Integrative genomics reveals the polygenic basis of seedlessness in grapevine

Xu Wang[1, 2,] †, Zhongjie Liu[1,] †, 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], Yiwen Wang[1], Yanling Peng[1], and Yongfeng Zhou[1, 4, *]

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

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

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

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

† These authors contributed equally to this work.

* *Corresponding authors:* *zhouyongfeng@caas.cn*

Abstract

Seedlessness is a crucial quality trait in table grape (*Vitis vinifera* L.) breeding. However, the development of seeds involved intricate regulations, while 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 four haplotype-resolved telomere-to-telomere (T2T) 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, tracing the origin of the seedlessness trait 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 novel candidate genes, such as three *11S 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 groups related to seed development. Machine learning based genomic selection achieved a remarkable 99% precision in predicting grapevine seedlessness. Our findings highlight the polygenic nature of seedless and provide novel candidate genes for molecular genetics and an effective prediction for seedlessness in grape genomic breeding.

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

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[6,7], and table grapes[8]. 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[9-11]. Although this process has already become a common practice for producing seedless grapes, it raises concerns about food safety and labor costs[12]. 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[13-16].

The development of various seed tissues in grapes involves intricate genomic regulations. Previous studies have identified specific genes associated with different tissues of seed development[17]. For instance, the formation of the seed coat (integument) has been affected by genes like *VviAGL11*[18-21], *VvMADS28*[22], *VviINO*[23,24], and *VvHB63*[25-27]. Nutrient storage in the endosperm is controlled by genes such as 7S and 11S globulin-like seed storage proteins[28-30], while normal embryo growth relies on several gibberellin (GA) genes[31,32]. Moreover, the growth of ovules (young seeds) is influenced by genes such as *VvMADS39*[33], *VvMADS45*[34], *VviABCG20*[35-37], *VvFUS3*[38], *VvNAC26*[39], *VvβVPE*[40,41], *VviASN1*[42]. 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 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[43]. Similarly, in the 'Muscat of Alexandria' and 'Crimson Seedless' progeny, six QTLs for SN and ten QTLs for SFW were detected[44]. 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[45]. However, the restricted genetic background in progenies and limited population samples hinders the investigation of the polygenic basis of seedlessness in grapes.

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

Results

Comparative genomics between seeded and seedless cultivars

To study the genetic basis of seedlessness, we generated haplotype-resolved T2T assemblies for two seedless cultivars: TS and BM (**Fig. 1b, c)**, utilizing high-depth PacBio HiFi sequencing (~120× coverage) and Hi-C sequencing (~116× coverage; **Extended Data Fig. 2c**). The evaluation of K-mer heterozygosity in the TS and BM genomes, based on HiFi data, measured 1.51% and 1.41%, respectively. The quality of these genomes meets the assessment standards of the T2T level[46], with all centromeres regions and mostly telomeres regions marked (**Fig. 1a and Supplementary Table 1, 3**). Statistical analysis of variants revealed that TS and BM, between their two haplotype genomes, harbor 5.35 Mb and 5.04 Mb of SNPs, 5.01 Mb and 4.54 Mb of insertions and deletions (InDels, < 50 bp), and 33.42 Mb and 31.87 Mb of SVs (≥ 50 bp), respectively (**Supplementary Table 4**). A unique heterozygous inversion region (PN_T2T, Chr15: 10.7-12.0 Mb) specific to the TS hap2 genome was detected when aligning the four haplotypic genomes to PN_T2T, and several genes were found near the inversion breakpoints, such as *AGAMOUS* (*VvAG2*), *AGAMOUS-LIKE 62* (*VvAGL62*), *OIL BODY-ASSOCIATED PROTEIN 2B* (*VvOBAP2B*), and *GDSL esterase/lipase At1g29670*, which are involved in stamen and carpel determining, early endosperm development, and oil body synthesis (**Extended Data Fig. 3b and Supplementary Table 5**). The differential chromosomes between the two haplotype genomes explains the polymorphism of alleles and the variation in the number of genes (**Supplementary Table 1**).

To further investigate the SVs associated with seedless and seeded grapes, we aligned a total of 15 genomes, including the five seedless genomes and ten seeded genomes, to the PN_T2T (**Extended Data Fig. 3a**). We detected a heterozygous inversion (PN_T2T, Chr10: 23.8-25.4 Mb) in seedless grape varieties (**Fig. 1d**). The authenticity of these inversions was confirmed by Hi-C heatmaps and IGV[47] (**Fig. 1e**

and Extended Data Fig. 4). 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 (**Fig. 1f and Supplementary Table 7-8**). 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. *TT16/ABS* controls the formation of the maternal-derived endothelial cells by interacting with *AGL11/ SEEDSTICK* (*STK*) in the previous studies[48-51]. *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 (**Supplementary Table 6-7**). These findings suggest that the power of comparative genomics of haplotype-resolved T2T genomes in uncovering overlooked new candidate genes underlying crucial agronomic traits.

Introgression rather than convergent evolution underlying the evolvement 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 (**Supplementary Table 8)**. A total of 4,462,797 SNPs, 443,812 InDels, 487,204 SVs were identified by aligning WGS data to ‘Cabernet Sauvignon’ (CS) genome[52]. 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*, n = 92), and outgroup grapes (OG, n = 7; **Fig. 2c and Extended Data Fig. 5**). The results showed two independent lineages of seedless grapes nested in the VV×VL and VV branches, respectively, which is also supported by PCA (**Fig. 2b, c**). 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[53], following previous studies[54,55]. Interestingly, we detected significant genomic signals of introgression at seedless associated locus (see the GWAS 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* (SDInew, 30.36-31.86 Mb) locus[20] and the newly detected QTL on Chr07 (SDI2, 8.85-8.86 Mb; **Fig. 3c, d**). 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 Identity-by-Descent (IBD) analysis (**Supplementary Table 8**). The result revealed that TS group and BM group serve as bridges for gene flow between VV and VV×VL clusters (**Fig. 2d, e**). In fact, BM grape traces its ancestry back to 'Sultania' (TS) and 'Ichkimar'[56], with an IBD score of 0.50 supporting this observation (**Supplementary Table 9**). 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 (**Supplementary Table 9**). 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[8,57]. Furthermore, the seed abortion could also be caused by cytoplasmic male sterility (CMS)[58]. We analyzed the chloroplast and mitochondrial genomic variation and found nuclear inheritance, rather than CMS inheritance, played a crucial role in controlling seed abortion (**Extended Data Fig. 6).**

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

Genome-wide Association Study 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; **Supplementary Table 8**). 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 (**Extended Data Fig. 7 and Supplementary Table 10, 11**). 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 enriched in embryo development ending in seed dormancy and xylan metabolic process ($P < 0.05$; **Extended Data Fig. 8 and Supplementary Table 12**).

Remarkably, two consensus regions exhibited high consistence in three populations: Chr07: 8.85-8.86 Mb and Chr18: 29.40-35.54 Mb (**Fig. 3a, b**). 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 (**Fig. 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] (**Supplementary Table 11**). Additionally, within the 50 kb upstream region of SDI2, we identified a closely linked cluster of three tandem-duplicated genes, *11S GLOBULIN SEED STORAGE PROTEIN*[28-30,60] (**Fig. 3c**), and designated them as *11S globulin G1*, *G2*, and *G3* based on their genomic positions. Notably, two nonsynonymous mutations and one deletion related to seedlessness were detected across 14 grape genomes (**Fig. 4a and Extended Data Fig. 9**). Among them, both the heterozygous Asp-to-Val and Leu-to-Val mutations were specific in seedless grapes, except for the somatic mutations of ‘Black Corinth’ (BC) seeded grapes. However, the heterozygous deletion of the 18 amino acids in *11S globulin G3* was only detected in seedless grapes (**Fig. 4a**). The relative expression values of these genes showed a significant correlation with seed phenotypes, especially from 40-50 days after flowering (DAF; **Fig. 4b**).

