nature genetics

Article

https://doi.org/10.1038/s41588-024-01967-5

Grapevine pangenome facilitates trait genetics and genomic breeding

Received: 24 July 2023

Accepted: 1 October 2024

Published online: 4 November 2024

Check for updates

Zhongjie Liu[1,9], Nan Wang[1,9], Ying Su[1,9], Qiming Long[1,9], Yanling Peng[1,9], Lingfei Shangguan[2,9], Fan Zhang[1], Shuo Cao[1], Xu Wang[1], Mengqing Ge[2], Hui Xue[1], Zhiyao Ma[1], Wenwen Liu[1], Xiaodong Xu[1], Chaochao Li[1,3], Xuejing Cao[1], Bilal Ahmad[1], Xiangnian Su[1], Yuting Liu[1], Guizhou Huang[1], Mengrui Du[1], Zhenya Liu[1], Yu Gan[1], Lei Sun[4], Xiucai Fan[4], Chuan Zhang[5], Haixia Zhong[5], Xiangpeng Leng[6], Yanhua Ren[2], Tianyu Dong[2], Dan Pei[2], Xinyu Wu[5], Zhongxin Jin[1,3], Yiwen Wang[1], Chonghuai Liu[4], Jinfeng Chen[7], Brandon Gaut[8], Sanwen Huang[1,3], Jinggui Fang[2,6] ✉, Hua Xiao[1] ✉ & Yongfeng Zhou[1,3] ✉

Grapevine breeding is hindered by a limited understanding of the genetic basis of complex agronomic traits. This study constructs a graph-based pangenome reference (Grapepan v.1.0) from 18 newly generated phased telomere-to-telomere assemblies and 11 published assemblies. Using Grapepan v.1.0, we build a variation map with 9,105,787 short variations and 236,449 structural variations (SVs) from the resequencing data of 466 grapevine cultivars. Integrating SVs into a genome-wide association study, we map 148 quantitative trait loci for 29 agronomic traits (50.7% newly identified), with 12 traits significantly contributed by SVs. The estimated heritability improves by 22.78% on average when including SVs. We discovered quantitative trait locus regions under divergent artificial selection in metabolism and berry development between wine and table grapes, respectively. Moreover, significant genetic correlations were detected among the 29 traits. Under a polygenic model, we conducted genomic predictions for each trait. In general, our study facilitates the breeding of superior cultivars via the genomic selection of multiple traits.

The cultivated grapevine (*Vitis vinifera* ssp. *vinifera* L.) is an economically important perennial fruit crop that is grown widely for winemaking and fresh fruit in ~94 countries[1,2]. Previous studies have suggested that grapevine originated from a single domestication event in the Black and Caspian Sea regions more than 10,000 years ago, which subsequently spread across the northern hemisphere with gene flow from local wild populations[1–6]. However, other studies have suggested the potential for multiple domestication events[7–9]. Since domestication, grapevine cultivars have accumulated deleterious genomic variants, including single-nucleotide polymorphisms (SNPs) and SVs, in a heterozygous state, resulting in strong inbreeding depression[2,10]. Recent studies have highlighted the potential contribution of hidden genomic variants, including SVs[10–15], to phenotypes, but the quantitative genetic basis of complex agronomic traits in grapevine has rarely been investigated at the genome scale.

Long-read sequencing technologies have revealed the prevalence of SVs in plant genomes. It is increasingly evident that SVs are more likely than SNPs to influence the phenotype of domestication traits[13,16–18]. At the population level, SVs tend to occur at low frequencies, reflecting negative selection signals[10,13,19]. Furthermore, the frequency of SVs may be related to their recent origin. For example, recent transposable element (TE) activity can generate new SVs that are initially present in only one individual or lineage[20]. In part because of their low population frequencies, SVs are typically in low linkage disequilibrium

A full list of affiliations appears at the end of the paper. ✉e-mail: fanggg@njau.edu.cn; xiaohua01@caas.cn; zhouyongfeng@caas.cn

(LD) with SNPs. One practical implication of low LD is that SVs may encompass substantial missing heritability for quantitative traits[10,13]. Consistent with this viewpoint, the addition of SVs to population and quantitative genetic analyses has yielded new insights into local adaptation and agronomic traits[12,13,17,21].

Grapevine genomes are highly heterozygous, partly because of the accumulation of genetic variation during clonal propagation, which has been carried out for thousands of years[6,10,22,23]. For example, the genomes of diploid Chardonnay and Cabernet Sauvignon contain more than 10% heterozygous sites including SNPs, insertion–deletions (indels) and SVs[10,24,25]. Although the commonly used reference genome from PN40024 was highly homozygous after nine generations of selfing, it is missing >10% of genes compared with heterozygous cultivars[26,27]. Across cultivars, only ~7% of the genes are shared, whereas ~8% are unique to each individual[28]. The high level of variability in grapevine merits the construction of a pangenome reference that incorporates presence–absence variation, improves the detection of genomic variants, including SV, and reduces reference biases[29–34].

Here we assembled 18 haplotype-resolved telomere-to-telomere (T2T) assemblies representing eight diploid grapevine cultivars and one diploid wild grape. We then constructed a graph-based pangenome, which we call Grapepan v.1.0, using these new assemblies and 11 previously published chromosomal assemblies. These genotypes represent the global genetic diversity of grapes. Using Grapepan v.1.0, we built a variation map that includes SNPs, indels (2 bp ≤ indel < 50 bp) and SVs (≥50 bp) across a larger sample of 466 accessions, including 324 that were newly sequenced. We utilized this variation map in a genome-wide association study (GWAS) and the genomic prediction of 29 complex agronomic traits. This exercise identified quantitative trait loci (QTLs) for these agronomic traits, provided unique insights into the contribution of SVs to quantitative genetic variation and demonstrated the feasibility of breeding superior cultivars via genomic selection for multiple traits. The pangenome reference (Grapepan v.1.0), variation map, QTLs and our genomic selection models facilitate genomic breeding of grapevine.

Results

The graph pangenome reference for grapevine (Grapepan v.1.0)

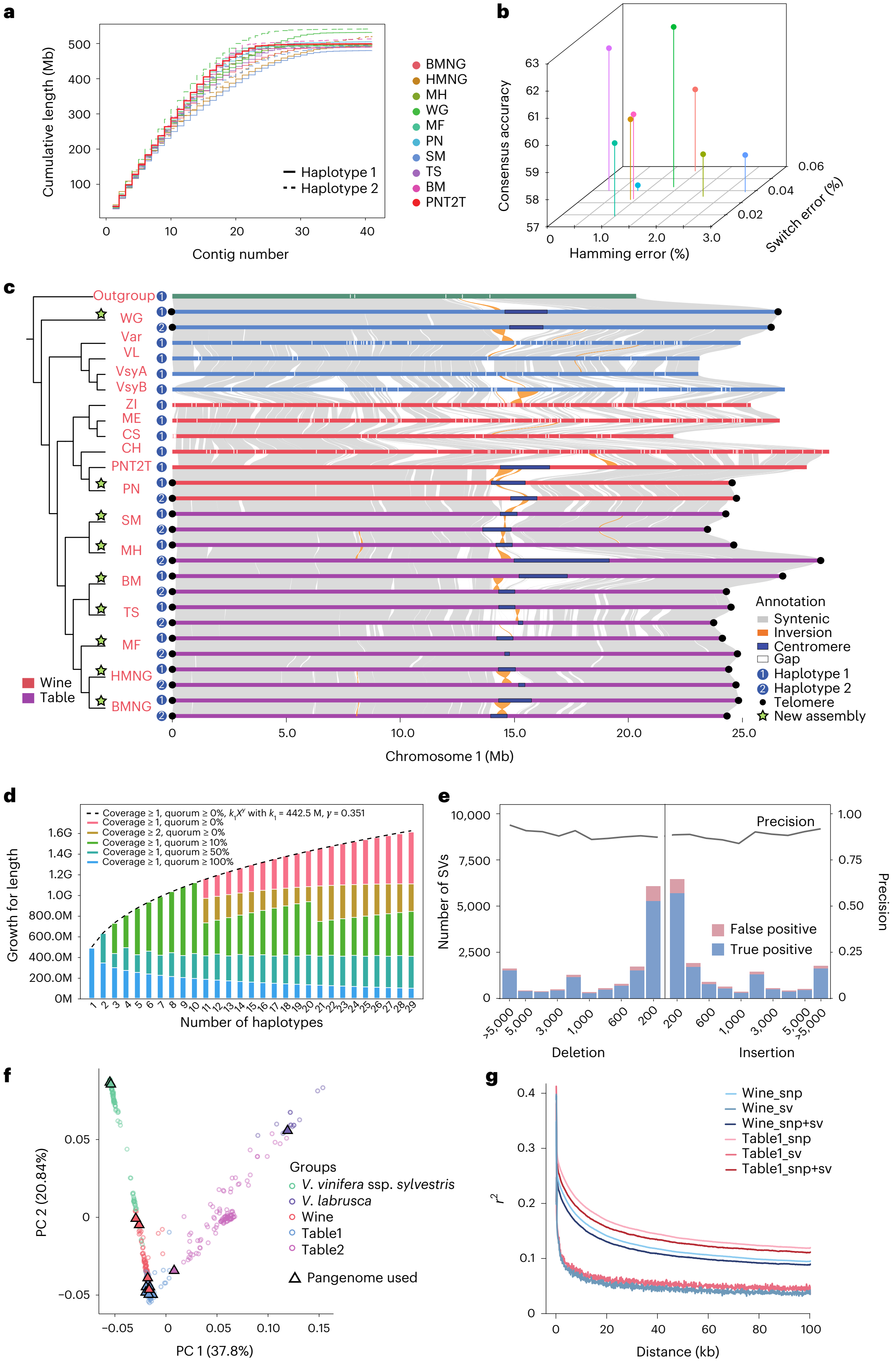
HiFi reads, Hi-C reads and ultra-long nanopore reads were collected for nine representative diploid samples, including one accession of *Vitis retordii*, a wild species endemic to Asia, and eight grapevine cultivars (seven table grapes and one wine grape) (Supplementary Table 1). The nine samples resulted in 18 haplotypes that reached T2T-level assembly after gap filling (Supplementary Fig. 1). Genome sizes ranged from 479.15 to 539.30 Mb (Supplementary Table 2). The quality of haplotype assembly was confirmed by high contiguity (>99.9%), minimal switching error (<0.05%) and low Hamming error[35] (<2.83%) (Fig. 1a,b and Supplementary Table 3). Benchmarking universal single-copy orthologs evaluation indicated an average completeness of 98.4% for these haplotypes (range 98.07% to 98.64%; Supplementary Fig. 2). We used the same pipeline to annotate all haplotypes and to ensure consistent results. Across the 18 haplotypes, the number of protein-coding genes ranged from 34,536 to 38,526 (Supplementary Table 2), and the TE sequence length per haplotype ranged from 263.86 Mb (54.68%) to 312.10 Mb (59.03%) (Supplementary Table 4). In addition, we identified centromere and telomere sequences in all assemblies (Fig. 1c and Supplementary Table 5). Consistent with previous studies[27], the predominant repeat unit of the centromere was 107 bp long. Overall, these 18 assembled haplotypes and their annotations represent one of the highest-quality grapevine genomic datasets generated to date.

