

Impacts of reproductive systems on grapevine genome and breeding

Received: 31 October 2023

Accepted: 31 January 2025

Published online: 03 March 2025



Hua Xiao^{1,9}, Yue Wang^{1,9}, Wenwen Liu^{1,9}, Xiaoya Shi^{1,2,9}, Siyang Huang^{1,9}, Shuo Cao¹, Qiming Long¹, Xu Wang¹, Zhongjie Liu^{1,3}, Xiaodong Xu¹, Yanling Peng¹, Pengfei Wang⁴, Zhonghao Jiang⁵, Summaira Riaz⁶, Andrew M. Walker⁶, Brandon S. Gaut⁷, Sanwen Huang^{1,8} & Yongfeng Zhou^{1,8}✉

Diversified reproductive systems can be observed in the plant kingdom and applied in crop breeding; however, their impacts on crop genomic variation and breeding remain unclear. Grapevine (*Vitis vinifera* L.), a widely planted fruit tree, underwent a shift from dioecism to monoecism during domestication and involves crossing, self-pollination, and clonal propagation for its cultivation. In this study, we discover that the reproductive types, namely, crossing, selfing, and cloning, dramatically impact genomic landscapes and grapevine breeding based on comparative genomic and population genetics of wild grapevine and a complex pedigree of Pinot Noir. The impacts are widely divergent, which show interesting patterns of genomic purging and the Hill-Robertson interference. Selfing reduces genomic heterozygosity, while cloning increases it, resulting in a “double U-shaped” site frequency spectrum (SFS). Crossing and cloning conceal while selfing purges most deleterious and structural burdens. Moreover, the close leakage of large-effect deleterious and structural variations in repulsion phases maintains heterozygous genomic regions in 4.3% of the grapevine genome after successive selfing for nine generations. Our study provides new insights into the genetic basis of clonal propagation and genomic breeding of clonal crops by purging deleterious variants while integrating beneficial variants through various reproductive systems.

Diversified reproductive systems can be observed in plant kingdoms and applied for crop breeding, including sexual crossing, self-pollination, and asexual clonal propagation. Crossing involves the deliberate combination of distinct varieties or even species to generate new

genetic combinations. These hybrid offspring often exhibit heterosis, leading to enhanced yield and quality¹. Self-pollination (selfing) includes pollination within the same plant or variety. This process promotes homozygosity and may aid in the elimination of recessive

¹National 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.

²College of Enology, Heyang Viti-Viniculture Station, Ningxia Helan Mountain's East Foothill Wine Experiment and Demonstration Station, Northwest A&F University, Yangling, China. ³Institute of Life and Health, China Resources Research Institute of Science and Technology, Hong Kong, China. ⁴Shandong Academy of Grape, Jinan, China. ⁵Guangdong Provincial Key Laboratory for Plant Epigenetics, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, China. ⁶San Joaquin Valley Agricultural Center, United States Department of Agriculture, Parlier, CA, USA. ⁷Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, USA. ⁸National Key Laboratory of Tropical Crop Breeding, Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, China. ⁹These authors contributed equally: Hua Xiao, Yue Wang, Wenwen Liu, Xiaoya Shi, Siyang Huang. ✉e-mail: zhouyongfeng@caas.cn

deleterious mutations through purging^{2,3}. Nevertheless, selfing can lead to inbreeding depression, resulting in compromised growth potential, reduced yield, and diminished biotic and abiotic resistances⁴. The third reproductive type, clonal propagation, avoids the introduction of additional genetic material and offers the advantage of preserving favorable agronomic traits present in specific varieties⁵. Clonal propagated lines can, however, accrue somatic mutations. Some somatic mutations have produced new breeding resources^{6–8}, such as bud sport varieties⁹, but somatic mutations can also be deleterious and negatively influence plant fitness^{10–12}. These plant reproductive systems can shape the landscapes of genomic variation both within and among populations¹³, but more empirical work is necessary to characterize the magnitude of the effects of reproductive systems on genomic landscapes¹³. Such investigations could facilitate advancements in crop genomic breeding^{14,15}.