Another consensus region is located on the Chr18: 29.40-35.54 Mb, encompassing 17 QTLs and the reported SDI locus (Chr18: 29.83-31.34 Mb; **Fig. 3d**). Due to the narrow genetic background of the samples used in the previous study[20], we redefined the SDI locus (SDInew, Chr18: 30.36-31.86 Mb) through GWAS analysis, revealing a total of eight QTLs (**Fig. 3d and Extended Data Fig. 7**). In this region, the population differentiation was lower than the genomic background between VV (n = 35) and VV×VL (n = 11) seedless grapes, as showed by the fixation indices (F_{ST}) results, suggesting a relatively close genetic distance between the two populations (SDInew F_{ST} = 0.073 vs. genome-wide F_{ST} = 0.126; **Supplementary Table 13**). This finding was also supported by genetic diversity (π) statistics. The two seedless populations showed similar genetic diversity on the SDInew locus (**Fig. 3d**). The f_d statistics[53] revealed numerous sites showing evidence of introgressions (f_d = 0.266, P = 1.28e-39), including numerous genes related to seedlessness, such as *CYP716A94*, *CYP716A17*, *VviAGL11*, etc. (**Fig. 3d**). Notably, several SNPs and InDels were highly associated with these candidate genes, especially in the promoter and coding sequence region (**Fig. 4c, d and Extended Data Fig. 10**), while several published molecular markers for seedlessness prediction, including e7_VviAGL11[44], 5U_VviAGL11[44], P3_VviAGL11[18], and VMC7f2[61], showed low predictive accuracy due to the absence of significant genotyping quality (**Fig. 4d and Supplementary Table 14**). We selected the top 10% of associated sites based on -$\log_{10}[P]$ values within the SDInew locus (**Fig. 4e**), and the most promising site for seedlessness prediction is Chr18_30874059, exhibiting a predicted accuracy of 97.8% for seedless grapes and 94.22% for seeded grapes in natural population.

Interestingly, we also detected two genomic regions for specific to each population (**Extended Data Fig. 7**). In the VV population, a specific region Chr01: 17.85-20.42 Mb harbored 13 QTLs and 67 genes, including seven primary genes involved in seed development, such as *NON-SPECIFIC LIPID TRANSFER PROTEIN GPI-ANCHORED 1* (*LTGP1*), *ARABIDOPSIS HISTIDINE KINASE 3* (*AHK3*), *B3 DOMAIN-CONTAINING TRANSCRIPTION FACTOR FUS3* (*FUS3*), *XYLOGLUCAN ENDOTRANSGLUCOSYLASE PROTEIN* (*XTH*), *11-BETA-HYDROXYSTEROID DEHYDROGENASE B* (*SOP3*), as well as previously reported *VvMADS4*[62] and *VvARF2-1*[63] (**Fig. 3a**). In the VV×VL population, a specific region Chr18: 15.14-20.57 Mb harbored eight QTLs and 33 associated genes. Among these genes, *SERINE DECARBOXYLASE 1* (*SDC1*), *ABC TRANSPORTER G FAMILY MEMBER 22 ABCG22* or *VvPNWBC22.2*(TANG, 2018 #20), *SUS2*, *LACCASE-14* (*LAC14*) and *TT10*/*LAC15* were highly associated with seed development (**Fig. 3a**). Our results suggest that seed abortion can be regulated by the collaborative effects of multiple genes, and the mapped candidate genes and variable sites hold valuable potential applications in seedless grapes breeding.

Integrative Genomic Analysis Identified 339 Seedless Candidate Genes

To elucidate the polygenic basis of seed abortion, we further utilized an integrative genomic analysis using transcriptomic analyses, seed development associated GO term genes, previously reported family genes and molecular markers, and GWAS candidate genes, to identify the core candidate genes associated with seed abortion (**Supplementary Table 17**). 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 (**Supplementary Table 15**). For the 'Italia' and 'Hongju Seedless' (HS) groups, we identified a total of 2,680 significantly upregulated genes and 1,835 significantly downregulated genes from the six time points (**Extended Data Fig. 13c**). Similarly, in the 'Pinot Noir' (PN) and TS groups, we identified a total of 3,969 significantly upregulated genes and 2,695 significantly downregulated genes (**Extended Data Fig. 13c**). 'Himrod Seedless' (Himrod) and 'Jinzao Wuhe' (Jinzao) were used serve as a control for cross-validation during four fruit development stage. Interestingly, we found *VviAGL11* was exclusively in the downregulated DEGs in the PN and TS comparison, but not in the 'Italia' and HS comparison (**Extended Data Fig. 13a, b**). In addition, 1,301 core upregulated genes and 616 core 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. (**Fig. 5a and Supplementary Table 11**).

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 (**Extended Data Fig. 12a, b**). This yielded 6,529 homologous genes (see Methods), with 5,061 genes only present in the GO pathway, including the *TT16/FBP24* gene identified in the comparative genomics (**Fig. 5a**). Notably, the intersection between the RNA-seq related genes and the GO homologous genes revealed 163 downregulated genes and 294 upregulated genes (**Extended Data Fig. 13b**), such as *LTP*, *11S GLOBULIN*, *OLEOSIN*, *CELLULOSE SYNTHASE A CATALYTIC SUBUNIT* (*CESA4*, *CESA7* and

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

Through integrative genomic analysis, we screened 339 core candidate genes, categorized into 13 groups, by condition-based filtering that exhibited significant differential expression between seedless and seeded grape cultivars (see Methods, **Supplementary Table 11**). 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 (**Fig. 5b**). Nutrient deficiency could be primary factor leading to seed abortion in later embryo development; Secondly, genes involved in the regulation of lignin and cellulose synthesis/degradation in seed coat exhibit higher activity levels in seeded grapes (**Fig. 5c**); Thirdly, candidate genes manage the synthesis and transport of oil bodies ensuring efficient lipid accumulation and utilization during seed development (**Fig. 5d**). Overall, our findings indicate that multiple genes have an accumulative effect in the process of seed abortion, resulting varying degrees of seedlessness[67]. 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 from GWAS analysis, including 77 InDels and 717 SNPs (**Supplementary Table 18**). Using these variations, an unrooted phylogenetic tree was constructed based on admixed populations (n = 444), revealing that the majority of seedless individuals clustered together (**Fig. 6a**). However, some seeded samples, like seeded hybrid progeny[20] and Rizamat[44] 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; **Fig. 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 (**Extended Data Fig. 14**). We used 794 variant sites and phenotypic data from the admixed populations (444 samples) as the training dataset, and evaluated the ability of nine different classical models to predict the seedlessness trait based on 100 rounds of random cross-validations (**Fig. 6b**). Among these models, machine learning based methods, including SVR-poly and ElasticNetCV, demonstrated a strong performance in predicting the phenotype, yielding prediction accuracies of 85.36% and 84.57%, respectively (**Fig. 6b**). As a result, we applied SVR-poly and ElasticNetCV to genomic prediction on the testing set data (39 samples not included in the model building) (**Supplementary Table 8**). We observed that the SVR-poly model and the ElasticNetCV model yielded high levels of accuracy, with correlation coefficient R values of 0.99 (P < 2.2e-16) and 0.96 (P < 2.2e-16), respectively (**Fig. 6c**), suggesting the efficiency of machine learning based genomic selection of seedless in grapes.

Discussion

Seedless is an important quality trait in table grape breeding. Previous genetic investigated the functions of multiple genes, including *VvAGL11*, however, quantitative genetic analyses of natural population 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 were built to accurately predict the seedless phenotypes. Importantly, these findings could efficiently save the cost and time in table grape breeding.