To represent genetic diversity among grapevines, we collected 11 previously published assemblies from five cultivars (one with haplotype-resolved assemblies and four primary assemblies) and four wild accessions (one with haplotype-resolved assemblies and three primary assemblies) (Fig. 1c and Supplementary Table 1). Using these samples, we investigated the heterozygosity within the genome. We found that the heterozygosities of *Vitis retordii, V. arizonica*, *V. labrusca* and *V. vinifera* ssp. *sylvestris* are significantly lower than those of the cultivars (excluding PNT2T, telomere-to-telomere grape genome of PN40024 (ref. 27)) ($P < 0.05$, Student's *t*-test). For example, genome heterozygosity was 0.40% in *V. labrusca* compared with an average of 1.42% across *V. vinifera* cultivars, whereas it was 1.66% in an interspecific hybrid (Shine Muscat, *V. labrusca* × *V. vinifera*) (Supplementary Fig. 3a). We also identified genome-wide presence and absence variations between the two haplotypes from our newly sequenced accessions. The results showed that an average of 12.57% (4,645) of genes were in a hemizygous state, which was identified as the entire gene structure affected by SVs (Supplementary Fig. 3b). Utilizing gene annotations obtained from these 18 grape accessions, we identified core and variable gene families. Our analysis revealed a total of 30,268 gene families, with 19.07% identified as core families and 77.00% categorized as variable families, including 20.40% soft-core families and 56.60% nonessential families. In addition, 3.93% of the identified gene families were classified as private (Supplementary Fig. 4).

An unbiased pangenome reference is crucial for discovering global genetic diversity, including SVs, among grape genomes. Using 29 haplotypes from the sequenced samples, we constructed two graph-based pangenomes based on the PanGenome Graph Builder (PGGB) and Minigraph-Cactus (MC) programs. The total length of the MC-based pangenome (Grapepan v.1.0) reached 1.43 Gb, which is 2.88 times that of the PNT2T genome (Fig. 1d). We identified SVs from graph deconstruction and integrated assembly alignments to validate the sensitivity and precision of SV detection and then repositioned SVs using PNT2T coordinates. Altogether, we detected 236,449 reliable SVs with high precision (Fig. 1e), and we verified large SVs (>10 kb) by mapping HiFi reads (Supplementary Fig. 5). The PGGB pangenome provided a similar result to the MC pangenome (Supplementary Fig. 6). Based on the Grapepan v.1.0 SV map, we found a biased distribution of SVs throughout the genome, including two SV hotspots near the centromeres of chromosome (Chr) 9 and Chr19 (Supplementary Fig. 7a). A significant fraction of the observed SVs had low population frequencies (Supplementary Fig. 7b). On average, 88.6% of the SVs overlapped TEs, whereas 30.9% intersected gene structures (Supplementary Fig. 7c,d). The high concordance between TEs and SVs indicates that the former play a critical role in driving genetic variation events.

To expand the genome-wide SV map, we integrated short-read sequencing data from 466 accessions, including 324 that were newly sequenced (Supplementary Tables 6 and 7). The genotypes of pangenome-based SVs in these accessions were characterized,

Fig. 1 | T2T genome assemblies and the construction of Grapepan v.1.0. **a**, NGx plot showing the assembly continuity of the 18 newly assembled haplotypes compared with the published PNT2T assembly. Two haplotypes (haplotype 1 and haplotype 2) of the same individual are distinguished. **b**, Assessment of the assembly for nine sequenced grape accessions (BMNG, HMNG, MH, WG, MF, PN, SM, TS and BM). The quality values demonstrate the base-level accuracy of each sample. The phasing accuracy is indicated by the percentages of switch errors and hamming errors. **c**, Comparative genomics of 27 (published genomes only selected the primary haplotype) assemblies and one assembly of *Muscadinia rotundifolia* as outgroup for chromosome 1. **d**, Total length of MC pangenome sequences with different numbers of haplotypes; M represents megabase pairs and G represents gigabase pairs. **e**, Validation of pangenome deletions and insertions involved counting SVs of varying lengths and calculating the accuracy. **f**, PCA of the first two components of the 466 sequenced grape accessions. Different grape groups are distinguished by different colors. The samples used to construct Grapepan v.1.0 represent a wide range of genetic diversity. PC, principal component. **g**, The decay of LD was calculated based on three different datasets: SVs, SNPs and SVs + SNPs.


a
Cumulative length (Mb)
500
400
300
200
100
0
10
20
30
40
Contig number
Haplotype 1
Haplotype 2
BMNG
HMNG
MH
WG
MF
PN
SM
TS
BM
PNT2T
b
Consensus accuracy
63
62
61
60
59
58
57
0
1.0
2.0
3.0
Hamming error (%)
0.02
0.04
0.06
Switch error (%)
c
Outgroup
WG
Var
VL
VsyA
VsyB
ZI
ME
CS
CH
PNT2T
PN
SM
MH
BM
TS
MF
HMNG
BMNG
Wine
Table
Annotation
Syntenic
Inversion
Centromere
Gap
Haplotype 1
Haplotype 2
Telomere
New assembly
0
5.0
10.0
15.0
20.0
25.0
Chromosome 1 (Mb)
d
Coverage ≥ 1, quorum ≥ 0%, k_1X^γ with k_1 = 442.5 M, γ = 0.351
Coverage ≥ 1, quorum ≥ 0%
Coverage ≥ 2, quorum ≥ 0%
Coverage ≥ 1, quorum ≥ 10%
Coverage ≥ 1, quorum ≥ 50%
Coverage ≥ 1, quorum ≥ 100%
Growth for length
1.6G
1.4G
1.2G
1.0G
800.0M
600.0M
400.0M
200.0M
0M
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Number of haplotypes
e
Number of SVs
10,000
7,500
5,000
2,500
0
Precision
1.00
0.75
0.50
0.25
0
False positive
True positive
>5,000
5,000
3,000
1,000
600
200
200
600
1,000
3,000
5,000
>5,000
Deletion
Insertion
f
PC 2 (20.84%)
0.05
0
−0.05
−0.05
0
0.05
0.10
0.15
PC 1 (37.8%)
Groups
V. vinifera ssp. sylvestris
V. labrusca
Wine
Table1
Table2
Pangenome used
g
r^2
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
Distance (kb)
Wine_snp
Wine_sv
Wine_snp+sv
Table1_snp
Table1_sv
Table1_snp+sv

providing a more thorough understanding of SV frequencies across a diverse set of accessions. In total, the variation map contained 8,591,818 SNPs, 513,969 indels and 236,449 SVs.

Grapepan v.1.0-based population structure of grapevine

We used the genome-wide SNPs to explore genetic relationships among grapes (Fig. 1f and Supplementary Fig. 8). Population structure analyses identified three major grape species or subspecies (*V. labrusca*, *V. vinifera* ssp. *sylvestris* and *V. vinifera* ssp. *vinifera*) within 466 collected grape accessions (Supplementary Figs. 9 and 10). We further analyzed modern cultivars of *V. vinifera* ssp. *vinifera*, which separated into three groups upon ADMIXTURE and PHATE analyses: winemaking varieties from Europe and the Middle East (Wine), and two groups of table grapes, one from Europe and Eastern Asia (Table1) and one from hybridization with *V. labrusca* (Table2, *V. labrusca* × *V. vinifera*). The Wine group had a high identity-by-state value (0.82), indicating a long history of shared genomic segments (Supplementary Fig. 11). The recent hybrid origin of the Table2 group may contribute to its lower identity-by-state value (0.79), low recessive deleterious burden and relatively high heterozygous burden compared with other groups. These observations support the vigorous phenotypes of Table2 grapes, because their deleterious burden is hidden in the heterozygous state owing to their hybrid origin (Supplementary Fig. 12). We assessed LD decay using both SNPs and SVs within the Wine and Table1 groups. LD decay was consistent across groups and rapid, but although was even more rapid among SVs (Fig. 1g and Supplementary Fig. 13). This rapid decay likely reflects the fact that SVs were typically observed in low frequencies (Supplementary Figs. 14 and 15). However, the rapid decay of LD between SVs and SNPs suggests that part of the missing heritability for quantitative traits could be hidden among SVs in the grape genome[10].

GWAS of complex agronomic traits and the importance of SVs

To investigate the contribution of SVs to quantitative traits, we performed a phenotypic survey of 29 traits over 2 years (2016 and 2017) for the 324 newly sequenced accessions (Fig. 2a). The sample of 324 cultivars included 106 Wine grapes, 108 Table1 grapes and 110 Table2 grapes (Supplementary Fig. 16). The 29 traits comprised five phenotypic categories: bunch (six), contents (eight), berry traits (eight), fruit size (four) and skin (three) (Fig. 2a). GWAS has been used to investigate phenotypic traits related to the composition and dimensions of fruit using ~6,000 SNPs[36]. Our larger sample size and genome-scale variants facilitate genomic selection for multiple traits simultaneously. We began by analyzing correlations among quantitative traits over the 2 years (Supplementary Fig. 17) and by mapping phenotypes in principal component analysis (PCA) (Fig. 2b). In the berry content category, pairwise correlations for the traits fructose (Fru), glucose (Glu) and soluble solids content (SSC) were significantly positive ($P < 0.001$). Traits were also correlated between different categories. For example, titratable acid (TAC) of the content category had a significant negative correlation with four measurements of fruit size: berry weight (BeWe), berry volume (BV), berry length (BL) and berry width (BeWi) ($P < 0.001$; Supplementary Fig. 17). These correlations likely follow from the fact that the synthesis of acids typically ceases during veraison, and this cessation contributes to the dilution of acid concentrations as the fruit continues to ripen and expand[36]. These correlations among phenotypes may lead to an overlap of some candidate GWAS loci and have potential implications for the genomic selection of elite grape varieties with multiple desirable traits.

To elucidate differences in 29 agronomic traits at the population level, we performed uniform manifold approximation and projection (UMAP) analyses. The first two components (UMAP1 and UMAP2) established trait differences between the Wine, Table1 and Table2 groups. Within-group trait distances were significantly lower than between-group distances ($P < 0.001$) (Fig. 2d). We then mapped the normalized values of agronomic traits onto the UMAP analysis to characterize the distribution of each trait among three groups. We found that 25 of 29 agronomic traits were evenly distributed without obvious group differences (Fig. 2c and Supplementary Fig. 18).

