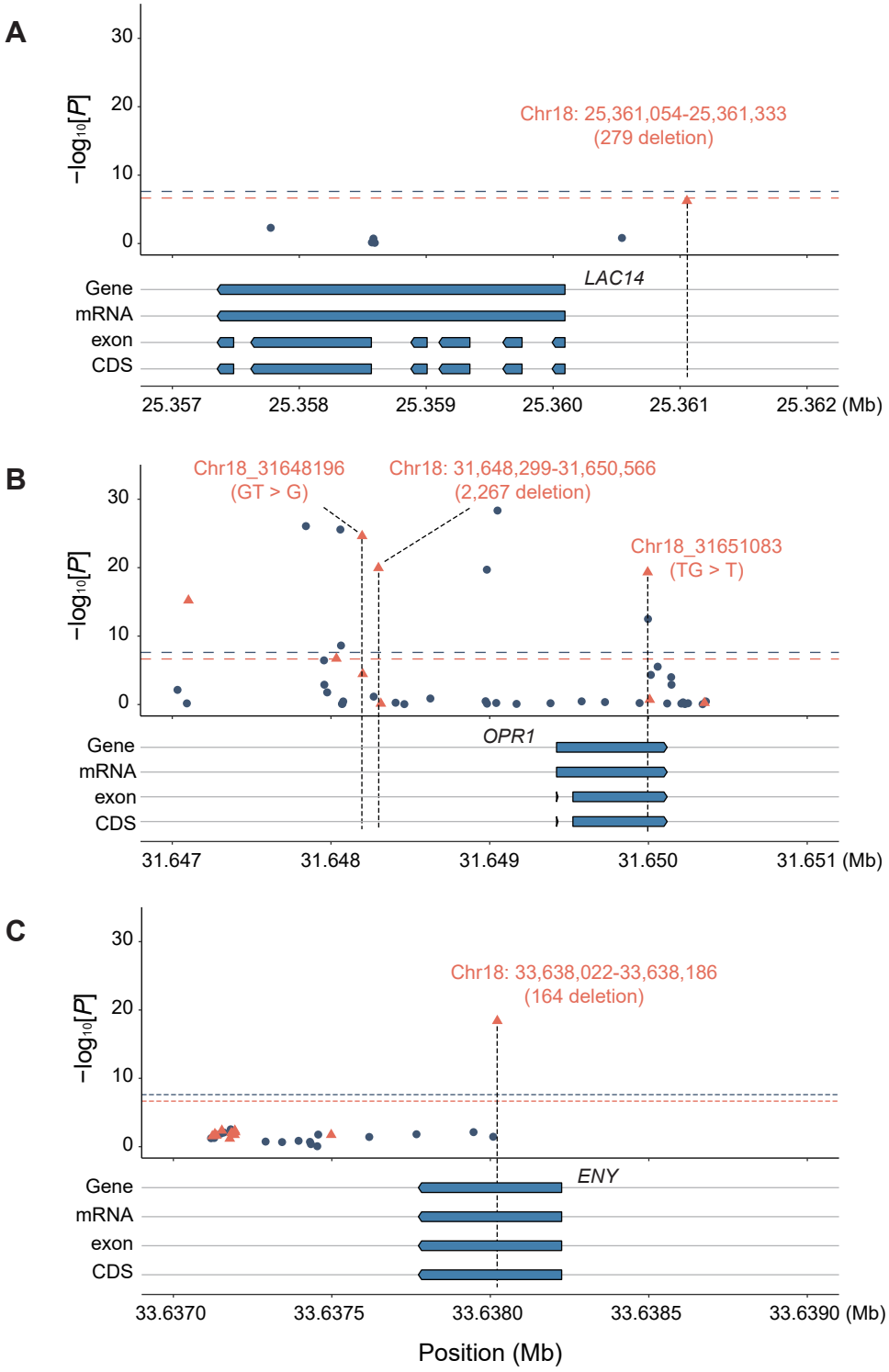


Figure S7. Three candidate genes related to seedlessness in Chr18, Related to Figure 3. Candidate genes were identified through GWAS analysis using the admixed population, referred to Figure 3(B), with a significant threshold of 7.61 for SNPs and 6.66 for SVs (and InDels).