The polygenic nature for seedlessness in grapes

The occurrence of seedlessness phenotypes from a cumulative polygenic effect associated with different tissues and development stages[17]. 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 (**Fig. 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[17,50]. 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 **(Fig. 3a, b and Supplementary Table 11**). Additionally, three transcriptomic analyses, GO homologous genes, and previously reported family genes provide us with numerous significant candidate genes (**Fig. 5a and Supplementary Table 11**).

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 (**Fig. 5a and Supplementary Table 17**). Most of these genes are associated with three main tissues related to seed coat, endosperm, and embryo, as shown in **Fig. 5b-d**, and ten other development processes (**Supplementary Table 11**). For example, we identified numerous candidate genes involving in seed hormone regulation, such as *MOTHER OF FT AND TFL1* (*MFT*) in ABA and GA pathways, *VvABI3-1*[68] in ABA pathway, *VvGH3.9*[69] in auxin pathways, and *VvMJE1*[64] 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*[70]. Except for lipid accumulation (**Fig. 5d**), the processes of sugar and amino acid synthesis also contribute to seed development, such as *BGLU15*, *VviGAPDH*[71], 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)[44,72-74] and marker set for genomic selection generated in this study could be efficiently used in table grape breeding (**Fig. 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[75], e7_VviAGL11[44], 5U_VviAGL11[44], P3_VviAGL11[18], VMC7f2[61], and VrSD10[76] (**Fig. 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 (**Fig. 4e**).

Quantitative genetics analysis revealed 110 high-quality QTLs associated with seedlessness in grapevines, including 634 candidate genes (**Supplementary Table 11**). Through extracting GWAS significant variants from admixed population, we obtained detail information on all 794 significant variant sites for training models. Genome selections, employing machine learning algorithms, achieved an impressive precision of 99%. This approach could facilitate early genomic selection of natural germplasms and hybrid progeny. Many crops, such as tomato[77], potato[78,79], cereal[80], rice[81,82], wheat and maize[83,84], have successfully applied genomic selection and prediction in their breeding programs. However, genomic selection has rarely been used on grape breeding. In the future, numerous agronomic traits like fruit aroma, disease resistance, soluble solids content, and so on, could be integrated into a single GS chip, offering a powerful genomic tool for genomic design of grapevine breeding.

Methods

Plant materials and genome sequencing

To enrich the genetic diversity of seedless grape varieties, we collected fresh tissues of two seedless grape varieties, ‘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. Genomic DNA was extracted from grape leaves using CTAB method, followed by purification with the QIAGEN Genomic kit (CAT#13343). For each sample, a total of 15 μg DNA was utilized for HiFi (SMRTbell) library preparation. The sheared DNA fragments (gTUBEs, Covaris, USA) underwent treatment with the SMRTbell Enzyme Cleanup Kit (Pacific Biosciences, CA, USA) 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 in Grandomics, 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 (Illuminia, USA) 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 **Extended Data Fig. 1**. The full pipeline for genome assembly and gap filling can be found on our lab GitHub@zhouyflab (see Code availability). In brief, we utilized the Hi-C and HiFi data integrated assembly algorithm to generate contig reads by Hifiasm (v. 0.16.1-r375)[85]. The contig reads were oriented and ordered to scaffold level using RagTag (v. 2.1.0)[86] with default parameters. Hi-C reads were also employed to anchor scaffolds onto chromosomes by Juicer 2.0[87] and Juicebox (v. 1.11.08)[88]. The adjusted genome at the chromosome level was generated using 3D-DNA (v. 201008)[89]. The detailed workflow of haplotype-resolved genome assembly and annotation is described in **Extended Data Fig. 1**. 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)[90] 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)[91] with HiFi long reads. Furthermore, using PN_T2T genome[46], we utilized RagTag to assemble three chromosomes-level genomes based on scaffold level: 'Cabernet Sauvignon' (CS)[52], 'Black Corinth Seedless' (BC seedless)[92], and 'Black Corinth Seeded' (BC seeded)[92].

Assembly assessment

HiFi data from TS and BM were performed to assess genome heterozygosity based on k-mer using GenomeScope2.0[93]. Meanwhile, basic genome statistics were calculated using seqkit (v. 2.2.0)[94], which included genome length, N50, and GC content. The completeness of the haplotype genomes was evaluated using BUSCOs (v. 5.3.0)[95] with the embryophyta_odb10 database. Merqury (v. 1.3, best_k=19)[96] was employed to evaluate the quality value (QV) and completeness of the haplotype genomes based on whole-genome sequencing data. The Hi-C interactive signals on grape genome were visualized using Juicebox.

Genome annotation

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

Comparative genomics

For genome level variant calling, we selected a total of 15 grape genomes, including ten seeded accessions: ‘PN40024’ (PN_T2T), ‘Cabernet Sauvignon’ (CS, hap1 and hap2)[92], ‘Muscat Hamburg’ (MH, hap1 and hap2), ‘Shine Muscat’ (SM, hap1 and hap2), ‘Muscadinia rotundifolia’ (MR, hap1 and hap2) [99], BC Seeded[92], as well as five seedless: TS (hap1 and hap2), BM (hap1 and hap2), BC Seedless[92]. 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.

All chromosome-level genomes were aligned with PN_T2T genome using Mummer4 (v. 4.0.0rc1)[100], and the results were visualized using Plotsr (v. 0.5.4)[101] 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)[102] 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)[103] 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)[104] with default parameters. The filtered data were then mapped to the CS reference genome using BWA (v. 0.7.17-r1188)[105]. Non-uniquely mapped and duplicated reads were excluded using SAMtools and GATK (v.4.2.3.0)[106]. 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)[107] was used to SVs calling with default parameters. Basic filtering of the VCF file was performed using VCFtools (v. 0.1.16)[108] 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)[109] 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, 38,267 variation sites (SNPs and InDels) from 314 grape accessions in the MT genome, and 2,247 variation sites (SNPs and InDels) from 314 grape accessions in the Pltd genome.

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)[110] with parameters: -m GTR+I+G -bb 1000 -bnni -alrt 1000 -st DNA, and the result were visualized using iTOLs[111]. Similarly, the phylogenetic analysis of the MT and PT 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)[112], 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 populations[113]. 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 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. GWAS analysis utilized the mixed linear model in GEMMA (v. 0.98.3)[114] 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/Variant Numbers).

GO enrichment analysis was employed using the web tool DAVID[115], and the results were visualized with R script. The SDI locus was identified through molecular markers SNP-25.24 and SNP-26.93[20]. The LD linkage heatmap was visualized using LDBlockshow (v. 1.40)[116]. Moreover, DIAMOND (v. 2.0.15)[117] was employed for sequence comparisons of whole-genome homologous protein from 14 genomes (Seeded: Seedless: Outgroup = 8:5:1). Sequence alignments of candidate proteins were conducted using ClustalW in MEGA11 [118], and sequence visualization was accomplished using GeneDoc (v. gd322700)[119]. The visualization of FPKM values, gene structures, and mutation ratios was achieved using R scripts (see Code availability).