Previous GWAS analyses have faced challenges in simultaneously accommodating a large variety of cultivars, multiple phenotypic traits and high-resolution data analysis[37–39], and no previous study of grapevine has included SVs in GWAS analyses. We performed GWAS analyses of SNPs and SVs using Grapepan v.1.0, based on a joint dataset containing 2 years of GWAS results (Fig. 3a). A total of 148 loci were significantly associated with agronomic traits, including 136 genomic regions that were detected by SNPs and 12 that were captured by SVs. Altogether, 27.61 Mb (~5.58%) of the genome was associated with at least one of the 29 agronomic traits (Supplementary Table 8). Of the 148 candidate regions, 26 (~17.57%) overlapped with loci identified by previous functional studies (Supplementary Table 9). For example, based on the SNP dataset, we detected a locus associated with the seedless or seedness trait on grapes in Chr18 (31.41–31.45 Mb), which contains a MADS-box gene, agamous-like 11 (*AGL11)*, responsible for the development of ovules into seeds after fertilization[40]. Similarly, we identified a 95-bp deletion in the BL1 locus that was significantly associated with grape berry length. This variant, which was located in the exon region of *Vitvi011427* and encodes a NAD(P)-linked oxidoreductase superfamily protein, had a phenotypic variation explained (PVE) value of 6.31% and was present in 20.4% of sequenced grapes (Fig. 3b and Supplementary Table 8). In addition, a significant 1.1 kb deletion in the SN6 locus had a PVE of 6.08%, was associated with a photolyase coding gene *Vitvi030206* and had a frequency of 13.0% across sequenced grapes (Fig. 3c). A 139 bp insertion specific to Table2 grapes was located in the Suc1 locus associated with sucrose content, demonstrated a PVE of 6.60% and was present in 56.9% of Table2 grapes (Fig. 3d). This insertion is close to a gene homologous to *AtRHM1*, which encodes an enzyme involved in UDP-beta-L-rhamnose biosynthesis. We also performed GWAS of the SNP dataset from the PNT2T single reference genome. We found that 124 of 136 (91.18%) loci were detected by both the pangenome SNP and PNT2T reference SNP (Supplementary Fig. 19). Collectively, our GWAS analysis based on the pangenome integrated SVs and SNPs to improve mapping of important traits.

We compared candidate GWAS loci for traits from different phenotypic categories and found a candidate locus (SSC7, Chr17:6.47–6.53 Mb) for SSC that was close to a candidate locus (BeWi9, Chr17:6.47–6.65 Mb) responsible for berry width. There were two most significant SNPs (17_6489512 and 17_6484258) with PVE values of 6.05% and 5.91% for SSC7 and BeWi9 loci, respectively (Fig. 3e,f). We constructed a local phylogenetic tree using the variations from the combined region (6.47–6.65 Mb). The tree featured a tight cluster within cultivated grape groups that had an extremely short inner branch length (Supplementary Fig. 20). The combined region had low genetic diversity compared with the rest of the genome, suggesting a selective sweep (Fig. 3g). The homozygous genotype of SNP 17_6489512 showed significantly lower soluble solids content ($P = 6.778 \times 10^{-4}$) and the homozygous genotype of SNP 17_6484258 showed significant higher berry width ($P < 1.12 \times 10^{-15}$) (Fig. 3h). Based on the genome annotation in this locus, we identified two gene clusters, the NEPS ([−]-isopiperitenol and/or [−]-carveol dehydrogenase) family (five members) and NRT1 (nitrate or dipeptide and/or tripeptide transporters) family (four members), and examined the expression of these genes in berries (Fig. 3i). *Vitvi031750* of the NEPS family and *Vitvi031760* of the NRT1 family had significantly high expression in four grapevine cultivars of both haplotypes, whereas the expression of *Vitvi031756* in two Wine grapes (Merlot and Cabernet Sauvignon) was higher than Table1 grapes (Fujiminori and Venus Seedless).

Divergent selection on agronomic traits in grapevine

To determine whether the QTLs associated with complex traits were under selection during divergence between populations, we performed

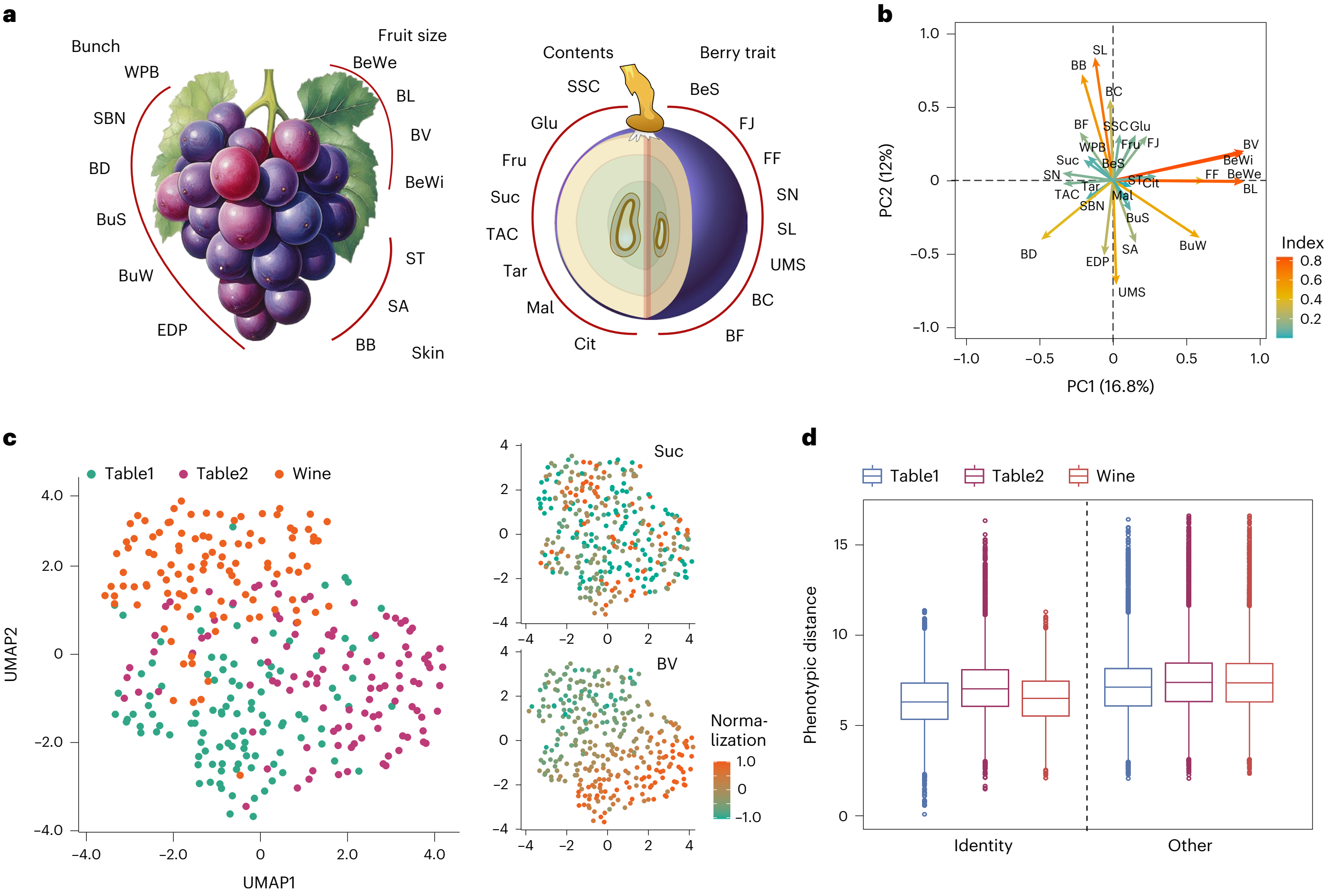


Fig. 2 | The correlation of 29 agronomic traits among different grape populations. a, Schematic diagram of agronomic traits of grape fruits investigated, including five identified categories. Traits in each category are labeled. **b**, PCA map showing the relationships among all agronomic traits. The distance between variables and the origin measures the quality (index by cos2 value) of the variables. **c**, UMAP plot (base map) generated from 29 trait scores for three grape populations. Scores were scaled and centered (*Z*-score) across all individuals for each trait independently. The points indicate individual grapes and are colored by different populations. UMAP plots from content of Suc and BV traits were generated by mapping these scores to the UMAP base map. **d**, Box plot of mean pairwise phenotypic distances within the population (identity), and between all other populations (other). Sample sizes are 5,778, 5,778, 5,565, 23,112, 23,112 and 22,896 pairs. Boxes, 25% to 75% quartiles; horizontal line, median; whiskers, inner fence within 1.5× box height; circles, outliers within 1.5× box height; asterisks, outliers beyond 1.5× box height. Statistical significance was determined by two-sided Student's *t*-tests. BB, berry bloom; BC, berry color; BD, bunch density; BeS, berry shape; BF, particularity of flavor; BuS, bunch shape; BuW, weight of a single bunch; Cit, content of citric acid; EDP, ease of detachment from pedicel; FF, firmness of flesh; FJ, juiciness of flesh; Mal, content of malic acid; SA, astringent of skin; SBN, number of subsidiary bunch; SL, length of seeds; SN, number of seeds; ST, thickness of skin; UMS, uniformity of time of physiological stage of full maturity of the berry; WPB, number of wings of the primary bunch.

XP-EHH (cross population extended haplotype homozogysity)[41] analyses between the two table grape groups (Table1 and Table2). We found a total of 21.45 Mb (4.4%) regions that were significantly differentiated ($P < 0.05$) (Fig. 4a). The top 5% outliers in fixation statistics (F_{ST}) analysis showed a similar pattern (Supplementary Fig. 21). Gene set enrichment analysis (GSEA) revealed enrichment of four Gene Ontology (GO) terms associated with hormone responses and stress responses among the set of genes in the diverged genomic regions (Fig. 4c). Comparing the highly differentiated regions with the regions associated with phenotypes, six GWAS candidate loci were located in divergent genomic regions, and these were associated with berry color (BC4 locus), skin astringent (SA1), berry shape (BeS2), bunch weight (BuW5), flesh firmness (FF6) and tartaric acid content (Tar4) (Supplementary Fig. 22). Among them, a BC4 locus containing multiple *MYB* genes on Chr2 associated with berry color[42] based on the pangenome SV dataset (Supplementary Table 10). The FF6 locus explained 7.35% of the variation for flesh firmness, which differed between groups because the Table2 group had an 11.7% increase in flesh firmness compared with the Table1 group ($P < 0.01$) (Supplementary Fig. 23).

The different usage of grapevine cultivars (for winemaking or consumption as fresh fruit) might drive genetic and phenotypic divergence between cultivated populations. We found that approximately 21.35 Mb of the genomic region was significantly different between the Wine group and the Table1 group based on XP-EHH analysis ($P < 0.05$) (Fig. 4b). The GSEA results indicate that amino sugar, glutathione and chitin metabolic processes, and the toxin catabolic process are enriched in the differentiated genomic regions (Fig. 4d). We detected 45 candidate GWAS loci associated with population divergent regions that determine berry size across five traits: BeWi, BL, BV, BeWe and weight of a single bunch (Supplementary Table 8). Metabolites provide rich flavors to wine, and 32 related candidate loci were enriched based on eight metabolic phenotypes (TAC, SSC, Glu, Fru, Suc, Tar, malic acid and citric acid). Within the diverged regions, we also identified five GWAS candidate loci (BV12, BeWe6, BuW2, BV15 and BeWi9) associated with fruit size, and two GWAS candidate loci (TAC3 and SSC7) associated with metabolites (Supplementary Fig. 22 and Supplementary Table 11). Overall, our analysis suggests that divergent selection on agronomic traits is associated with different breeding targets.