Grapevine (*Vitis vinifera* L.) is one of the most widely planted and economically important perennial fruit trees. There are two subspecies of grapevine: domesticated *V. vinifera* subsp. *vinifera* and its wild progenitor *V. vinifera* subsp. *sylvestris*^{16–18}. The initial steps of *V. vinifera* domestication may date as far back as ~15,000 years ago, and it included a shift from a dioecious, obligately outcrossing mating system to monoecy, with the potential for self-pollination^{18,19}. Despite this shift, the cultivation of domesticated grapes has primarily been based on hybridization and on clonal propagation of highly heterozygous genotypes, in part due to substantial inbreeding depression for selfed materials²⁰. In some cases, clonal propagation has been prolonged; some modern cultivars possess genetic information identical to grape seeds from the medieval period²¹. Thus, the evolutionary history and breeding of domesticated grapes have been influenced by diverse reproductive systems, significantly shaping their genomic makeup²⁰.

Pinot Noir (PN) is a premium red wine variety popular for its flavorful, aromatic wines. With a history spanning over a millennium, PN has served as a focus of research and as a foundational parent in grapevine breeding. In its over 900 years of clonal propagation²¹, numerous variants have been derived from PN through bud sports, including Pinot Grigio, Pinot Meunier, Pinot Gris, and Pinot Blanc²². The initial grapevine reference genome was determined from a highly homozygous descendant of PN called PN40024²³. This accession originated from successive self-pollinations of Helfensteiner (HE)²⁴, an offspring resulting from a cross between PN and Schiava Grossa (SG) performed in 1931 (www.vivc.de). Recently, a complete telomere-to-telomere (T2T) genome assembly of PN40024 has also been made available²⁵.

In this study, we construct a haplotype-resolved chromosomal genome of PN and characterize the haplotypic diversity present within it, including significant structural variations and gene families unique to each haplotype. Additionally, we examine both somatic and fixed variants within the PN population to enhance our understanding of the cultivar's development. Our analysis of various grapevine germplasm samples indicates that reproductive types have a substantial impact on genomic landscapes and grapevine breeding practices. These effects differ in terms of genome heterozygosity, as well as deleterious and structural burdens. We find that selfing significantly purges heterozygous deleterious SNPs (dSNPs) and structural variants (by 62% and 65%, respectively, compared to HE) in PN40024. The close linkage of large-effect deleterious and structural variations in repulsion phases maintains 4.3% of the genomic regions in a heterozygous state even after successive selfing. Our study explores the evolutionary genomics underlying the transitions of reproductive systems in forming grapevine lineages, which sheds insights on genomic breeding of grapevine.

Results

Comparative genomics of clonal Pinot Noir and selfing PN40024

To construct the haplotype-resolved PN genome, we generated a total of 33 Gb (33,349,412,693 bp, ~65x coverage) of HiFi reads and 84 Gb

(83,958,664,800 bp, ~160x coverage) of Hi-C reads (see “Methods”). Two haplotypes of PN were assembled: PN1 (496.43 Mb) and PN2 (489.78 Mb) (Fig. 1A), with a contig N50 of 23.60 Mb and 24.19 Mb, respectively. The BUSCO score of PN1 (98.3%) and PN2 (98.4%) was comparable to the complete assembly of PN40024 (PN40024_T2T, 98.5%, Supplementary Fig. 1). These statistics suggested that employing a similar methodology as PN40024_T2T²⁵, all centromeres and most of the telomeres in the two haplotypes' genome were identified. For genome annotation, PN1 (37,038) and PN2 (37,350) yielded a similar number of identified genes. We also identified repetitive sequences representing 66.64% and 66.21% of the PN1 and PN2 assemblies, respectively (Supplementary Table 1). Interestingly, we detected a large inversion on chromosome 19 between PN1 and PN2, which is almost 4 Mb in size (Fig. 1A and Supplementary Fig. 2). Further analysis of long reads, complemented by PCR analyses, supported the inversion between the two haplotypes at the breakpoints (Supplementary Table 2, Supplementary Figs. 3–5).