Multi-transcriptome analysis

Three independent transcriptome datasets were downloaded from the NCBI database (see **Supplementary Table 15**). The 76 samples encompassed six grape varieties: PN and TS, which underwent repeated testing from 20 DAF to 50 DAF over a two-year period; 'Italia' and HS, continuously evaluated from 7 to 42 DAF within a single year; and Himrod (VV×VL population) and Jinzao (VV population), continuously monitored from full flowering stage to grape maturity within a single year. These raw fastq data underwent quality control and data cleaning using Fastp with default parameters. Using the PN_T2T genome as a reference, transcriptomic assembly was conducted on the processed data using STAR (v. 1.5.2)[120]. The R package DESeq2 (v.1.36.0)[121] was utilized for PAC analysis and pairwise comparisons between seeded and seedless grape samples at each development time point. The thresholds for differential expression genes were |Log2FoldChange| ≥ 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 ≥ 50%, e-value ≤ 1e-5). Additionally, the primer sequences collected from previous studies in the past decade, including 451 genes and 7 molecular markers associated with seed abortion processes. All of them were then mapped to the PN_T2T genome based on sequence alignment using BALST in TBtools (v. 1_113)[122]. The results were visualized using Venn diagrams and whole-genome density plots using the R package ggvenn (v0.1.10) and RIdeogram[123], respectively. The integrated results in this study, including GWAS results, RNA-seq results, GO enrichment results, and previously reported family genes, were overlapped and filtered manually. The core candidate genes were screened based on an average expression value across all time points (AVE FPKM ≥ 100) and Fold-Changes (Fold = Seeded AVE FPKM / Seedless AVE FPKM, Fold ≤ 0.5 or Fold ≥ 2.0). Data visualization was performed using an R script, which included UpSet plot analysis and gene expression heatmaps (see Code availability).

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 **Extended Data Fig. 14**. Beagle (v. 5.2_21Apr21.304.jar)[124] was used to impute the VCF file. Subsequently, the 794 variants, above the significance threshold of $-\log_{10}(0.05/\text{Variant Numbers})$ (~ 7.66), were extracted from imputed VCF using BCFtools (v. 1.13)[125], including 717 SNPs and 77 InDels. 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 rounds of random cross-validation with the training set, we chose the models with best performance, SVR-poly and ElasticNetCV, for the prediction of testing set. 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.

Figure Legends

Fig. 1 | Comparative genomic 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.

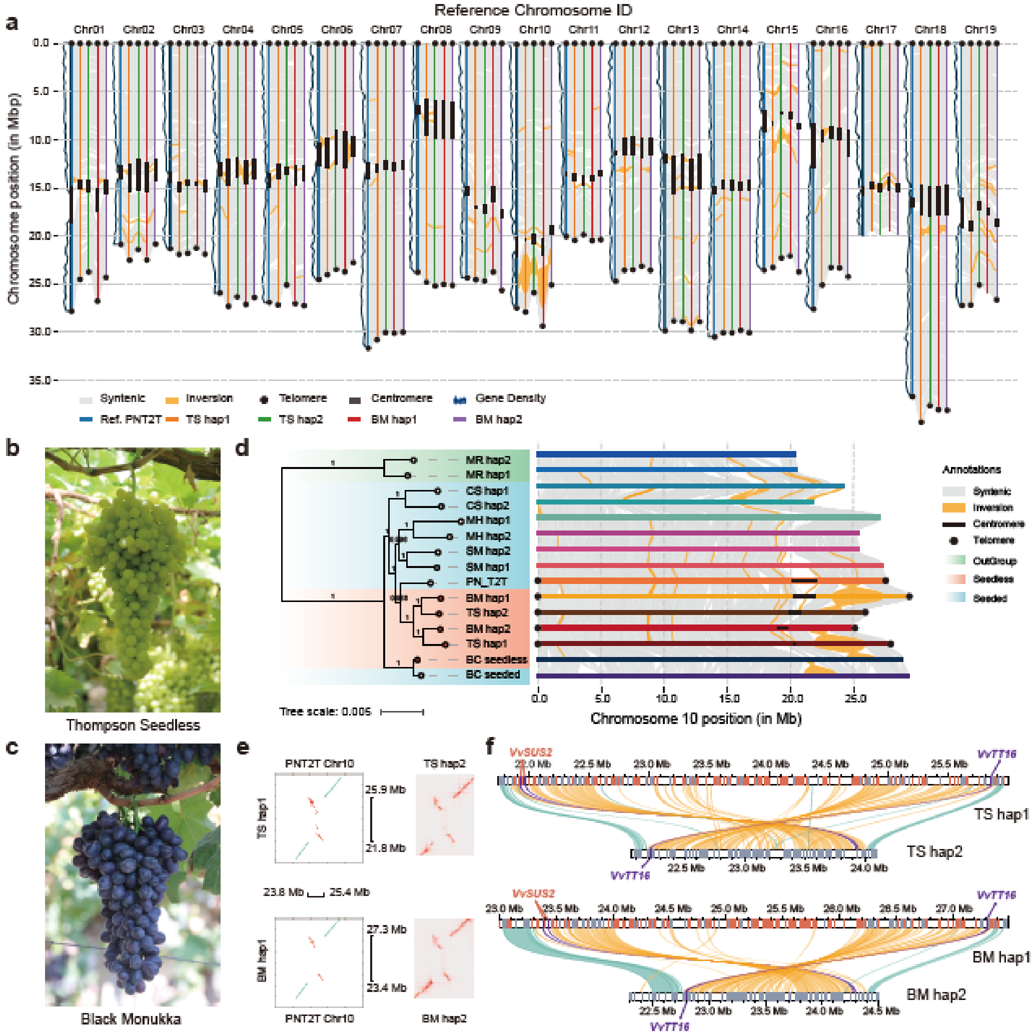
Fig. 2 | Evolutionary genomics of seedlessness in grapevine. a, Comparison of grape clusters and cross-sectional views of berry from three cultivated grape varieties: TS, BM, and CS. The scale bar indicates 5 cm for grape clusters and 1 cm for individual berries. **b**, PCA analysis on the whole-genome 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 Extended Data Fig. 6 for detailed information of the phylogenetic tree). **d**, IBD analysis of 46 seedless grapes, filtering the results with a threshold of 0.40. 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.

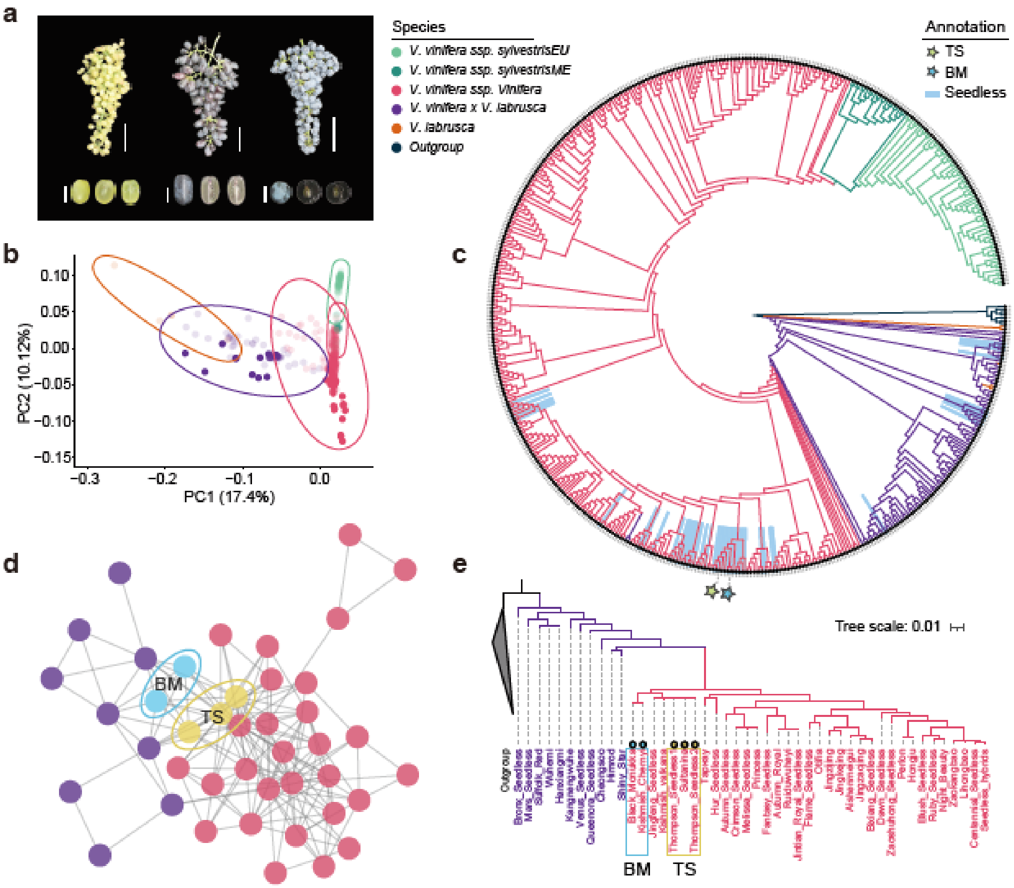
Fig. 3 | GWAS mapping of the polygenic basis of seedlessness. a and **b**, Seedless-associated genomic loci and genes across three populations: the VV population (in red, n=352), the VL population (in purple, n=92), and the admixed population (in yellow and dark-blue, n = 444). The horizontal dashed 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 grey 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.