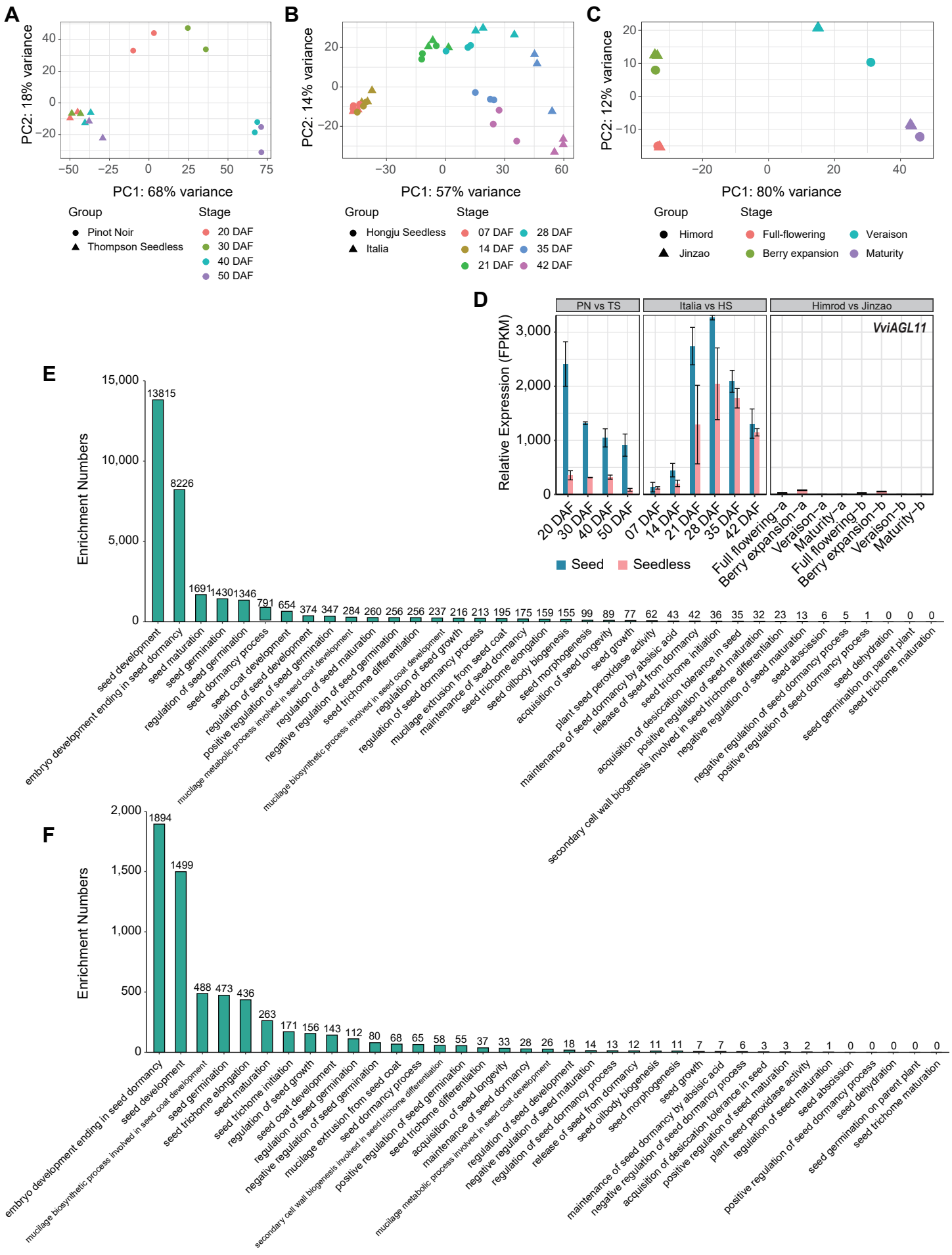


Figure S8. Transcriptomic analysis and GO homologous gene enrichment related to seed development, Related to STAR Methods. (A-C) Principal Component Analysis (PCA) of the three transcriptomic datasets. (D) Relative expression values (FPKM) for VviAGL11. (E) Enrichment analysis of 14,650 seed development genes from the GO database. (F) Enrichment analysis of 6,529 GO homologous genes associated with grape seed development, related to STAR Methods for integrative genomic analysis.