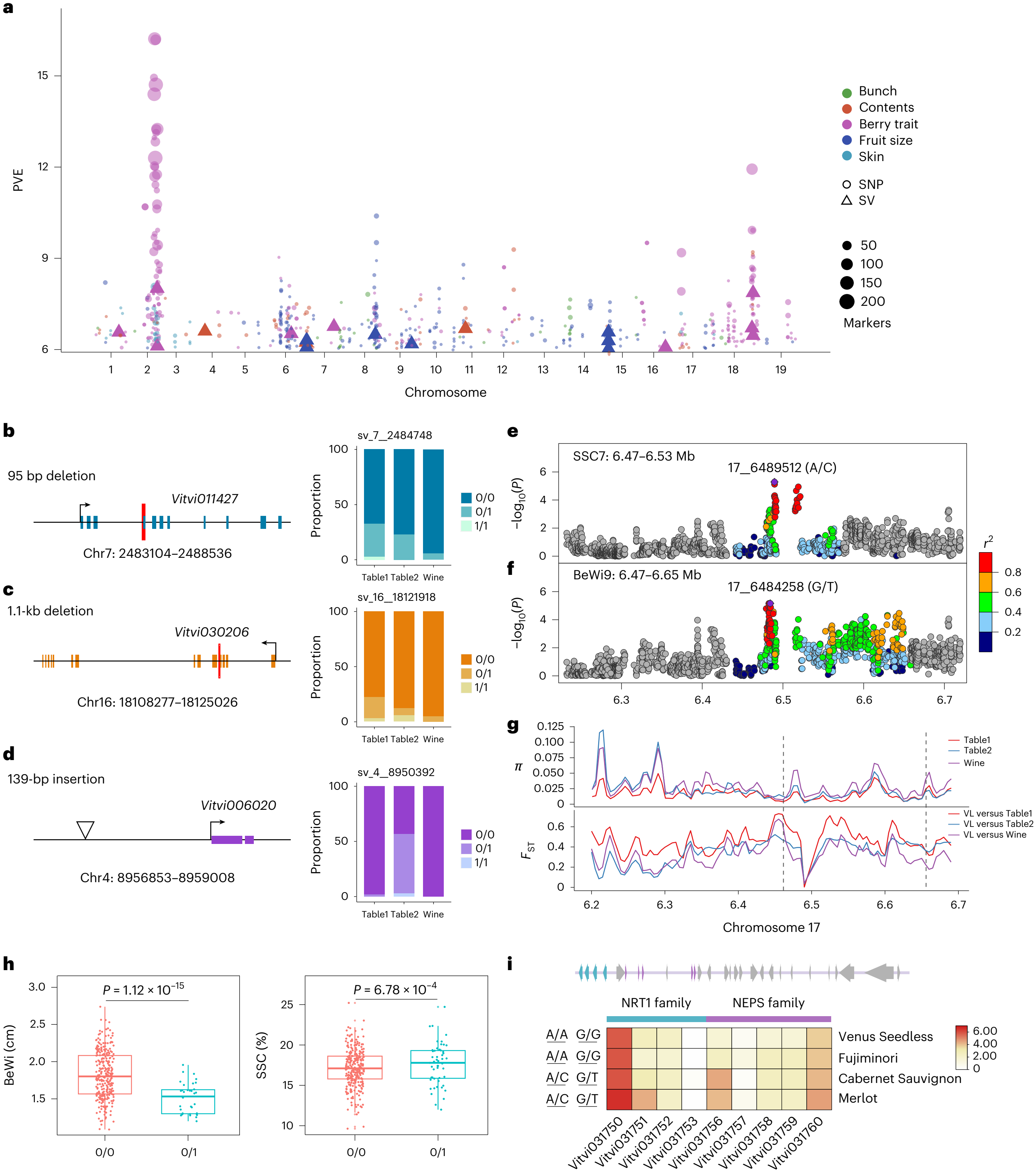


Fig. 3 | Candidate loci associated with agronomic traits and their genomic footprints of artificial selection. a, Integrated GWAS map for 29 grape agronomic traits. The ordinate represents the PVE of the trait. **b**–**d**, Three significant SVs for BL1 (**b**), SN6 (**c**) and Suc1 (**d**) GWAS loci and their populational frequencies. Left, Gene models with coding regions and transcription direction. The corresponding deletions and insertions are highlighted. Right, Proportions of different genotypes in three populations. **e**,**f**, The GWAS results and linkage for SSC7 (**e**) and BeWi9 (**f**) loci. Statistical significance was determined by generalized least squares *F*-test. **g**, The nucleotide diversity (π) and F_{ST} around this candidate region. The vertical dashed lines indicate the combined GWAS loci of BeWi or SSC traits, VL represents *V. labrusca*. **h**, Proportions of different genotypes at significant SNP sites in BeWi ($nGT_{0/0} = 272$, $nGT_{0/1} = 51$) and SSC ($nGT_{0/0} = 288$, $nGT_{0/1} = 58$) in 2016 (center line, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range). $nGT_{0/0}$ and $nGT_{0/1}$ refer to the counts of different genotypes at significant SNP sites. Statistical significance was determined by two-sided Student's *t*-tests. **i**, Gene annotation in candidate region (upper) and expression level (transcripts per kilobase of exon model per million mapped reads) of genes in the candidate region in different grape cultivars (lower).

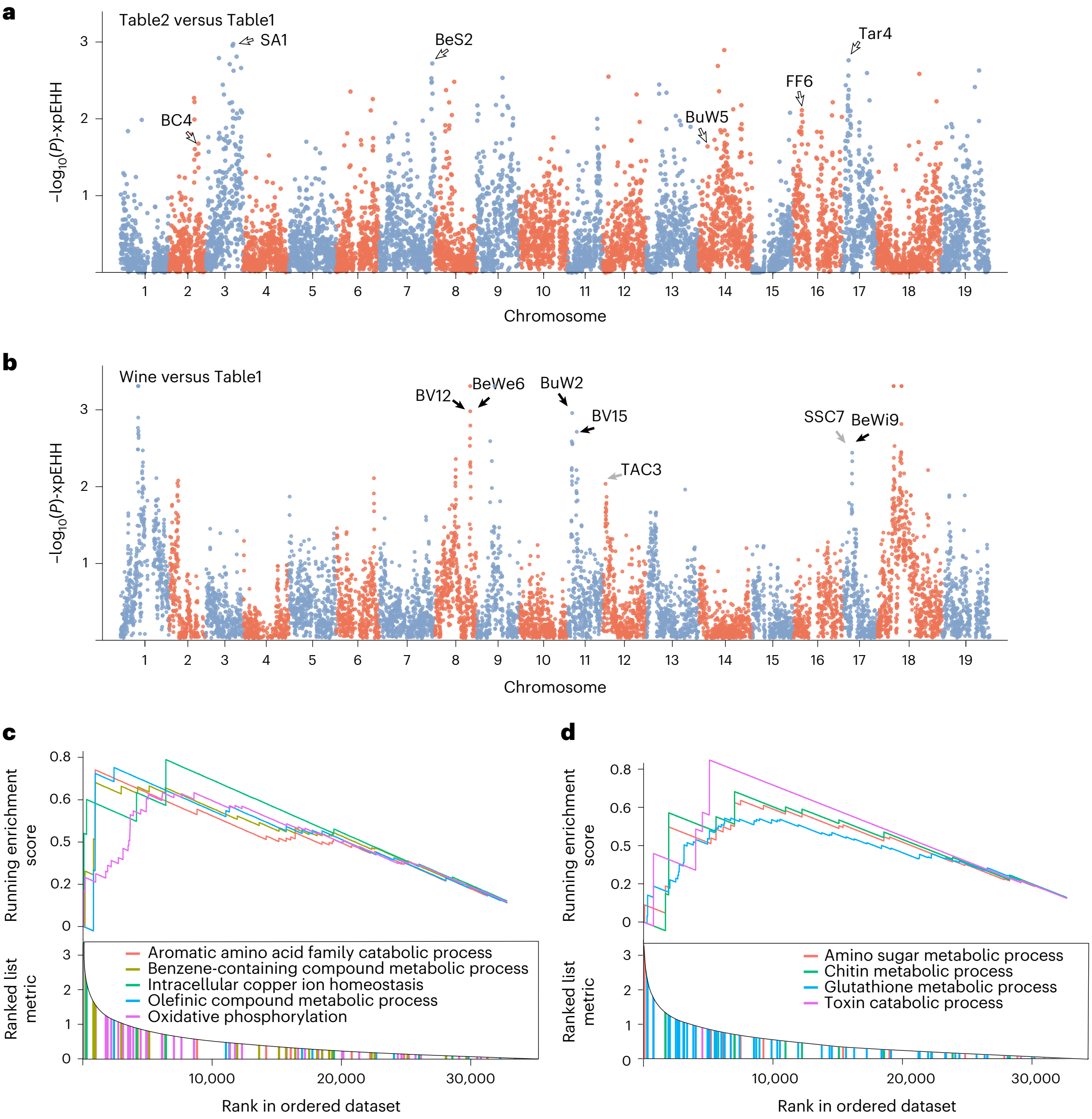
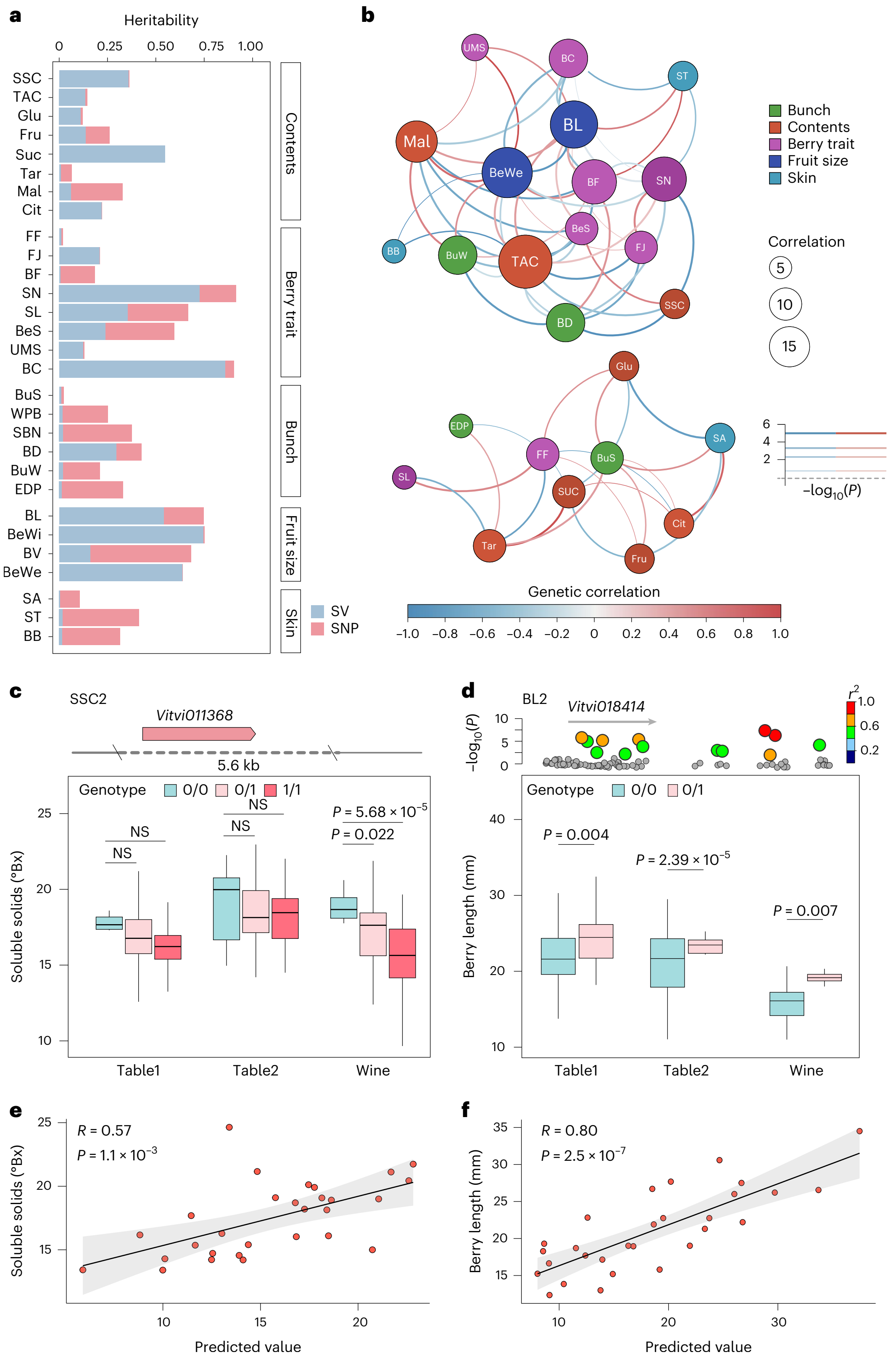