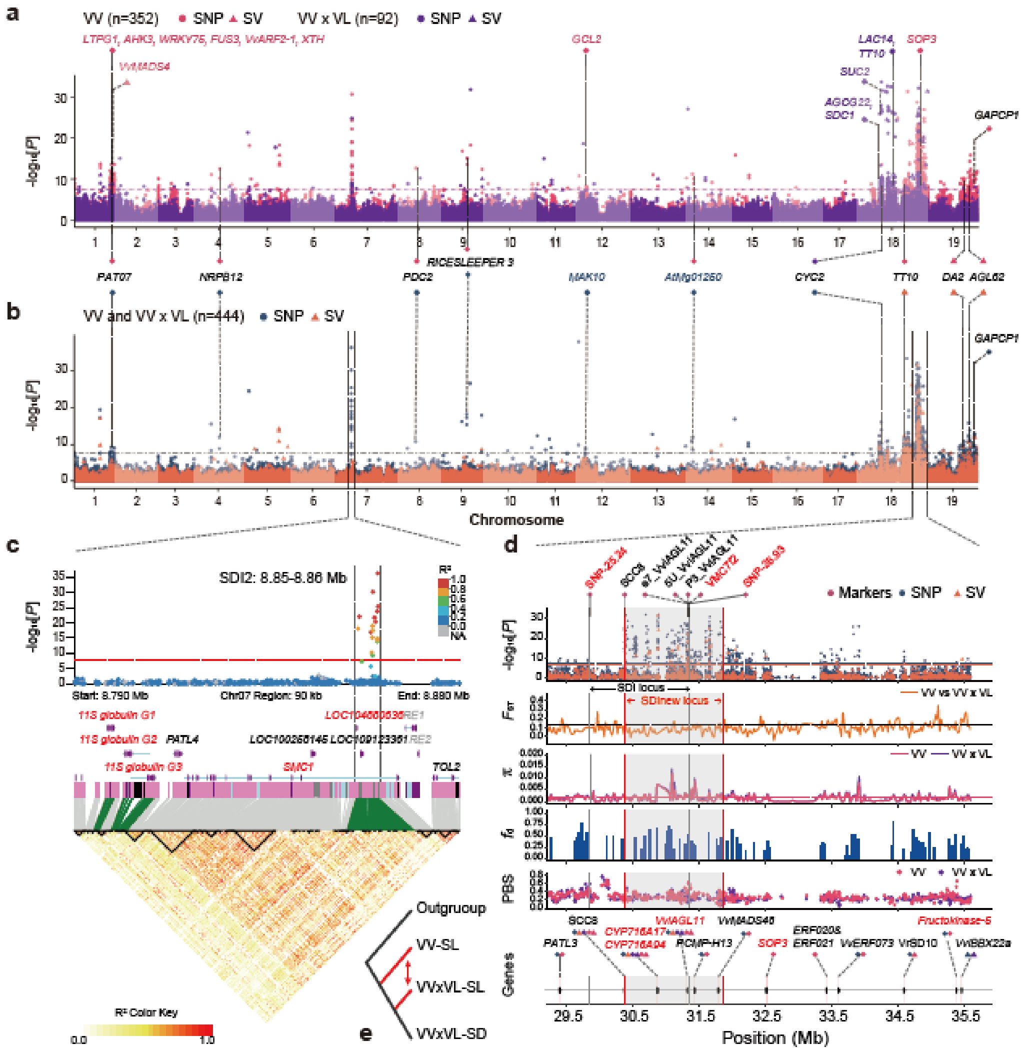
Fig. 4 | Deep mining of key loci associated with seed abortion. a, Sequence alignment of three *11S GLOBULIN SEED STORAGE PROTEIN* genes in Chr07 across 14 grape varieties. Black blocks represent deletion and nonsynonymous mutations. **b**, Relative expression values of these three genes at different time points across six grape varieties. 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, ‘Seeded’ includes the case with genotype 0/0.

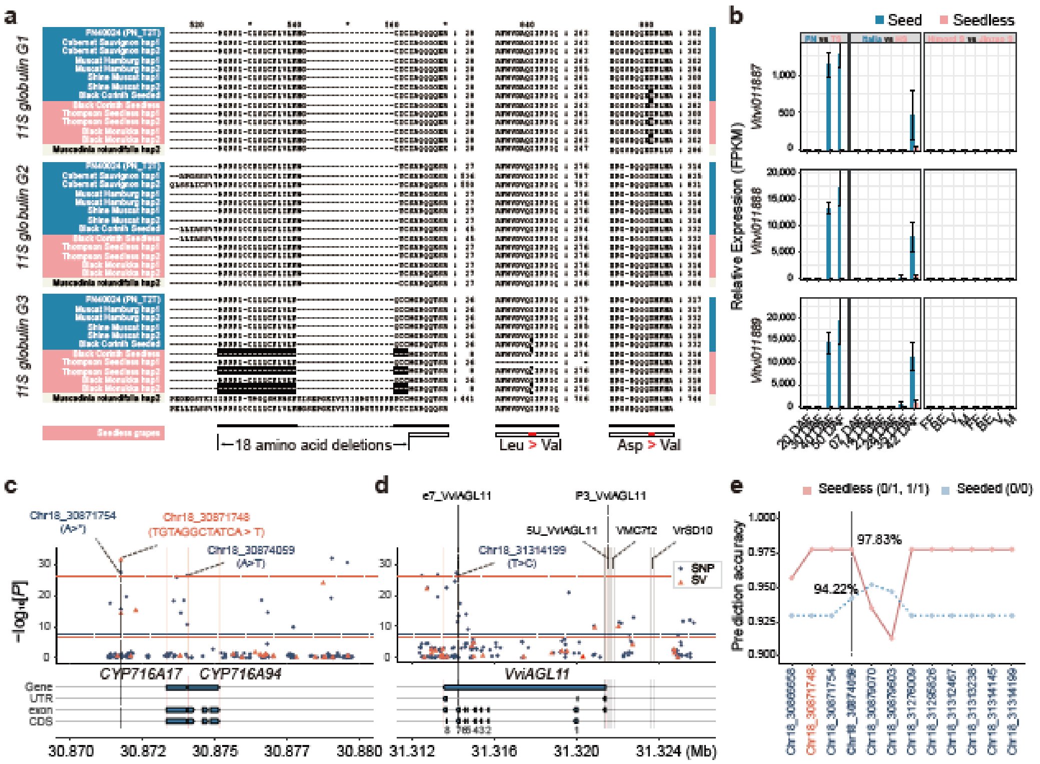
Fig. 5 | Integrative genomic analyses for grapevine seed abortion. a, Results from integrative genomic analyses: GWAS, transcriptomic analysis, reported genes and markers mapping, and GO homologous genes overlapping. Blue bars represent the number of genes uniquely and overlappedly through this approach. **b** and **c**, Heatmap of Log_2(FPKM) for candidate genes related to embryo, endosperm, and seed coat development, resulting from integrative genomic analyses. 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).