To study the impacts of reproduction systems on genomic landscapes, we also conducted comparative genomics among PN1, PN2, and PN40024_T2T (Fig. 1B). *K*-mer analyses estimated the genome heterozygosity to be 1.43% in PN (Supplementary Fig. 1), which is higher than PN40024_T2T (0.18%)²⁵. The three genomes were highly diverged at the sequence level, including gene content. If no homologous genes were identified for a gene family within a genome, that gene family was considered absent from that genome. Using this approach, we identified 20,800 shared gene families among PN40024_T2T, PN1, and PN2. However, we also detected extensive variation at gene contents: 2869, 2864, and 3026 gene families were exclusive to PN1, PN2, and PN1+PN2, respectively, compared to PN40024_T2T, while smaller numbers (2243, 2214 and 581 gene families) were exclusive to PN40024_T2T compared to PN1, PN2 and PN1+PN2 (Fig. 1C), respectively. These observations could have two causes. First, it is possible that the SG, parent of HE, had fewer gene families than the PN parent. An alternative hypothesis is that selfing led to the loss of many gene families, potentially through a process that favors shorter haplotypes, as suggested for the rapid loss of genome size in selfed maize lineages³.

Selection and introgression shaped the characteristics of Pinot Noir

To assess evolutionary processes contributing to the formation of distinguishing characteristics of PN, we gathered resequencing data from 38 *V. vinifera* samples (Supplementary Data 1 and 2), including 18 PN clones (nine generated in this study and nine from previous publications), 20 previously published wild grapes accessions (ten from Europe (EU) and ten from the Middle East and Caucasus region (ME), respectively) and three muscadine grapes used as outgroup. These sequences constitute three groups with which to investigate introgression and selection signals in PN. After SNP and SV calling and filtering, we counted 4,687,377 SNPs and 18,469 SVs across the entire dataset. Genome-wide selection signatures were observed throughout the PN clones, especially strong peaks on chromosomes 1, 3, 4, 5, and 18, by applying the population branch statistic (PBS)²⁶ with ME and EU populations as controls to identify genomic regions with exaggerated divergence relative to the controls (Fig. 2A). Gene set enrichment analysis (GSEA)²⁷ revealed an enrichment of organelle assembly, glycerolipid metabolic process, regulation of protein metabolic process, response to auxin, response to abiotic stimulus and beta-glucan biosynthetic process, and so on (Fig. 2B, Supplementary Data 3). Note that the PBS is likely identifying extended lengths of the branch leading to PN, and thus it is likely that the inferred selection events occurred prior to the diversification of PN clonal lineages.

As EU wild population has a significant contribution to the origin of modern wine grapes²⁸, we explored how such introgression events shaped the genome of PN clones using *f*-statistic (f_d)²⁹. Although the f_d