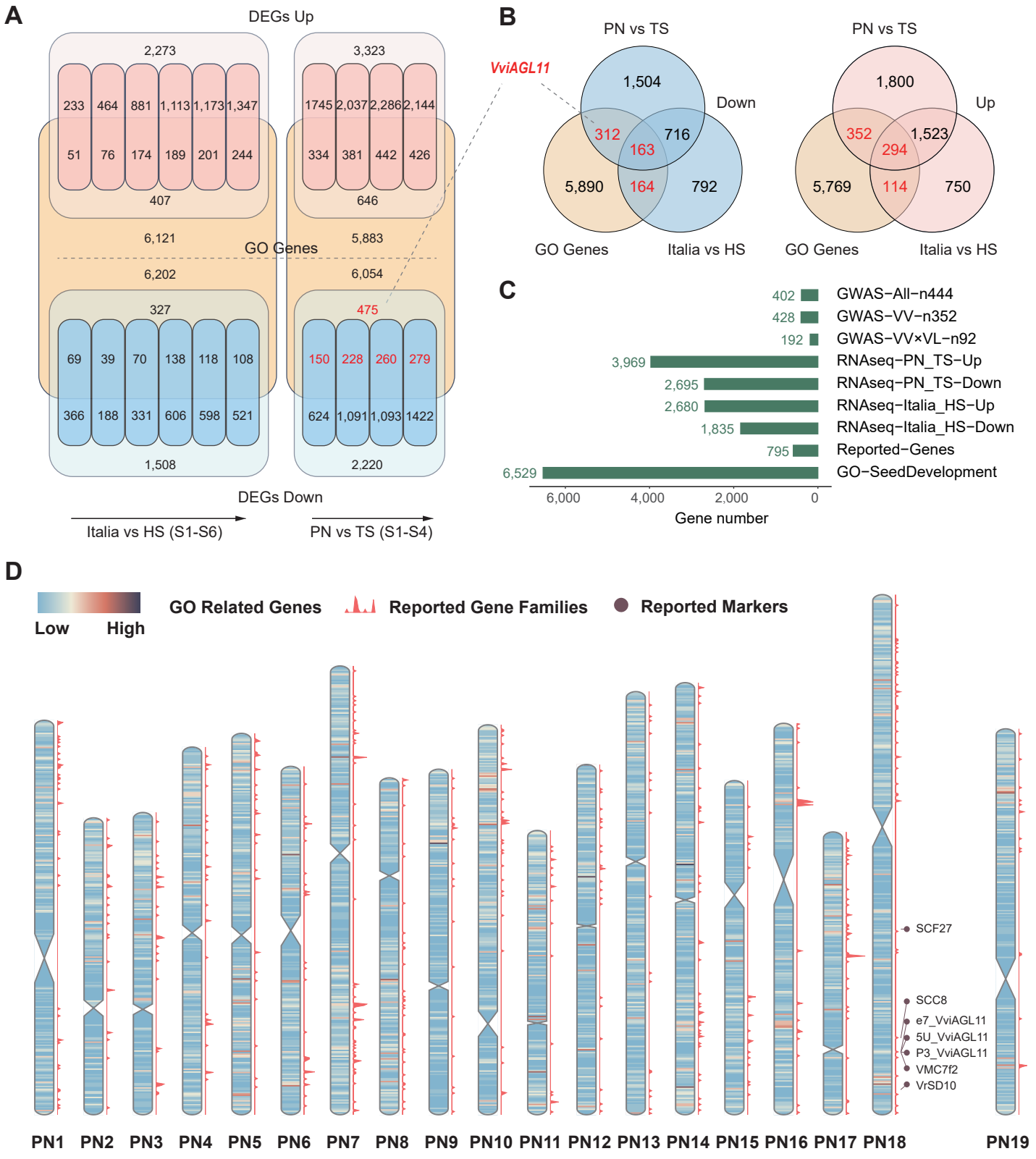