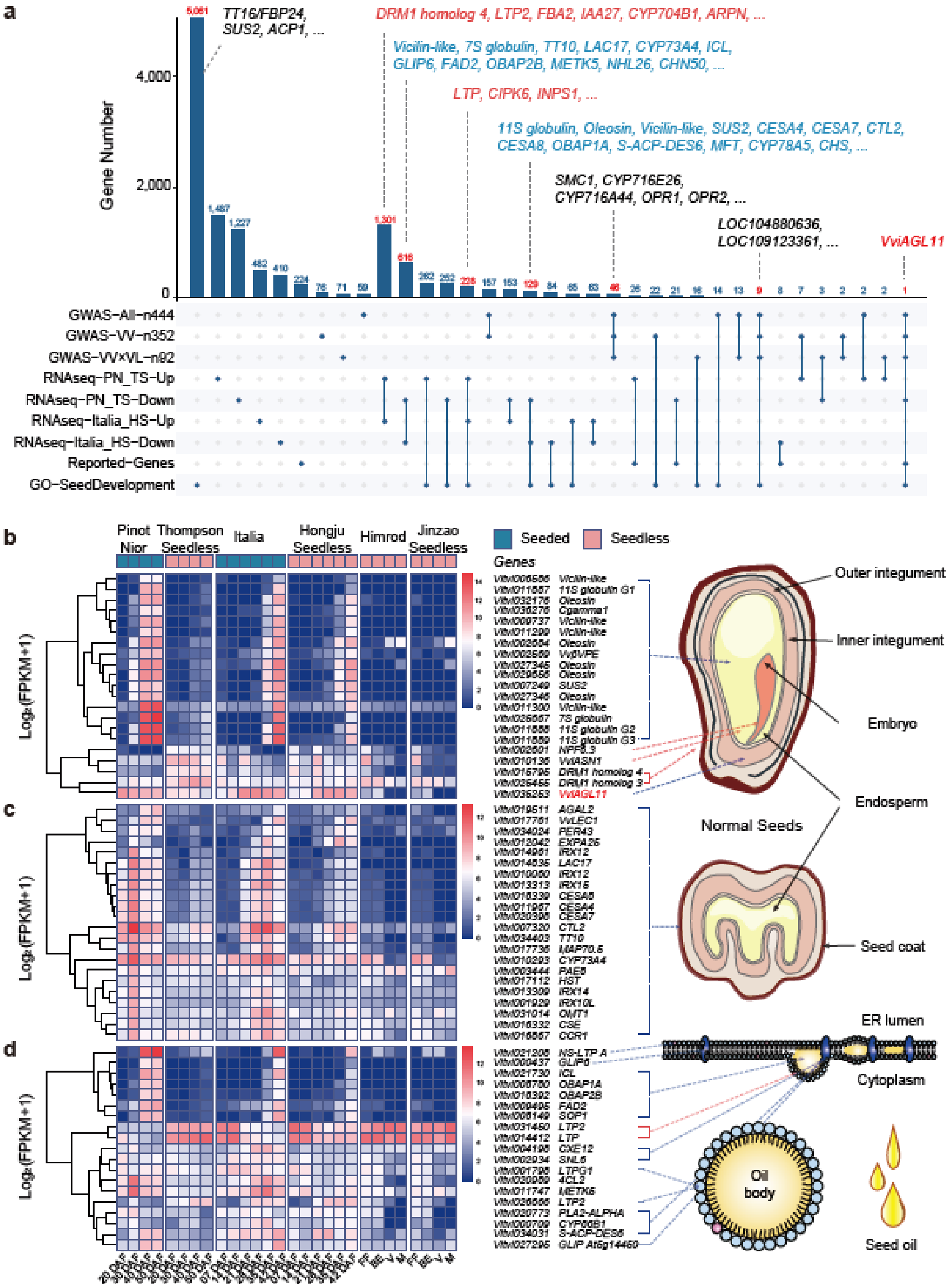
Fig. 6 | Machine learning based genomic selection on seedlessness in grapevine breeding. a, Phylogenetic clustering based on 794 significant variants (77 InDels and 717 SNPs) derived from the GWAS analysis. **b**, Comparison of seedlessness prediction accuracy across nine classical models for genome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