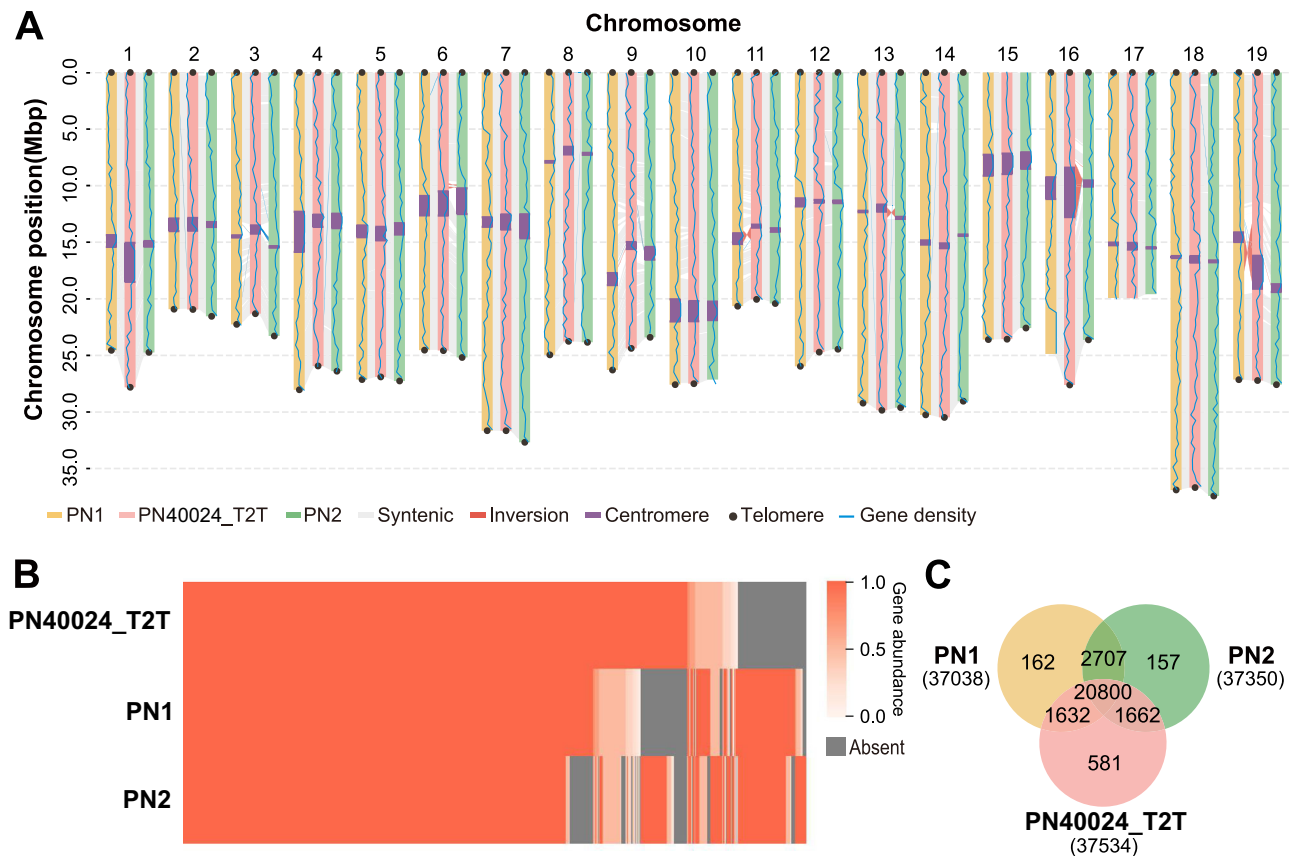


Fig. 1 | Haplotype-resolved assemblies of the clonal Pinot Noir genome.

A Collinearity between the reference genome PN40024_T2T and the two haplotypes of Pinot Noir: haplotype 1 (PN1) and haplotype 2 (PN2). Collinear regions are indicated by gray lines. **B** Presence and absence of gene families in the PN and PN40024_T2T genomes. The absence of a gene family in a genome is indicated by

gray coloring. The red gradients represent gene abundance within each gene family when present in the genome. **C** Shared and unique gene families among each genome assembly. The numbers in parentheses indicate the total number of genes identified within each assembly.