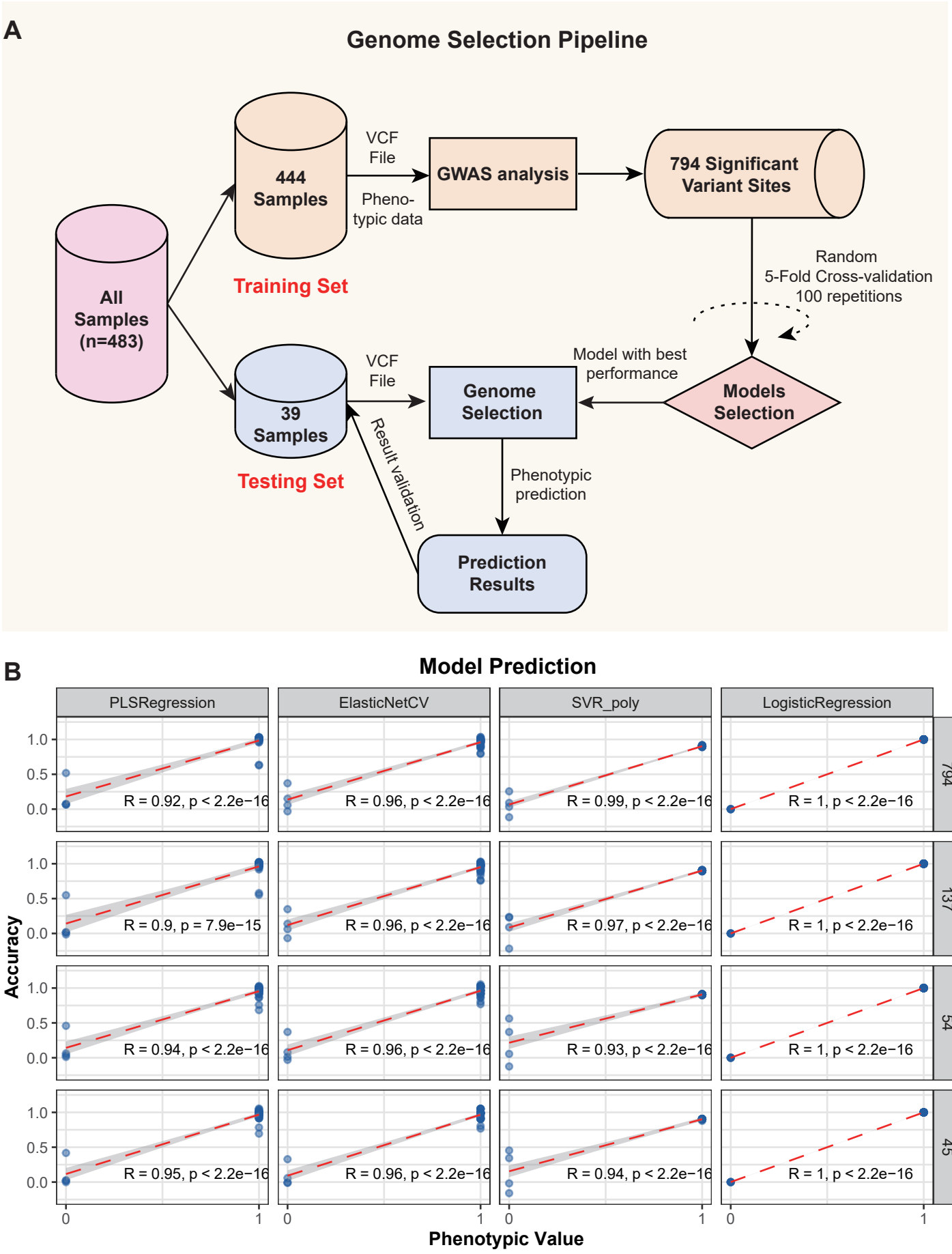