Fig. 4 | Divergent selection on agronomic traits among subpopulations. **a**,**b**, Selection of the XP-EHH genomic scan for the Table1 versus Table2 (**a**) and Wine versus Table1 (**b**). $n_{\mathrm{Sites}} = 8{,}508$, FDR (Benjamini–Hochberg) correction. Arrows indicate the highly correlated loci associated with the agronomic traits analyzed by GWAS. **c**,**d**, GSEA analyzes genes ranked in divergent regions of the genome and included two comparisons, Table1 versus Table2 (**c**) and Wine versus Table1 (**d**). The vertical dashed lines indicate the leading-edge subset and the max rank of the enriched gene.

SVs enhance heritability estimates for grape traits

Because most SVs are not linked with SNPs (Fig. 1g), they could potentially contribute to the missing heritability in association analyses and genomic scanning tests[13]; in fact, we have already shown that a few candidate GWAS regions were identified with only SVs and not SNPs. We further investigated the contribution of SVs to phenotypic traits by using the LDAK model to estimate the proportion of phenotypic variance that is explained by genetic variants. The use of only SVs or SNPs limited the power of prediction for most agronomic traits, with the heritability contributed by SNPs ranging from 0.01% to 52.3% and that contributed by SVs ranging from 0.5% to 86.1% (Fig. 5a). The response to quantitative traits was dominated by SNPs, which implies polygenic architecture with minor effects from many loci, whereas qualitative traits were mostly affected by SVs with potentially large effects. Our analysis indicated that SVs contributed more to the heritability of 15 traits than SNPs. For example, genome-wide SVs explained 74.6% of the variance in BeWi but genome-wide SNPs explained only 0.5% (Fig. 5a). Similarly, SVs contributed 35.8% of the captured heritability in the SSC, but SNPs contributed only 0.6%. One 5.6 kb deletion on Chr7 explained 6.23% of the PVE for SSC (SSC2, Chr7:2029369–2032050) (Fig. 5c). The Wine grape accessions with a heterozygous deletion of this SV had a significantly lower SSC than accessions without the deletion. We suspect this might be related to the regulation of *Vitvi011368*, a gene encoding an isoamylase. The GWAS result revealed a significant association between BL and an SNP (BL2 locus, Chr10_9052243, PVE: 6.63%) (Fig. 5d). SNPs contributed to 20.9% of the captured heritability in BL, whereas SVs further improved the captured heritability to 64.9%.

We calculated the genetic correlation between traits to assess the genomic selection of multiple traits for grape breeding. We combined pangenome SVs and SNPs to calculate pairwise genetic correlations among 29 traits, 20.7% of which showed a significant signal



($P < 0.05$). The genetic correlations (rG value) between various traits of fruits ranged from 0.41 to 0.97, suggesting a potential for concurrent selection of multiple traits in future breeding efforts (Fig. 5b). Single BeWe (SV heritability = 63.7%) and BL (SV heritability = 54.0%) were the main hubs of pairwise genetic correlations. Polygenic scores (PGS) can be used to aggregate effects across many genetic variants into a single predictive score, enabling the assessment of genome selection[43]. We evaluated PGS on the basis of GWAS summary statistics (pangenome SNPs + SVs) based on two derivation methods for all 29 traits (Supplementary Fig. 24 and Supplementary Table 12). The PGS prediction accuracy averaged >50% across all traits. As expected, traits with higher captured heritability tend to show improved prediction accuracy. The predictive accuracies obtained in this study showed improvements of at least 16%, and often higher, relative to previously

Fig. 5 | Missing heritability, genetic correlations and genomic predictions of agronomic traits. a, The heritability of 29 traits contributed by genomic SVs and SNPs. The contributions from SNP and SV were distinguished. **b**, The genetic correlations among 29 traits. FDR (Benjamini–Hochberg) corrected. **c**, The genotypes of the gene *Vitvi011368* and SSC in each group. Top, schematic diagram of a deletion in the region including the gene *Vitvi011368*. Bottom, SSC2 values in each group with different genotypes (center line, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range). The sample sizes, from left to right, are [6, 50, 54], [6, 40, 61] and [10, 42, 50]. Statistical significance was determined using two-sided Student's *t*-tests. **d**, GWAS result and LD analysis of the BL2 locus, which contained a candidate gene, *Vitvi018414*. Differences in BL between populations and genotypes were estimated (center line, median; box limits, first and third quartiles; whiskers, 1.5× interquartile range). The sample sizes, from left to right, are [38, 70], [12, 96] and [103, 3]. Statistical significance was determined using two-sided Student's *t*-tests. **e**, Linear regression analysis for phenotype prediction between SSC2 values and values predicted by LDpred2-auto. The confidence interval (CI) is shown by gray shading. The smoothed line represents a linear regression fit of the actual data, and the shading represents the CI. Sample size $n = 29$. **f**, Linear regression analysis for phenotype prediction between BL values and predicted values by lassoum2. The significance of the linear relationship between variables was evaluated through Pearson correlation coefficients. The smoothed line represents a linear regression fit of the actual data, and the shading represents the CI. Sample size $n = 29$. °Bx, degrees Brix; NS, nonsignificant.

reports for grapevines[36,37]. In particular, we found that BL has a higher heritability (74.9%), exceeding the 36.4% heritability captured in SSC. The PGS prediction showed an accuracy of 57.46% in SSC and 79.53% in berry length (BL) (Fig. 5e,f). Collectively, our analyses enable substantial prediction of agronomic traits in the grape breeding program.

Discussion

Accelerating the innovation of grape varieties is urgently required to adapt to future planting, rapidly changing market demands and climate change. Grape breeding exhibits a degree of reliance on older varieties; in particular, clonal reproduction allows the preservation of genotypes over extended periods, some of which are older than 900 years[22]. Advances in grapevine breeding lag far behind those made in annual cereal crops because of their long generation times (~3 years on average), high deleterious burden that leads to inbreeding and/or hybrid depression, high genomic heterozygosity, inefficient genetic transformations and limited knowledge about the genetic basis of complex agronomic traits.

Progress in understanding the complexity of the grapevine over the past two decades, from phenotypic characterization to marker identification and association analysis, has greatly benefited breeding efforts. Early breeding emphasized correlation analysis between phenotypic traits and low-density genetic markers, and selected phenotypic traits through marker-assisted selection[36,39,44]. Using association analyses, researchers have linked specific genetic variations to desirable phenotypic traits, providing breeders with valuable tools for the targeted selection of multiple phenotypes, including berry size, color and sugar content[37–39,45]. These efforts have led to significant contributions such as the development of seedless grape varieties and the enhancement of disease resistance in grapevines[46,47]. However, limitations inherent in detecting variations within a single reference genome hinder the identification of crucial variants associated with breeding traits and a comprehensive analysis of agronomic trait inheritance.

Advanced pangenome-based approaches underscore broader efforts aimed at discovering genetic variants in crop breeding[48]. Recent research has focused on North American wild grapevines and has established a nonreference pangenome inclusive of nine wild accessions[32]. Their sequencing encompasses the diversity of wild grapevine species, aiming to integrate resistance variants from wild species for use in rootstock improvement. By contrast, our pangenome (Grapepan v.1.0) focuses on discovering variants associated with agronomic traits in domesticated grapevines. We selected representative cultivated varieties to construct the pangenome. We also included table grape varieties to expand diversity across grape populations with different uses[49]. Therefore, our pangenome may contain more advantageous genotypes related to domesticated traits, thus directly serving breeding programs. We utilized a graph-based approach in which any variant is integrated as a node within the pangenome reference. Indeed, the most significant enhancement of the pangenome lies in the discovery of SVs[30,50]. The number of newly discovered SNPs differs only slightly compared with alignment with a single reference genome or previous pangenome versions[32]. Thus, our grape pangenome places greater emphasis on uncovering traits associated with SVs and revealing their inheritance patterns.

In Grapepan v.1.0, SVs often associated with repetitive sequences and TEs, suggesting that TE-mediated events are an important evolutionary force[49,51,52]. The low frequency of SVs in the grapevine genome can be attributed to recent TE activity and the evolutionary constraints imposed by natural selection. This poses challenges in precisely controlling the breeding process when relying solely on SNP for trait selection. This challenge is exacerbated by the incomplete capture of heritability for multiple traits, particularly from SVs, which might be related to two factors. First, LD decay can influence the resolution of genetic mapping and the identification of causal variants[53]. Second, SVs are often found to generate and explain a greater proportion of phenotypic variation in numerous traits compared with SNPs[54]. The rarity of SVs also makes it difficult to accurately estimate their frequency and effect size within a population[10]. Consequently, the statistical power to detect associations involving rare SVs is lower than that for SNPs. In addition, SVs are larger relative to SNPs and can engender more immediate functional consequences, such as perturbations in gene dosage or the disruption of critical gene regulatory elements[16,31]. For example, SVs contribute the largest share of heritability for approximately half of the molecular traits in tomatoes, the identification of SVs based on pangenome has greatly increased estimates of the heritability of metabolic traits[17]. In foxtail millet, the precision of 73.9% of traits with both SNP and SV markers increased by between 0.04% and 12.67% compared with SNP-only markers[14]. Fruit color serves as a key trait in grape breeding, renowned for its association with SV determination[10,55,56]. We confirmed the higher heritability in fruit color contributed by SV and emphasized the power and accuracy of SV-based GWAS and genomic selection. We have found that the inheritance of an isoamylase gene associated with a 5.6-kb deletion explained 6.23% of the variance in SSC. Collectively, a deep understanding of SVs based on the pangenome will greatly improve the efficiency of SV-associated analysis for grapevine breeding[44,48].

The ultimate goal of genomic breeding is to build superior cultivars by combining beneficial alleles underlying multiple agronomic traits of interest, while purging or hiding moderate-to-strong deleterious variants (including SNPs and SVs)[57,58]. Interestingly, we found strong genetic correlations among the 29 grape agronomic traits investigated, which allowed us to predict multiple traits during the same breeding cycle and decrease the time and monetary costs of breeding. In particular, SV-based genomic selection will play an excellent role in programs integrating multiple breeding traits. Overall, our Grapepan v.1.0, variation map, and the genomic prediction will greatly facilitate grapevine genomic breeding.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01967-5.