-test is primarily used in sexual populations, it focuses on shared alleles rather than allele frequencies. Therefore, it could be an efficient way for detecting introgressed fragments in clonal PN. By using ME as the sister population, we detected strong introgression signals in the PN genomes from EU, particularly on chromosomes 1, 2, 3, and 19 (Fig. 2C). The genes in the top 1% of regions with the highest f_d values were enriched in pathways related to plant physiological processes, such as photosynthesis and generation of precursor metabolites and energy (Supplementary Fig. 6, Supplementary Data 4). One locus, *CONSTANS-like 5* (*COL5*), related to plant flowering (Supplementary Fig. 7)³⁰, stood out as one of the top outlier in introgression analysis. Both the nucleotide diversity (π) of PN clones and sequence similarity (D_{xy}) between PN clones and EU were significantly lower at *COL5* and surrounding regions (Fig. 2D). These observations strongly suggest that both selection and introgression shaped *COL5* in PN clones, conferring local adaptation to adjust flowering time in new climates after the spread of cultivars globally after domestication. We further validated the reliability of this introgression in PN by conducting a phylogenetic analysis (Supplementary Fig. 8, Supplementary Data 2). The results indicated that ME and EU were mainly grouped into two separate clades. Instead of forming an independent clade, some cultivars, including PN, clustered with the EU clade, suggesting an introgression from EU in certain wine grapevines in this region.

Germline and somatic mutations in Pinot Noir clonal lineage

To investigate the impact of clonal propagation on PN, we analyzed genetic variants among the PN clones and their genetic differentiation from other grapevine populations. In total, 3,915,920 SNP variants and

17,035 SV variants were identified in grapes (PN, EU, and ME populations) in the absence of the outgroup data (used the inferred ancestral alleles, Fig. 3A). Among them, 70,291 SNPs and 605 SVs were unique to PN (specific variants), while 2,227,889 SNPs and 8635 SVs were shared with wild EU, ME or both populations, which are primarily likely to be germline mutations (Fig. 3A). The frequency of both shared SNPs and specific SNPs in PN clones displayed a “double U-shaped” site frequency spectrum (SFS) (Fig. 3B). However, the distribution of PN-specific SNPs showed an excess of rare variants (16.1% compared to 11.1% of all PN SNPs that were observed only once in PN clones) and a likely heterozygous state (69.2% compared to 21.1% of all PN SNPs that were observed 18 times across 18 PN clones), suggesting that many of these PN-specific variants are mostly somatic mutations in heterozygous states (Fig. 3B).

Given the duration of its cultivation, there has been ample opportunity for somatic mutations to occur in different clonal lineages of PN. To assess the number and type of these mutations, we divided the PN-specific SNPs into two groups: (i) 20,328 mutations that vary among PN clones (28.9%), which reflect the accumulation of somatic mutations and (ii) mutations that were observed across PN clones (71.1%), representing germline mutations that occur during the formation of the PN cultivar prior to the diversification of the clonal lineages, or somatic mutations fixed during clonal propagation (Fig. 3C). We defined the first group as “somatic mutations” and the second as “fixed mutations (including both somatic and germline mutations)”. Previous studies identified somatic mutations across clones of other varieties^{31–39}, but most of them failed to distinguish somatic and germline mutations. We applied control populations to