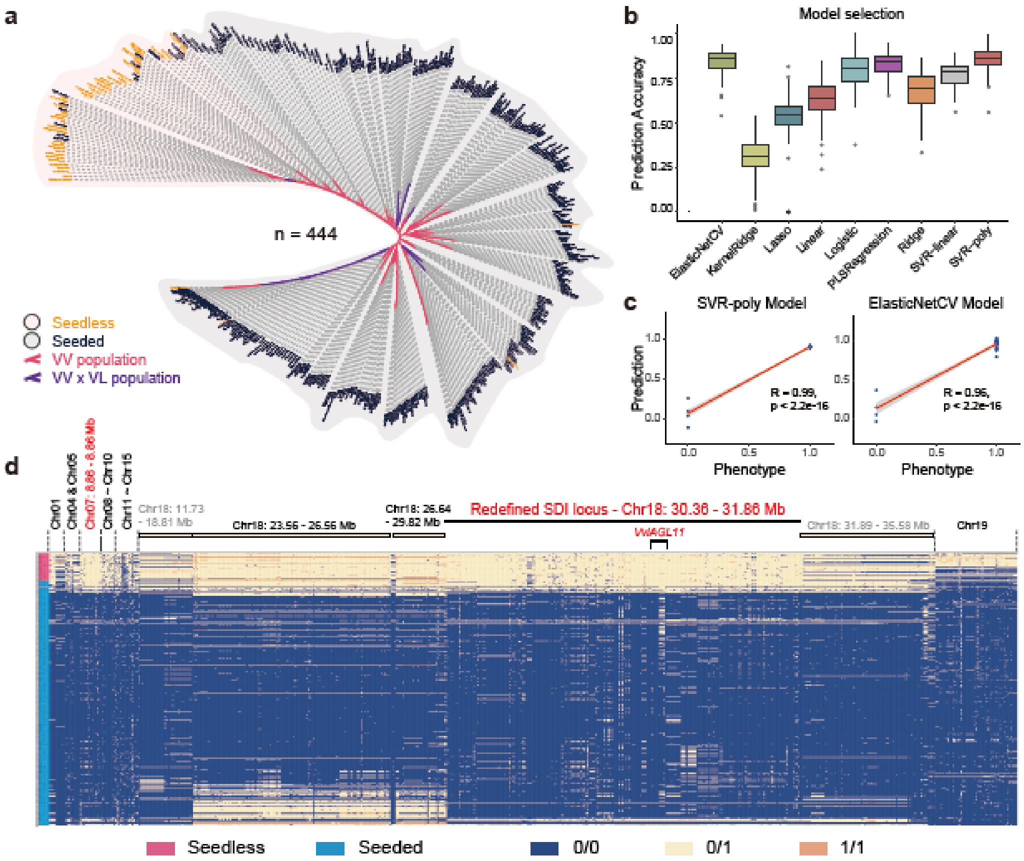

a
Reference Chromosome ID
Chr01 Chr02 Chr03 Chr04 Chr05 Chr06 Chr07 Chr08 Chr09 Chr10 Chr11 Chr12 Chr13 Chr14 Chr15 Chr16 Chr17 Chr18 Chr19
Chromosome position (in Mbp)
0.0
5.0
10.0
15.0
20.0
25.0
30.0
35.0
Syntenic
Inversion
Telomere
Centromere
Gene Density
Ref. PNT2T
TS hap1
TS hap2
BM hap1
BM hap2
b
Thompson Seedless
c
Black Monukka
d
MR hap2
MR hap1
CS hap1
CS hap2
MH hap1
MH hap2
SM hap2
SM hap1
PN_T2T
BM hap1
TS hap2
BM hap2
TS hap1
BC seedless
BC seeded
Tree scale: 0.005
0.0 5.0 10.0 15.0 20.0 25.0
Chromosome 10 position (in Mb)
Annotations
Syntenic
Inversion
Centromere
Telomere
OutGroup
Seedless
Seeded
e
PNT2T Chr10
TS hap2
TS hap1
25.9 Mb
21.8 Mb
23.8 Mb
25.4 Mb
BM hap1
27.3 Mb
23.4 Mb
PNT2T Chr10
BM hap2
f
VvSUS2
VvTT16
22.0 Mb 22.5 Mb 23.0 Mb 23.5 Mb 24.0 Mb 24.5 Mb 25.0 Mb 25.5 Mb
TS hap1
TS hap2
22.5 Mb 23.0 Mb 23.5 Mb 24.0 Mb
VvTT16
VvSUS2
VvTT16
23.0 Mb 23.5 Mb 24.0 Mb 24.5 Mb 25.0 Mb 25.5 Mb 26.0 Mb 26.5 Mb 27.0 Mb
BM hap1
BM hap2
22.5 Mb 23.0 Mb 23.5 Mb 24.0 Mb 24.5 Mb
VvTT16


a
Species
● V. vinifera ssp. sylvestrisEU
● V. vinifera ssp. sylvestrisME
● V. vinifera ssp. Vinifera
● V. vinifera x V. labrusca
● V. labrusca
● Outgroup
Annotation
☆ TS
☆ BM
Seedless
b
0.10
0.05
0.00
−0.05
−0.10
−0.15
PC2 (10.12%)
−0.3
−0.2
−0.1
0.0
PC1 (17.4%)
c
d
BM
TS
e
Tree scale: 0.01
Outgroup
Bronx_Seedless
Mars_Seedless
Suffolk_Red
Wuhemi
Hanxiangmi
Kangxiangwuhe
Venus_Seedless
Queens_Seedless
Chengzao
Himrod
Shine_Star
Black_Monukka
Kishmish_Chernyi
Jingfeng_Seedless
Kishmish_vatkana
Thompson_Seedless
Sultanina
Thompson_Seedless2
Tapsay
Hur_Seedless
Autumn_Seedless
Crimson_Seedless
Melissa_Seedless
Princess
Fantasy_Seedless
Autumn_Royal
Jintian_Royal_Seedless
Flame_Seedless
Otilia
Jingxing
Aishenmeigui
Jingzaojing
Bixiang_Seedless
Dawn_Seedless
Zaoheibao
Honglu
Blush_Seedless
Ruby_Seedless
Night_Beauty
Zaokangbao
Lihongbao
Centennial_Seedless
Seedless_hybrids
BM
TS


a
VV (n=352) ● SNP ▲ SV VV x VL (n=92) ● SNP ▲ SV
LTPG1, AHK3, WRKY75, FUS3, VvARF2-1, XTH
VvMADS4
GCL2
LAC14, TT10
SOP3
SUC2
AGOG22, SDC1
GAPCP1
-log10[P]
0
10
20
30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
RICESLEEPER 3
PAT07
NRPB12
PDC2
MAK10
AtMg01250
CYC2
TT10
DA2
AGL62
b
VV and VV x VL (n=444) ● SNP ▲ SV
GAPCP1
-log10[P]
Chromosome
c
SDI2: 8.85-8.86 Mb
R²
1.0
0.8
0.6
0.4
0.2
0.0
NA
-log10[P]
0 5 10 15 20 25 30 35
Start: 8.790 Mb
Chr07 Region: 90 kb
End: 8.880 Mb
11S globulin G1
LOC104880636
RE1
11S globulin G2
PATL4
LOC100268145
LOC109123361
RE2
11S globulin G3
SMC1
TOL2
R² Color Key
0.0
1.0
d
SNP-25.24
SCC8
e7_VvAGL11
5U_VvAGL11
P3_VvAGL11
VMC7f2
SNP-26.93
● Markers ● SNP ▲ SV
-log10[P]
SDI locus
SDInew locus
F_{ST}
VV vs VV x VL
π
VV
VV x VL
f_d
PBS
VV
VV x VL
SCC8
VvAGL11
VvMADS48
ERF0208
Fructokinase-5
CYP716A17
CYP716A94
PATL3
RCMP-H13
SOP3
ERF021
VvERF073
VrSD10
VvBBX22a
Genes
29.5 30.5 31.5 32.5 33.5 34.5 35.5
Position (Mb)
e
Outgruoup
VV-SL
VVxVL-SL
VVxVL-SD


a
11S globulin G1
11S globulin G2
11S globulin G3
PN40024 (PN_T2T)
Cabernet Sauvignon hap1
Cabernet Sauvignon hap2
Muscat Hamburg hap1
Muscat Hamburg hap2
Shine Muscat hap1
Shine Muscat hap2
Black Corinth Seeded
Black Corinth Seedless
Thompson Seedless hap1
Thompson Seedless hap2
Black Monukka hap1
Black Monukka hap2
Muscadinia rotundifolia hap2
Seedless grapes
18 amino acid deletions
Leu > Val
Asp > Val
b
Seed
Seedless
PN vs TS
Italia vs
Vitvi011687
Vitvi011688
Vitvi011689
Relative Expression (FPKM)
1,000
500
0
20,000
15,000
10,000
5,000
0
c
Chr18_30871754
(A>*)
Chr18_30871748
(TGTAGGCTATCA > T)
Chr18_30874059
(A>T)
−log₁₀[P]
30
20
10
0
CYP716A17
CYP716A94
Gene
UTR
exon
CDS
30.870
30.872
30.875
30.877
30.880
d
e7_VviAGL11
P3_VviAGL11
5U_VviAGL11
VMC7f2
VrSD10
Chr18_31314199
(T>C)
SNP
SV
VviAGL11
8 765 432 1
31.312
31.316
31.320
31.324 (Mb)
e
Seedless (0/1, 1/1)
Seeded (0/0)
Prediction accuracy
1.000
0.975
0.950
0.925
0.900
97.83%
94.22%
Chr18_30865658
Chr18_30871748
Chr18_30871754
Chr18_30874059
Chr18_30879070
Chr18_30879603
Chr18_31276009
Chr18_31295826
Chr18_31312467
Chr18_31313238
Chr18_31314145
Chr18_31314199