References

1. Myles, S. et al. Genetic structure and domestication history of the grape. *Proc. Natl Acad. Sci. USA* **108**, 3530–3535 (2011).
2. Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D. & Gaut, B. S. Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. *Proc. Natl Acad. Sci. USA* **114**, 11715–11720 (2017).
3. McGovern, P. et al. Early neolithic wine of Georgia in the South Caucasus. *Proc. Natl Acad. Sci. USA* **114**, E10309–E10318 (2017).
4. Freitas, S. et al. Pervasive hybridization with local wild relatives in Western European grapevine varieties. *Sci. Adv.* **7**, eabi8584 (2021).
5. Magris, G. et al. The genomes of 204 *Vitis vinifera* accessions reveal the origin of European wine grapes. *Nat. Commun.* **12**, 7240 (2021).
6. Xiao, H. et al. Adaptive and maladaptive introgression in grapevine domestication. *Proc. Natl Acad. Sci. USA* **120**, e2222041120 (2023).
7. Arroyo-García, R. et al. Multiple origins of cultivated grapevine (*Vitis vinifera* L. ssp. *sativa*) based on chloroplast DNA polymorphisms. *Mol. Ecol.* **15**, 3707–3714 (2006).
8. Dong, Y. et al. Dual domestications and origin of traits in grapevine evolution. *Science* **379**, 892–901 (2023).
9. Sivan, A. et al. Genomic evidence supports an independent history of Levantine and Eurasian grapevines. *Plants People Planet* **3**, 414–427 (2021).
10. Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
11. Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B. & Hirsch, C. N. How the pan-genome is changing crop genomics and improvement. *Genome Biol.* **22**, 3 (2021).
12. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
13. Gaut, B. S., Seymour, D. K., Liu, Q. & Zhou, Y. Demography and its effects on genomic variation in crop domestication. *Nat. Plants* **4**, 512–520 (2018).
14. He, Q. et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* **55**, 1232–1242 (2023).
15. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
16. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
17. Zhou, Y. et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).
18. Chen, S. et al. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. *Nat. Plants* **9**, 1986–1999 (2023).
19. Kou, Y. et al. Evolutionary genomics of structural variation in Asian rice (*Oryza sativa*) domestication. *Mol. Biol. Evol.* **37**, 3507–3524 (2020).
20. Munasinghe, M. et al. Combined analysis of transposable elements and structural variation in maize genomes reveals genome contraction outpaces expansion. *PLoS Genet.* **19**, e1011086 (2023).
21. Shi, T. et al. The super-pangenome of *Populus* unveil genomic facets for its adaptation and diversification in widespread forest trees. *Mol. Plant* **17**, 725–746 (2024).
22. Ramos-Madrigal, J. et al. Palaeogenomic insights into the origins of French grapevine diversity. *Nat. Plants* **5**, 595–603 (2019).
23. Calderón, L. et al. Diploid genome assembly of the Malbec grapevine cultivar enables haplotype-aware analysis of transcriptomic differences underlying clonal phenotypic variation. *Hortic. Res.* **11**, uhae080 (2024).
24. Massonnet, M. et al. The genetic basis of sex determination in grapes. *Nat. Commun.* **11**, 2902 (2020).
25. Vondras, A. M. et al. The genomic diversification of grapevine clones. *BMC Genomics* **20**, 972 (2019).
26. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
27. Shi, X. et al. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Hortic. Res.* **10**, uhad061 (2023).
28. Long, Q. et al. Population comparative genomics discovers gene gain and loss during grapevine domestication. *Plant Physiol.* **195**, 1401–1413 (2024).
29. Shang, L. et al. A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
30. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
31. Li, N. et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat. Genet.* **55**, 852–860 (2023).
32. Cochetel, N. et al. A super-pangenome of the North American wild grape species. *Genome Biol.* **24**, 290 (2023).
33. Tang, D. et al. Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).
34. Kang, M. et al. The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat. Commun.* **14**, 6259 (2023).
35. Porubsky, D. et al. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.* **8**, 1293 (2017).
36. Migicovsky, Z. et al. Patterns of genomic and phenomic diversity in wine and table grapes. *Hortic. Res.* **4**, 17035 (2017).
37. Flutre, T. et al. A genome-wide association and prediction study in grapevine deciphers the genetic architecture of multiple traits and identifies genes under many new QTLs. *G3 (Bethesda)* **12**, jkac103 (2022).
38. Guo, D.-L. et al. Genome-wide association study of berry-related traits in grape [*Vitis vinifera* L.] based on genotyping-by-sequencing markers. *Hortic. Res.* **6**, 11 (2019).
39. Zhang, C., Cui, L. & Fang, J. Genome-wide association study of the candidate genes for grape berry shape-related traits. *BMC Plant Biol.* **22**, 42 (2022).
40. Malabarba, J. et al. Manipulation of VviAGL11 expression changes the seed content in grapevine (*Vitis vinifera* L.). *Plant Sci.* **269**, 126–135 (2018).
41. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
42. Walker, A. R. et al. White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J.* **49**, 772–785 (2007).
43. Choi, S. W., Mak, T. S. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
44. Brault, C. et al. Across-population genomic prediction in grapevine opens up promising prospects for breeding. *Hortic. Res.* **9**, uhac041 (2022).
45. Lin, H. et al. Berry texture QTL and candidate gene analysis in grape (*Vitis vinifera* L.). *Hortic. Res.* **10**, uhad226 (2023).
46. Mejía, N. et al. Molecular, genetic and transcriptional evidence for a role of VvAGL11 in stenospermocarpic seedlessness in grapevine. *BMC Plant Biol.* **11**, 57 (2011).
47. Riaz, S., Tenscher, A. C., Ramming, D. W. & Walker, M. A. Using a limited mapping strategy to identify major QTLs for resistance to grapevine powdery mildew (*Erysiphe necator*) and their use in marker-assisted breeding. *Theor. Appl. Genet.* **122**, 1059–1073 (2011).

48. Schreiber, M., Jayakodi, M., Stein, N. & Mascher, M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat. Rev. Genet.* **25**, 577 (2024).
49. Cardone, M. F. et al. Inter-varietal structural variation in grapevine genomes. *Plant J.* **88**, 648–661 (2016).
50. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
51. Di Genova, A. et al. Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.* **14**, 7 (2014).
52. Maestri, S. et al. 'Nebbiolo' genome assembly allows surveying the occurrence and functional implications of genomic structural variations in grapevines (*Vitis vinifera* L.). *BMC Genomics* **23**, 159 (2022).
53. Slatkin, M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
54. Gabur, I., Chawla, H. S., Snowdon, R. J. & Parkin, I. A. P. Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.* **132**, 733–750 (2019).
55. Azuma, A. et al. Genomic and genetic analysis of Myb-related genes that regulate anthocyanin biosynthesis in grape berry skin. *Theor. Appl. Genet.* **117**, 1009–1019 (2008).
56. Carbonell-Bejerano, P. et al. Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiol.* **175**, 786–801 (2017).
57. Zhang, C. et al. Genome design of hybrid potato. *Cell* **184**, 3873–3883.e12 (2021).
58. Wu, Y. et al. Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313–2328.e15 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[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]College of Horticulture, Nanjing Agricultural University, Nanjing, China. [3]National Key Laboratory of Tropical Crop Breeding, Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, China. [4]Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou, China. [5]Institute of Horticultural Crops, Xinjiang Academy of Agricultural Sciences, Urumqi, China. [6]College of Horticulture, Qingdao Agricultural University, Qingdao, China. [7]State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. [8]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, USA. [9]These authors contributed equally: Zhongjie Liu, Nan Wang, Ying Su, Qiming Long, Yanling Peng, Lingfei Shangguan. ✉e-mail: fanggg@njau.edu.cn; xiaohua01@caas.cn; zhouyongfeng@caas.cn

Methods

Samples and genome sequencing

We collected nine grape accessions for genome assemblies, including seven *V. vinifera* ssp. *vinifera* (six Table1 grapes and one Wine grape), one *V. labrusca* × *vinifera* hybrid (Table2) and one *V. retordii* grape (Supplementary Table 1). We generated HiFi, Hi-C and ONT (Oxford Nanopore) ultra-long 150-kb reads for three grape accessions (Manicure Finger, Muscat Hamburg and Shine Muscat) and ONT ultra-long reads for the other six samples (the HiFi and Hi-C reads were obtained from other studies that have been published or will be published soon[59–61]).

We also obtained 11 previously published grape genome haplotypes from 9 grape accessions[10,24,25,27,62–64] (Supplementary Table 1). Notably, haplotype-resolved assemblies were achieved for Cabernet Sauvignon[63] and *V. arizonica* grape. These 18 grape accessions were utilized for core and/or variable gene family analysis, and 29 corresponding haplotypes were used to construct the graph pangenome.

We collected 324 modern cultivars of *V. vinifera* ssp. *vinifera*, including 106 Wine grapes, 108 Table1 grapes and 110 Table2 grapes, from the experimental orchard of Zhengzhou Fruit Research Institute of the Chinese Academy of Agricultural Sciences (Zhengzhou, China). The genomic DNA of these accessions was used for genome sequencing (15-fold coverage) (Supplementary Table 6). These newly sequenced data were used for subsequent analyses, including GWAS analysis, genomic prediction, heritability estimation and genetic correlation calculation. To improve population genome analysis, we also collected previously published resequencing data from 139 grape accessions and three outgroups[6,64–66]. These data were download from the National Center for Biotechnology Information (NCBI) database (Supplementary Table 6). Altogether, 466 grape varieties were used to investigate population structure, genetic diversity and genetic differentiation. This study provides a detailed description of the grape samples used for genome assembly and short-reads sequencing (Supplementary Notes).

De novo genome assembly

For nine genome assemblies, the HiFi and ONT reads with Hi-C reads were integrated for self-correction, trimmed and assembled using the Hifiasm[67] program (v.0.19.5-r587). For each accession, the contig-level assemblies were anchored and oriented to 19 chromosomes based on the reference-guided software RagTag[68] (v.1.0.1). The two sets of HiFi contigs were then validated, grouped, sorted and anchored with the Hi-C reads to generate two pseudochromosomes, by using Juicer (v.1.5)[69] and 3D-DNA (v.201008)[70]. In addition, we utilized Juicebox (v.1.11.08) to visualize and check the Hi-C data. We analyzed the gaps in these assembled haplotypes. These gaps were filled based on reads mapping and assembled scaffolds using the ONT data. We used NextDenovo[71] (v.2.5.0) to generate the ONT assemblies. We then used Minimap2 (v.2.26)[72] to map the ONT reads to the genome. Combined with Integrative Genomics Viewer visualization[73], we identified split reads and estimated gaps. The gaps were closed with ONT assemblies with HiFi reads polished. Therefore, we obtained 18 gap-free T2T haplotypes from 9 newly sequenced grapes. The completeness of genome assembly was checked according to the benchmarking universal single-copy orthologs score[74], and phase errors were verified by Hamming error and switch error. Genome heterozygosity was estimated with a *k*-mer-based approach using kmc[75] (v.3.2.2) and GenomeScope2.0 (ref. 76) programs.