Figure S9. Visualization of multiple seed development-related datasets, Related to Figure 5. (A-B) Results from two inde-pendent transcriptomic analyses, 'Italia' vs 'Hongju Seedless' (HS) and 'Pinot Noir' (PN) vs 'Thompson Seedless' (TS), overlapping with GO homologous gene. Red indicates up-regulated DEGs, blue represents down-regulated DEGs, and yellow denotes GO homologous genes associated with seedlessness. (C) Results from integrative genomic analyses: GWAS, transcriptomics, reported genes mapping, and GO homologous genes. Green bars represent the total number of genes identified through this approach. (D) Visualization of three datasets: GO homologous genes, reported gene families, and reported molecular markers. These results correspond to Figure 5(A).



Figiure S10. Workflow of genome selection and visualization of top model results, Related to Figure 6. (A) Identification of 794 significant variants from GWAS, comprising 77 InDels and 717 SNPs, used as features for model selection and building as shown in Figure 6(D). The top-performing models were utilized for phenotypic prediction. (B) Dimensionality reduction of input data was achieved by eliminating strong linkage disequilibrium (LD). A reduced number of variant sites maintained high predictive accuracy, particularly for binary classification problems using Logistic Regression. With LD < 0.8, 137 variant sites were retained; LD < 0.5, 54 sites remained; and LD < 0.2, 45 sites were conserved. Phenotypic values: 0 for seedless and 1 for seeded grapes. These results correspond to Figure 6(C).

Supplemental References

S1. Wang, L., Yin, X., Cheng, C., Wang, H., Guo, R., Xu, X., Zhao, J., Zheng, Y., and Wang, X. (2015). Evolutionary and expression analysis of a MADS-box gene superfamily involved in ovule development of seeded and seedless grapevines. Molecular Genetics and Genomics *290*, 825-846.

S2. Wang, Y., Liu, Z., Wu, J., Hong, L., Liang, J., Ren, Y., Guan, P., and Hu, J. (2021). MADS-Box Protein Complex VvAG2, VvSEP3 and VvAGL11 Regulates the Formation of Ovules in Vitis vinifera L. cv.'Xiangfei'. Genes *12*, 647.

S3. di Rienzo, V., Imanifard, Z., Mascio, I., Gasser, C.S., Skinner, D.J., Pierri, C.L., Marini, M., Fanelli, V., Sabetta, W., and Montemurro, C. (2021). Functional conservation of the grapevine candidate gene INNER NO OUTER for ovule development and seed formation. Horticulture research *8*.

S4. Zhang, S., Wang, L., Sun, X., Li, Y., Yao, J., Nocker, S.v., and Wang, X. (2019). Genome-wide analysis of the YABBY gene family in grapevine and functional characterization of VvYABBY4. Frontiers in plant science *10*, 1207.

S5. Tang, Y., Huang, C., Li, Y., Wang, Y., and Zhang, C. (2021). Genome-wide identification, phylogenetic analysis, and expression profiling of glycine-rich RNA-binding protein (GRPs) genes in seeded and seedless grapes (Vitis vinifera). Physiology and Molecular Biology of Plants *27*, 2231-2243.

S6. Wang, L., Yan, J., Zhang, X., Qu, Z., Yao, W., Tang, Y., Wang, R., Van Nocker, S., Wang, Y., and Li, Y. (2022). Genome-wide identification and comprehensive expression analysis of VviASN and VviGS gene families during seed development/abortion in grapevine. Scientia Horticulturae *292*, 110625.

S7. TANG, Y.-j., Qian, W., XUE, J.-y., Yan, L., LI, R.-m., Van Nocker, S., WANG, Y.-j., and ZHANG, C.-h. (2018). Gene cloning and expression analyses of WBC genes in the developing grapevine seeds. Journal of Integrative Agriculture *17*, 1348-1359.