a
Gene Number
4,000
2,000
0
5,061
TT16/FBP24, SUS2, ACP1, ...
DRM1 homolog 4, LTP2, FBA2, IAA27, CYP704B1, ARPN, ...
Vicilin-like, 7S globulin, TT10, LAC17, CYP73A4, ICL, GLIP6, FAD2, OBAP2B, METK5, NHL26, CHN50, ...
LTP, CIPK6, INPS1, ...
11S globulin, Oleosin, Vicilin-like, SUS2, CESA4, CESA7, CTL2, CESA8, OBAP1A, S-ACP-DES6, MFT, CYP78A5, CHS, ...
SMC1, CYP716E26, CYP716A44, OPR1, OPR2, ...
LOC104880636, LOC109123361, ...
VviAGL11
1,487
1,227
482
410
224
76
71
59
1,301
616
262
252
228
157
153
129
84
65
63
46
26
22
21
16
14
13
9
8
7
3
2
2
2
1
GWAS−All−n444
GWAS−VV−n352
GWAS−VV×VL−n92
RNAseq−PN_TS−Up
RNAseq−PN_TS−Down
RNAseq−Italia_HS−Up
RNAseq−Italia_HS−Down
Reported−Genes
GO−SeedDevelopment
b
Pinot Nior
Thompson Seedless
Italia
Hongju Seedless
Himrod
Jinzao Seedless
Seeded
Seedless
Genes
Log2(FPKM+1)
Vitvi006586 Vicilin-like
Vitvi011557 11S globulin G1
Vitvi032176 Oleosin
Vitvi036276 Cgamma1
Vitvi009737 Vicilin-like
Vitvi011299 Vicilin-like
Vitvi002684 Oleosin
Vitvi002569 VvβVPE
Vitvi027345 Oleosin
Vitvi029656 Oleosin
Vitvi007249 SUS2
Vitvi027346 Oleosin
Vitvi011300 Vicilin-like
Vitvi026667 7S globulin
Vitvi011558 11S globulin G2
Vitvi011559 11S globulin G3
Vitvi002601 NPF8.3
Vitvi010136 VviASN1
Vitvi015795 DRM1 homolog 4
Vitvi026456 DRM1 homolog 3
Vitvi036253 VviAGL11
c
Vitvi019511 AGAL2
Vitvi017761 VvLEC1
Vitvi034024 PER43
Vitvi012042 EXPA26
Vitvi014961 IRX12
Vitvi014535 LAC17
Vitvi010060 IRX12
Vitvi013313 IRX15
Vitvi018339 CESA8
Vitvi011957 CESA4
Vitvi020398 CESA7
Vitvi007320 CTL2
Vitvi034403 TT10
Vitvi017736 MAP70.5
Vitvi010293 CYP73A4
Vitvi003444 PAE8
Vitvi017112 HST
Vitvi013309 IRX14
Vitvi001929 IRX10L
Vitvi031014 OMT1
Vitvi016332 CSE
Vitvi016667 CCR1
d
Vitvi021208 NS-LTP A
Vitvi000437 GLIP6
Vitvi021730 ICL
Vitvi006780 OBAP1A
Vitvi016392 OBAP2B
Vitvi009495 FAD2
Vitvi006149 SOP1
Vitvi031450 LTP2
Vitvi014412 LTP
Vitvi004195 CXE12
Vitvi002934 SNL6
Vitvi001795 LTPG1
Vitvi020959 4CL2
Vitvi011747 METK6
Vitvi020666 LTP2
Vitvi020773 PLA2-ALPHA
Vitvi000709 CYP86B1
Vitvi034031 S-ACP-DES6
Vitvi027295 GLIP At5g14460
14
12
10
8
6
4
2
0
PF
BE
V
M
Outer integument
Inner integument
Embryo
Endosperm
Normal Seeds
Seed coat
ER lumen
Cytoplasm
Oil body
Seed oil


a
n = 444
○ Seedless
○ Seeded
VV population
VV x VL population
b
Model selection
Prediction Accuracy
1.00
0.75
0.50
0.25
0.00
ElasticNetCV
KernelRidge
Lasso
Linear
Logistic
PLSRegression
Ridge
SVR-linear
SVR-poly
c
SVR-poly Model
ElasticNetCV Model
Prediction
1.0
0.5
0.0
R = 0.99, p < 2.2e-16
R = 0.96, p < 2.2e-16
0.0
0.5
1.0
Phenotype
d
Chr01
Chr04 & Chr05
Chr07: 8.86 - 8.86 Mb
Chr08 ~ Chr10
Chr11 ~ Chr15
Chr18: 11.73 - 18.81 Mb
Chr18: 23.56 - 26.56 Mb
Chr18: 26.64 - 29.82 Mb
Redefined SDI locus - Chr18: 30.36 - 31.86 Mb
VviAGL11
Chr18: 31.89 - 35.58 Mb
Chr19
Seedless
Seeded
0/0
0/1
1/1

Supplementary Data

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

Extended Data Fig. 2 | Evaluation of four haplotype-resolved genomes. **a**, BUSCO assessment of genome completeness 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. Most single contig (green rectangle) are directly composed of chromosomes (blue rectangles).

Extended Data Fig. 3 | Comparative genomics results. a, Sequence alignment of 15 grape genomes with the PN_T2T genome. 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.

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

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

Extended Data Fig. 6 | Phylogenetic tree of mitochondrial and chloroplast genomes in 314 grapevine accessions. 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 chloroplast, encompassing six populations.

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

Extended Data Fig. 8 | Quantile-Quantile (Q-Q) plot and Gene Ontology (GO) enrichment analysis. Genome-Wide Association Study (GWAS) Q-Q plot for the three populations, 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 (refer to Supplementary Table 12).

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

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

InDels).

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

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

Extended Data Fig. 13 | Visualization of multiple seed development-related datasets. a and **b**, Results from two independent 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Acknowledgements

We express our gratitude to Zhiwu Zhang of Washington State University, USA, for providing invaluable suggestions regarding genome selection and prediction. 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.

Author contributions

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

Competing interests

The authors declare no competing interests.

Data 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.

Code availability

All scripts performed in this study are available on GitHub:

https://github.com/zhouyflab/Polygenetic_Basis_Seedless_Grapes

Funding

This work was supported by the National Key Research and Development Program of China (No. 2023YFF1000100; 2023YFD2200700), the National Natural Science Fund for Excellent Young Scientists Fund Program (Overseas) to Yongfeng Zhou, the National Natural Science Fund (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.

Reference

1. Varoquaux, F., Blanvillain, R., Delseny, M. & Gallois, P. Less is better: new approaches for seedless fruit production. *Trends in biotechnology* **18**, 233-242 (2000).
2. Sardos, J. *et al.* 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 (2016).
3. Sidhu, J.S. & Zafar, T.A. Bioactive compounds in banana fruits and their health benefits. *Food Quality and Safety* **2**, 183-188 (2018).
4. Ye, W. *et al.* Seedless mechanism of a new mandarin cultivar 'Wuzishatangju'(Citrus reticulata Blanco). *Plant Science* **177**, 19-27 (2009).
5. Zhang, S. *et al.* Comparative transcriptome analysis during early fruit development between three seedy citrus genotypes and their seedless mutants. *Horticulture research* **4**(2017).
6. Andrus, C.F., Seshadri, V. & Grimball, P.C. *Production of seedless watermelons*, (US Agricultural Research Service, 1971).
7. Wijesinghe, S., Evans, L., Kirkland, L. & Rader, R. A global review of watermelon pollination biology and ecology: The increasing importance of seedless cultivars. *Scientia Horticulturae* **271**, 109493 (2020).
8. Akkurt, M., Tahmaz, H. & Veziroğlu, S. Recent developments in seedless grapevine breeding. *South African Journal of Enology and Viticulture* **40**, 1-1 (2019).
9. Reynolds, A., Wardle, D., Zurowski, C. & Looney, N. Phenylureas CPPU and

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

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

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

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

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

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

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

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

17. Ehlers, K. *et al.* 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 (2016).

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

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

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

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

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

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

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

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

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

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

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

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

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

30. Chamizo-González, F. *et al.* 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 Chemistry* **413**, 135591 (2023).

31. He, H. *et al.* 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 (2019).

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

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

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

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

35. Tang, Y. *et al.* 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 (2020).

36. Wang, L., Dai, W., Shi, Y., Wang, Y. & Zhang, C. 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 (2022).

37. Wang, L. *et al.* 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. *Physiology and Molecular Biology of Plants*, 1-12 (2023).

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

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

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

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

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

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

44. Ocarez, N. *et al.* 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 (2020).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

62. Wang, L. *et al.* 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 (2015).

63. Sheng, Z. *et al.* 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 (2022).

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

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

65. Cui, M. *et al.* 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 (2020).

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

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

68. Ahmad, B. *et al.* 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 (2019).

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

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

71. Tang, Y., Huang, C., Li, Y., Wang, Y. & Zhang, C. 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 (2021).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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