The annotation of genome assemblies

We used the same pipeline to annotate gene structure and repetitive sequences in all 29 haplotypes. For gene annotation, expression data were collected from various tissues, including flowers, leaves, stems, roots and fruits (PRJNA565689 and PRJNA434655). To improve the annotation process, we used Hisat2 (v.2.10.2)[77] to align RNA sequencing reads against the repeat-masked assemblies. The mapping states were extracted using StringTie[78] (v.1.3.0). We mainly relied on customs scripts that integrated Braker[79] (v.3.0.2), PASA[80] (v.2.5.1) and MAKER[81] (v.3.01.03) programs for our genome annotations. We developed gene models and conducted subsequent searches utilizing AUGUSTUS[82] (v.3.4.0). Incomplete genes and low confidence gene structures were filtered based on hidden Markov models and Pfam[83] (v.1.6) database. TEs were identified using multiple combined programs. We generated the nonredundant TE catalogs using RepeatModeler[84] (v.2.0.4) based on all haplotypes. At the same time, we improved the nonredundant TE annotation based on the program EDTA[85] (v.2.0.1). RepeatMasker[86] (v.4.1.2) was used to execute homolog annotation.

Telomere repeat units were investigated utilizing TIDK (v.0.2.0). The entire genome underwent an in-depth search and we generated statistics pertaining to telomere regions and visualized the telomere peaks. To investigate tandem repetitive sequences, we used Tandem Repeats Finder[87] (v.4.09) to generate the statistics of the number of repeats and position information. These statistics were combined to annotate the centromeric repeats.

Comparative genomics

We performed genome alignments based on 27 haplotypes (18 newly generated and 9 previously published) from 18 representative grape accessions. We additionally collected one assembly from outgroup *Muscadinia rotundifolia*[88]. Haplotype sequences were aligned using SyRI[89] (v.1.6.3) and the resulting alignments visualized using Plotsr[90] (v.1.0.0). A set of homologous genes was estimated based on gene family clustering using Orthofinder[91] (v.2.5.2). The resulting statistical summaries were used for identifying the core and nonessential gene family sets.

Construction of the graph pangenome reference (Grapepan v.1.0)

To represent the genetic diversity of grapes, we constructed a graph pangenome that incorporated a total of 29 haplotype assemblies, including 18 newly sequenced haplotype assemblies from 9 phase-resolved accessions and 11 previously published assemblies based on continuous long read data (two phase-resolved and seven primary assemblies). Based on these genome assemblies, we used two tools to build graph pangenome, MC (v.2.6.11)[92] and PGGB (v.0.5.4)[93], respectively. We combined reads mapping and assembled alignments to validate the graph paths, edges and nodes in the grape pangenome. To address the issue of small fragments in assemblies, we implemented filtering steps to exclude minor, fragmented and diverse assemblies that could introduce the wrong structures.

Pangenome variation maps

The complex regions of the PGGB pangenome may introduce additional challenges and uncertainties during SV genotyping based on Illumina sequencing data. Therefore, we chose to use the MC pangenome (Grapepan v1.0) to generate the SV set for population-scale SV genotyping. We used the PNT2T assembly as the reference to order the position of variations. We deconstructed the MC pangenome to obtain an SV map. At the same time, we identified SVs by aligning assemblies using SVIM-asm[94] (v.1.0.3). To validate these pangenome SVs, we compared the SV generated by MC and SVIM-asm and obtained three validation metrics (precise, recall ratio and f1 score) using the program Truvari[95] (v.4.1.0). The high-quality pangenome SVs were filtered for downstream analysis. To performed GWAS using SVs in a large population, we genotype the pangenome SVs in 466 sequenced grape accessions. We used vcfbub (v.0.1.0) to filter SV, processing input variants. We used the re-genotyping tool PanGenie[50] (v2.1.0) to determine unique *k*-mers in the graph with PanGenie index. We then executed the PanGenie command on each accession and combined these VCF (variant call format) files to generate a merged SV map. We then distinguished biallelic and multiallelic variants. The resulting biallelic bubble-based VCF was used for downstream GWAS analysis, whereas the multiallelic bubble-based VCF was used for haplotype graph classification.

In addition, we generated pangenome SNPs and indels maps based on these sequenced grape accessions. We mapped 466 short reads to Grapepan v.1.0 using the Giraffe command from the vg program (v.1.51.0) and generated a BAM file for each accession. Small variants were identified by GTX (http://www.gtxlab.com/product/cat) (v.2.2.1) based on the BAM files. The decay of LD between SNPs and SVs was estimated using nonlinear regression of pairwise r^2. LD was calculated using PLINK[96] (v.3.31) and LD decay graphs were plotted using PopLDdecay[97] (v.3.26).

Population genetic analyses

We analyzed population structure based on 466 whole-genome sequencing data. The maximum likelihood (ML) tree was constructed using IQ-TREE[98] (v.1.6.619) based on the VT + F + R5 model. The reliability of the ML tree was estimated using the ultrafast bootstrap approach with 1,000 replicates. Figtree and an online tool iTOL (Interactive Tree of Life v.3, https://itol.embl.de) were used to display the ML tree. Population structure was analyzed using the ADMIXTURE[99] (v.1.3) program with a block-relaxation algorithm. To explore the convergence of individuals, we ran the cross-validation error procedure with K from 2 to 8. PCA was performed using PLINK (v.1.9). Two-dimensional PHATE embedding was generated using the first 20 principal components in the program phateR (v.1.0.7).

The genome divergence between different grape groups was estimate based on XP-EHH by using the selscan program[100] (v.2.0.0). To identify candidate regions potentially affected by selections, nucleotide diversity and population F_{ST} were calculated based on vcftools[101] (v.0.1.13) and genomics_general (https://github.com/simonhmartin/genomics_general).

Measurement of 29 agronomic traits

Phenotyping of traditional core-cultivated grapes was performed in the experimental fields of the Zhengzhou Institute of Fruit Trees (Zhengzhou, China) in the spring and summer of 2016 and 2017 (over two growing seasons). Three clones were investigated per grape accession. We investigated 29 agronomic traits in field experiments. The measurements and criteria for 21 phenotypes were based on standardized protocols from the International Organization of Vine and Wine, with minor adaptations made for certain traits as necessary. The remaining eight agronomic traits were defined based on our field experience. All specific methods for phenotypic measurement and definitions of phenotypes can be found in Supplementary Notes. Our 29 phenotypic traits included five phenotypic categories: six traits in bunch category (bunch shape, number of wings of the primary bunch, number of subsidiary bunches, bunch density, weight of a single bunch and ease of detachment from pedicel), eight traits in content category (SSC, Glu, Fru, Suc, TAC, Tar, malic acid and citric acid), eight traits in berry traits category (firmness of flesh, juiciness of flesh, particularity of flavor, number of seeds, length of seeds, berry shape, uniformity of time of physiological stage of full maturity of the berry and berry color), four traits in berry size category (BL, BeWi, BV and BeWe) and three traits in berry skin category (astringence of skin, thickness of skin and berry bloom). We used Scatterplot Matrix to plot the correlations between each two pairs of traits and highlight the significant differences.

Genome-wide association study

For two consecutive growing seasons of phenotype data, we conducted independent GWAS analysis and created a joint dataset by combining the GWAS results from both years. We have constructed a pangenome dataset encompassing SNPs and SVs. Based on these datasets, we independently conducted GWAS analyses: reference-based SNPs, pangenome-based SNPs and pangenome-based SNPs + SVs. PermGWAS (v.2023.05) was used to perform GWAS analysis of agronomic traits. To account for population structure, we implemented a mixed linear model approach to estimate heritability for each trait. The top five principal components have been shown to effectively capture population structure. To reduce the influence of overfit structure, we considered both fixed effects (using four PCA components to capture population structure) and random effects (using a kinship matrix to model individual covariance). Furthermore, we use permutation analysis to verify the result of each GWAS. The minimal P value from each permutation was used to calculate a permutation-based threshold using the maxT/minP multiple testing procedure.

The PVE of the most significantly associated single variant was estimated using the following formula[102]:

$$\mathrm{PVE} = [2\times(\mathrm{beta}^2)\times \mathrm{MAF}\times(1-\mathrm{MAF})]/\,[\,2\times(\mathrm{beta}^2)\times \mathrm{MAF}(1-\mathrm{MAF})$$
$$+((\mathrm{s.e.}\times(\mathrm{beta}))^2)\times 2\times N\times \mathrm{MAF}\times(1-\mathrm{MAF})\,]$$

Where N represents the sample size, s.e. is the standard error of the effect size of genetic variants, beta is the effect size of genetic variants, and MAF is the minor allele frequency of the target marker.

Heritability estimation

The genome SNP heritability ($G\_h^2_{\mathrm{SNP}}$) of a trait is the fraction of phenotypic variance explained by additive contributions from all pangenome SNPs. Similarly, we calculate the genome SV heritability ($G\_h^2_{\mathrm{SV}}$) using all pangenome SVs. The LDAK[103] (v.5.2) program provides a model for estimating $G\_h^2_{\mathrm{SNP}}$ and $G\_h^2_{\mathrm{SV}}$ by deriving approximate relationships between the heritability of a variant and MAF. Consider the heritability model of the form:

$$E[h_j^2] = \mathrm{tau}_i w_j [f_j(1-f_j)]^{(1+\mathrm{alpha})}$$

where w_j is the weighting for each SNP or SV (j), f_j is its MAF and tau are the corresponding coefficients and are estimated from the SNP or SV. The parameter alpha determines how the expected heritability contributed by a variant depends on its MAF (https://dougspeed.com/technical-details/).

Our approach involves filtering out genotyping error and nearly linked mutation sites, followed by utilization of the LDAK-Thin model. To improve robustness, we inferred $G\_h^2_{\mathrm{SNP}}$ and $G\_h^2_{\mathrm{SV}}$ based on a kinship matrix generated by pangenome SNPs or SVs. Because LDAK cannot directly estimate alpha, we adopt an alternative strategy involving a trial of multiple alpha values by comparing the highest log likelihood and identified value estimates from the Gaussian distribution.

Polygenic scores

To establish a connection between genetic variation and phenotype, we used a method that combines LD between loci and multiple genetic variations, calculating PGS through application of the LDpred2 program (v.1.12.4)[104]. To reduce the overfitting issues arising from excessive variable numbers, we initially clustered all genomic variations. We then partitioned the dataset into three subsets: a training dataset comprising 70%, a tuning dataset consisting of 20% and a test dataset constituting 10%. We used the training dataset to generate final summary statistics.

Transcriptome and GO enrichment analyses

To validate the candidate genes identified in our GWAS and divergent selection, we retrieved fruit expression data from two accessions at three distinct stages—green-hard, veraison and ripening—from the NCBI database (PRJNA565689), with each library consisting of three biological replicates. These sequenced reads were mapped to the PNT2T genome using the STAR program (v.2.4.2)[105]. Gene expression levels were quantified using featureCounts[106] (v.2.0.2) in terms of transcripts per kilobase of exon model per million mapped reads. Based on the GO database (http://geneontology.org/), we conducted the background GO-Terms using whole-genome proteins. GO and GSEA enrichment analysis was performed using clusterProfiler[107] (v.4.0) and a test threshold of false discovery rate (FDR) <0.05.