S8. Ahmad, B., Zhang, S., Yao, J., Rahman, M.U., Hanif, M., Zhu, Y., and Wang, X. (2019). Genomic organization of the B3-domain transcription factor family in grapevine (Vitis vinifera L.) and expression during seed development in seedless and seeded cultivars. International Journal of Molecular Sciences *20*, 4553.

S9. Zhu, Y., Li, Y., Zhang, S., Zhang, X., Yao, J., Luo, Q., Sun, F., and Wang, X. (2019). Genome-wide identification and expression analysis reveal the potential function of ethylene responsive factor gene family in response to Botrytis cinerea infection and ovule development in grapes (Vitis vinifera L.). Plant Biology *21*, 571-584.

S10. Zhang, X., Zhang, L., Ji, M., Wu, Y., Zhang, S., Zhu, Y., Yao, J., Li, Z., Gao, H., and Wang, X. (2021). Genome-wide identification and expression analysis of the B-box transcription factor gene family in grapevine (Vitis vinifera L.). BMC genomics *22*, 1-16.

S11. Wang, L., Xue, J., Dai, W., Tang, Y., Gong, P., Wang, Y., and Zhang, C. (2019). Genome-wide identification, phylogenetic analysis, and expression profiling of CONSTANS-like (COL) genes in Vitis vinifera. Journal of Plant Growth Regulation *38*, 631-643.

S12. Li, Z., Zhang, C., Guo, Y., Niu, W., Wang, Y., and Xu, Y. (2017). Evolution and expression analysis reveal the potential role of the HD-Zip gene family in regulation of embryo

abortion in grapes (Vitis vinifera L.). BMC genomics *18*, 1-16.

S13. Wang, L., Zhang, S., Zhang, X., Hu, X., Guo, C., Wang, X., and Song, J. (2018). Evolutionary and expression analysis of Vitis vinifera OFP gene family. Plant Systematics and Evolution *304*, 995-1008.

S14. Li, S., Geng, X., Chen, S., Liu, K., Yu, S., Wang, X., Zhang, C., Zhang, J., Wen, Y., and Luo, Q. (2021). The co-expression of genes involved in seed coat and endosperm development promotes seed abortion in grapevine. Planta *254*, 1-16.

S15. Wang, L., Ahmad, B., Liang, C., Shi, X., Sun, R., Zhang, S., and Du, G. (2020). Bioinformatics and expression analysis of histone modification genes in grapevine predict their involvement in seed development, powdery mildew resistance, and hormonal signaling. BMC Plant Biology *20*, 1-20.

S16. Li, Y., Zhu, Y., Yao, J., Zhang, S., Wang, L., Guo, C., Van Nocker, S., and Wang, X. (2017). Genome-wide identification and expression analyses of the homeobox transcription factor family during ovule development in seedless and seeded grapes. Scientific Reports *7*, 12638.

S17. Kong, J., Zhao, T., Liu, B., and Li, Y. (2022). Cloning and Expression Analysis of Seven VvGA2ox Genes in Ovule Development of Seeded and Seedless Grapes (Vitis vinifera).

S18. Sheng, Z., Xuan, X., Wang, F., Sadeghnezhad, E., Gong, P., Xiao, Y., Dong, T., Zhang, P., Wang, X., and Fang, J. (2022). Identification and Characterization of AUXIN Response Factor Gene Family Reveals Their Regulatory Network to Respond the Multi-Hormones Crosstalk during GA-Induced Grape Parthenocarpic Berry. International Journal of Molecular Sciences *23*, 11108.

S19. Wang, L., Dai, W., Shi, Y., Wang, Y., and Zhang, C. (2022). Cloning and activity analysis of the highly expressed gene VviABCG20 promoter in seed and its activity is negatively regulated by the transcription factor VviDof14. Plant Science *315*, 111152.

S20. Tang, Y., Wang, R., Gong, P., Li, S., Wang, Y., and Zhang, C. (2016). Gene cloning, expression and enzyme activity of Vitis vinifera vacuolar processing enzymes (VvVPEs). PloS one *11*, e0160945.