Statistical analysis

Details on all statistical analyses used in this paper, including the statistical tests used, the number of replicates and precision measures, are indicated in the corresponding figure legends. Statistical analysis of replicate data was performed using appropriate strategies in R (v.8.4.3).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All long-read and Hi-C sequencing data have been deposited in the NCBI database under the accession code BioProject: PRJNA1048106. All resequencing data generated have been deposited in the NCBI database under the accession code BioProject: PRJNA994294. All assemblies have been deposited under the accession codes BioProject: PRJNA1018808, PRJNA1018809, PRJNA1029477, PRJNA1029478, PRJNA1029479, PRJNA1029480, PRJNA1130629, PRJNA1130630, PRJNA1130639, PRJNA1130641, PRJNA1130642, PRJNA1130643, PRJNA1130644, PRJNA1130647, PRJNA1130648, PRJNA1130649, PRJNA1130650, PRJNA1130651. These data are also available at the National Genomics Data Center Genome Sequence Archive (https://ngdc.cncb.ac.cn/gsa/) with BioProject codes PRJCA024688 and PRJCA024753. The Grapepan v.1.0, all assembled genome sequences and annotations are available via Zenodo at https://doi.org/10.5281/zenodo.10851547 (ref. 108) and https://doi.org/10.5281/zenodo.10846425 (ref. 109).

Code availability

All scripts and codes associated with this project are available via GitHub at https://github.com/zhouyflab/GrapePan and Zenodo at https://doi.org/10.5281/zenodo.13308856 (ref. 110).

References

59. Wang, X. et al. Integrative genomics reveals the polygenic basis of seedlessness in grapevine. *Curr. Biol.* **34**, 3763–3777 (2024).
60. Zhang, T. H. et al. Population genomics highlights structural variations in local adaptation to saline coastal environments in woolly grape. *J. Integr. Plant Biol.* **66**, 1408–1426 (2024).
61. Zhong, H. et al. Haplotype-resolved assemblies provide insights into genomic makeup of the oldest grapevine cultivar (Munage) in Xinjiang. Preprint at *BioRxiv* https://www.biorxiv.org/content/10.1101/2024.09.11.612401v2 (2024).
62. Li, B. & Gschwend, A. R. *Vitis labrusca* genome assembly reveals diversification between wild and cultivated grapevine genomes. *Front. Plant Sci.* **14**, 1234130 (2023).
63. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
64. Morales-Cruz, A. et al. Multigenic resistance to *Xylella fastidiosa* in wild grapes (*Vitis* sps.) and its implications within a changing climate. *Commun. Biol.* **6**, 580 (2023).
65. Badouin, H. et al. The wild grape genome sequence provides insights into the transition from dioecy to hermaphroditism during grape domestication. *Genome Biol.* **21**, 223 (2020).
66. Ramos, M. J. N. et al. Portuguese wild grapevine genome re-sequencing (*Vitis vinifera sylvestris*). *Sci. Rep.* **10**, 18993 (2020).
67. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
68. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
69. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
70. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
71. Hu, J. et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).
72. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
73. Robinson, J. T., Thorvaldsdottir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **39**, btac830 (2022).
74. Simao, 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).
75. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
76. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
77. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
78. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
79. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, lqaa108 (2021).
80. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
81. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
82. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
83. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
84. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
85. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
86. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **5**, 4.10.1–4.10.14 (2004).
87. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
88. Park, M. et al. Chromosome-level genome sequence assembly and genome-wide association study of *Muscadinia rotundifolia* reveal the genetics of 12 berry-related traits. *Hortic. Res.* **9**, uhab011 (2022).
89. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
90. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).

91. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
92. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
93. Garrison, E. et al. Building pangenome graphs. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.05.535718 (2023).
94. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).
95. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
96. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
97. Zhang, C., Dong, S. S., Xu, J. Y., He, W. M. & Yang, T. L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
98. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
99. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
100. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).
101. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
102. Shim, H. et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS One* **10**, e0120758 (2015).
103. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).
104. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
105. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
106. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
107. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb.)* **2**, 100141 (2021).
108. ZhouLab. Grapepan v1.0. *Zenodo* https://doi.org/10.5281/zenodo.10851547 (2024).
109. Zhou lab. Haplotype-resolved telomere to telomere genomes and annotations for nine representative diploid grapes. *Zenodo* https://doi.org/10.5281/zenodo.10846425 (2024).
110. Liu Z. & Ying, S. lzjhehe/Grapepan: v1.0.0 (v1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.13308856 (2024).

Acknowledgements

We thank members of the Zhou lab at AGIS and Z. Zhang from Washington State University for discussion and comments on the project. This work was supported by the National Key R&D Program of China (grant no. 2023YFF1000100 to Y.Z.), the National Natural Science Foundation of China (grant no. 32372662 to Y.Z. and grant no. 32272647 to J.F.), the National Natural Science Foundation of China (Overseas) to Y.Z. and the Priority Academic Program Development of Jiangsu Higher Education Institutions (RAPD).

Author contributions

Y.Z. conceived of and designed the research. Zhongjie Liu, C. Liu, Y.L., M.G., X.F., L. Sun, L. Shangguan, M.D. and Y.R. participated in the material preparation and phenotype investigation. Z.J., X. Wu, C.Z., L. Shangguan, J.F. and D.P. provided metabolites data. Y.S., X. Wang, C.Z., S.C. and T.D. contributed to assembly and annotation. Zhongjie Liu, N.W., Y.S. and Q.L. constructed graph pangenome and detected genetic variations. G.H. and X.L. performed gene expression analysis. Zhongjie Liu, F.Z., Y.G., X.S., Zhenya Liu, H.Z., Y.P. and Y.W. contributed to the heritability estimation and association study. N.W. and Zhongjie Liu wrote the paper. Y.Z., B.G. and S.H. revised the paper with contributions from H. Xiao, J.C., H. Xue, S.H., Y.S., N.W., J.F., L. Shangguan, C. Li, B.A., Z.M., X.X., W.L. and X.C. All of the authors read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01967-5.

Correspondence and requests for materials should be addressed to Jinggui Fang, Hua Xiao or Yongfeng Zhou.

Peer review information *Nature Genetics* thanks Roberto Bacilieri and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): Dr. Yongfeng Zhou, Dr. Hua Xiao, Dr. Jinggui Fang

Last updated by author(s): Jul 1, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequencing platforms used to generate the raw data are listed as followed: PacBio SMRT, Oxford Nanopore, Illumina HiSeq, NovaSeq.
Data analysis	The softwares used in this manuscript include Hifiasm v0.19.5-r587, RagTag v1.0.1, Juicer v1.5, 3D-DNA v201008, Juicebox v1.11.08, NextDenovo v2.5.0, Minimap2 v2.26, Integrative Genomics Viewer v2.13.1, BUSCOs v5.2.2, kmc v3.2.2, GenomeScope2.0 v1.0.0, Hisat2 v2.10.2, StringTie v1.3.0, Braker v3.0.2, PASA v2.5.1, MAKER v 3.01.03, AUGUSTUS v3.4.0, RepeatModeler v2.0.4, EDTA v2.0.1, RepeatMasker v4.1.2, TIDK v0.2.0, Tandem Repeats Finder v4.09, SyRI v1.6.3, Plotsr v1.0.0, Orthofinder v2.5.2, Minigraph-Cactus v2.6.11, PanGenome Graph Builder v0.5.4, SVIM-asm v1.0.3, Truvari v4.1.0, PanGenie v2.1.0, vg v1.51.0, GTX v2.2.1, PLINK v3.31, PopLDdecay v3.26, IQ-TREE v1.6.619, ADMIXTURE v1.3, phateR v1.0.7, selscan v2.0.0, vcftools v0.1.13, PermGWAS v2023.05, LDAK v5.2, LDpred2 v1.12.4, STAR v2.4.2, featureCounts v2.0.2, clusterProfiler v4.0, calc_switchErr v1.0, Yak v0.1, WhatShap v2.0, CD-HIT v4.8.1, PHATE v1.0.11, BWA v0.7.17, seqkit v2.2.0, odgi v0.8.4, vcfbub v0.1.0, bedtools v2.30.0, SAMtools v1.3.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All long-read and Hi-C sequencing data have been deposited in the National Center for Biotechnology Information (NCBI) database under the accession code BioProject: PRJNA1048106. All re-sequencing data generated have been deposited in the NCBI database under the accession codes BioProject: PRJNA994294. All assemblies have been deposited under the accession codes BioProject: PRJNA1018808, PRJNA1018809, PRJNA1029477, PRJNA1029478, PRJNA1029479, PRJNA1029480, PRJNA1130629, PRJNA1130630, PRJNA1130639, PRJNA1130641, PRJNA1130642, PRJNA1130643, PRJNA1130644, PRJNA1130647, PRJNA1130648, PRJNA1130649, PRJNA1130650, PRJNA1130651. These data are also available at the National Genomics Data Center (NGDC) Genome Sequence Archive (GSA) (https://ngdc.cncb.ac.cn/gsa/) with BioProject codes PRJCA024688 and PRJCA024753. The Grapepan v1.0, all assembled genome sequences, and annotations have been deposited to Zenodo with the following DOIs: https://doi.org/10.5281/zenodo.10851547, https://doi.org/10.5281/zenodo.10846425.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No specific statistical methods were used to determine sample size for the experiments. 466 grape varieties representing the full range of phenotypic diversity and geographic distribution of grapes worldwide were collected, including varieties from China, Germany, France, Canada, Japan, the United States, and other countries. All of these genotypes fully reflect grape genetic diversity, were chosen for our mainstudy. We selected 9 representative accessions including all different Vitis morphotypes and some wild types for pan-genomeconstruction. All detailed information was provided with supplemental dataset and figures.
Data exclusions	No data were excluded from the analyses.
Replication	All experiments were performed once with at least three independent biological replicates. All replications were successful and were used.
Randomization	For each grape accession, the sampling process for genome DNA/RNA sequencing was randomly conducted
Blinding	Blinding is not necessary for genome sequencing and assembly, since the investigators know which grape accessions they were handing.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a | Involved in the study

☒ ☐ Antibodies
☒ ☐ Eukaryotic cell lines
☒ ☐ Palaeontology and archaeology
☒ ☐ Animals and other organisms
☒ ☐ Clinical data
☒ ☐ Dual use research of concern
☐ ☒ Plants

Methods

n/a | Involved in the study

☒ ☐ ChIP-seq
☒ ☐ Flow cytometry
☒ ☐ MRI-based neuroimaging

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No | Yes

☒ ☐ Public health
☒ ☐ National security
☒ ☐ Crops and/or livestock
☒ ☐ Ecosystems
☒ ☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No | Yes

☒ ☐ Demonstrate how to render a vaccine ineffective
☒ ☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒ ☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒ ☐ Increase transmissibility of a pathogen
☒ ☐ Alter the host range of a pathogen
☒ ☐ Enable evasion of diagnostic/detection modalities
☒ ☐ Enable the weaponization of a biological agent or toxin
☒ ☐ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	We collected all modern cultivars of V. vinifera ssp. vinifera from the experimental orchard of Zhengzhou Fruit Research Institute of the Chinese Academy of Agricultural Sciences (Zhengzhou, China).
Novel plant genotypes	No novel plant genotypes
Authentication	Each plant stock has authentication by Zhengzhou Fruit Research Institute of the Chinese Academy of Agricultural Sciences.