S21. Wang, L., Hu, X., Jiao, C., Li, Z., Fei, Z., Yan, X., Liu, C., Wang, Y., and Wang, X. (2016). Transcriptome analyses of seed development in grape hybrids reveals a possible mechanism influencing seed size. BMC genomics *17*, 1-15.

S22. Su, Z., Xuan, X., Sheng, Z., Wang, F., Zhang, X., Ye, D., Wang, X., Dong, T., Pei, D., and Zhang, P. (2022). Characterization and regulatory mechanism analysis of VvmiR156a–VvAGL80 pair during grapevine flowering and parthenocarpy process induced by gibberellin. The Plant Genome *15*, e20181.

S23. Li, Y., Zhang, S., Dong, R., Wang, L., Yao, J., van Nocker, S., and Wang, X. (2019). The grapevine homeobox gene VvHB58 influences seed and fruit development through multiple hormonal signaling pathways. BMC plant biology *19*, 1-18.

S24. Ahmad, B., Yao, J., Zhang, S., Li, X., Zhang, X., Yadav, V., and Wang, X. (2020). Genome-wide characterization and expression profiling of GASA genes during different stages of seed development in grapevine (Vitis vinifera L.) predict their involvement in seed development. International journal of molecular sciences *21*, 1088.

S25. Cui, M., Wang, W., Guo, F., Fan, X., Guan, L., Zheng, T., Zhu, X., Jia, H., Fang, J., and Wang, C. (2020). Characterization and temporal–spatial expression analysis of LEC1 gene in the

development of seedless berries in grape induced by gibberellin. Plant Growth Regulation *90*, 585-596.

S26. Upadhyay, A., Jogaiah, S., Maske, S., Kadoo, N., and Gupta, V. (2015). Expression of stable reference genes and SPINDLY gene in response to gibberellic acid application at different stages of grapevine development. Biologia Plantarum *59*, 436-444.

S27. Tang, Y., Liu, B., Li, Y., Van Nocker, S., Wang, Y., and Zhang, C. (2020). Differential expression of the seed-specific gene ABCG20 between seedless and seeded grapes and its roles in tomato seed development. South African journal of botany *131*, 428-436.

S28. Ahmad, B., Zhang, S., Yao, J., Chai, S., Yadav, V., Rahman, M.U., Wang, L., and Wang, X. (2022). Ectopic expression of VvFUS3, B3-domain transcription factor, in tomato influences seed development via affecting endoreduplication and hormones. Horticultural Plant Journal *8*, 351-360.

S29. Ocarez, N., Jiménez, N., Núñez, R., Perniola, R., Marsico, A.D., Cardone, M.F., Bergamini, C., and Mejía, N. (2020). Unraveling the deep genetic architecture for seedlessness in grapevine and the development and validation of a new set of markers for VviAGL11-based gene-assisted selection. Genes *11*, 151.

S30. Mejia, N., Soto, B., Guerrero, M., Casanueva, X., Houel, C., de los Angeles Miccono, M., Ramos, R., Le Cunff, L., Boursiquot, J.-M., and Hinrichsen, P. (2011). Molecular, genetic and transcriptional evidence for a role of VvAGL11 in stenospermocarpic seedlessness in grapevine. BMC plant biology *11*, 1-19.

S31. Lahogue, F., This, P., and Bouquet, A. (1998). Identification of a codominant scar marker linked to the seedlessness character in grapevine. Theoretical and Applied Genetics *97*, 950-959.

S32. Mejía, N., and Hinrichsen, P. (2002). A new, highly assertive SCAR marker potentially useful to assist selection for seedlessness in table grape breeding. pp. 559-564.

S33. Thomas, M., Edwards, K., and Pellerone, F. (2001). Grapevine microsatellite repeats: Isolation, characterisation and use ofr genotyping of grape germplasm from Southern Italy. Vitis: Journal of Grapevine Research *40*, 179-186.

S34. Ma, Y., Feng, J., Liu, C., Fan, X., Sun, H., Jiang, J., and Zhang, Y. (2018). Development and application of SSR new molecular marker for seedless traits in grape. Scientia Agricultura Sinica *51*, 2622-2630.