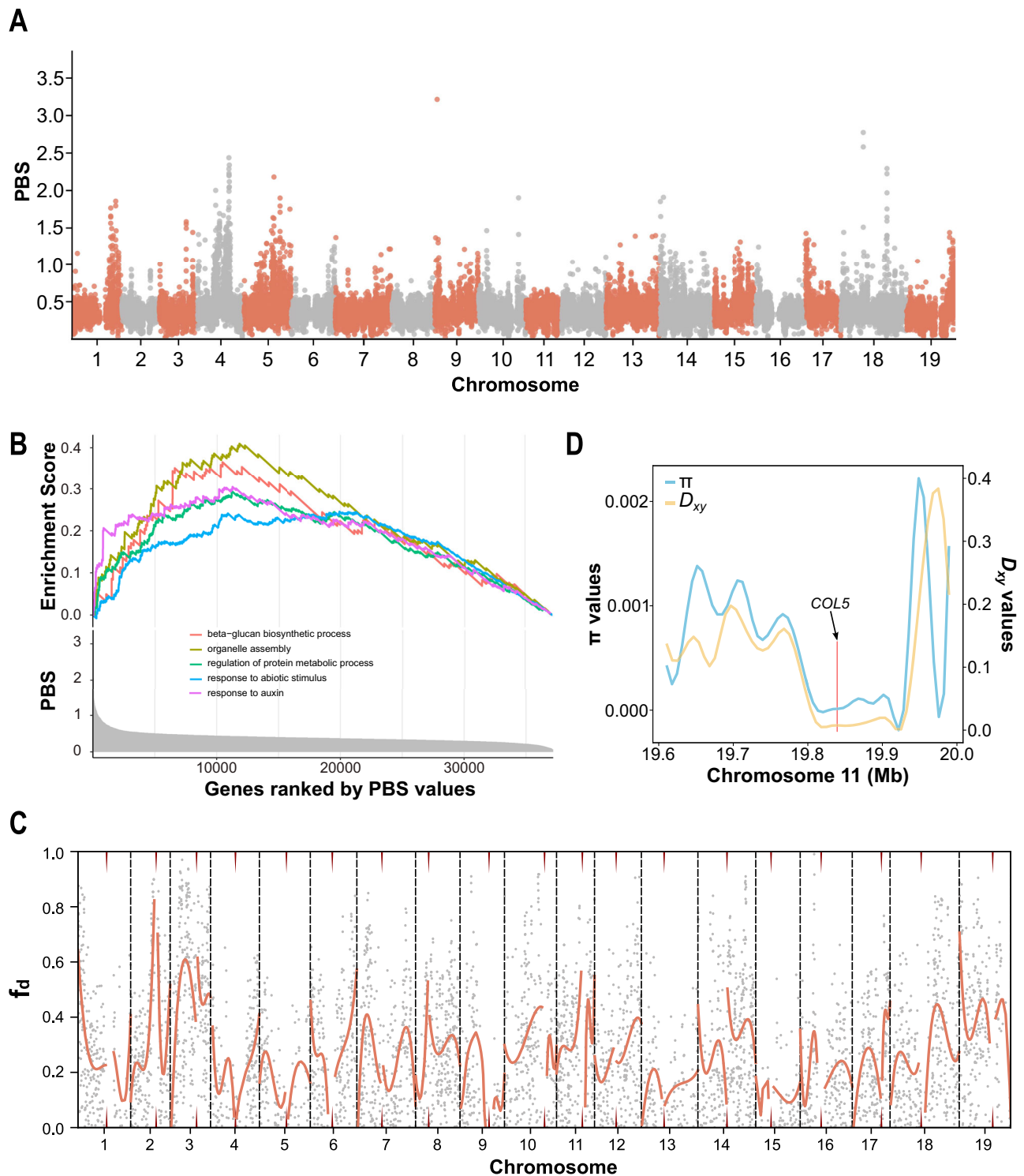


Fig. 2 | Signatures of introgression and selection in the Pinot Noir genome.

A PBS analysis of the PN group, utilizing the ME and EU groups as controls. **B** Gene set enrichment analysis for genes ranked by PBS values. Five biological processes are shown. **C** f_d analysis for the PN groups. **D** π and D_{xy} values for the region on

chromosome 11 that contains the *CONSTANS-like 5 (COL5)* gene. The red vertical line indicates the location of the *COL5* gene. Source data are provided as a Source Data file.

detect cultivar-specific mutations, which greatly improved the precision of classifying somatic and germline mutations (Supplementary Fig. 9). Interestingly, almost all (98.6%) of these fixed mutations were maintained as heterozygous in the 18 PN clones. Additionally, 55.5% of the remaining putative somatic SNPs were unique to a single individual, suggesting the possible accumulation of somatic mutations in each clone (Fig. 3C). We randomly selected ten somatic mutation sites

specific to one individual by examining the reads in IGV. The results showed that the mutation existed only in some reads of the individual predicted to have the mutation, and not in other individuals (Supplementary Figs. 10–14). The distribution of these somatic and fixed SNPs was then investigated in terms of gene structure. Somatic SNPs were found to occur more frequently in intergenic regions compared to fixed SNPs (Supplementary Fig. 15). A similar pattern has

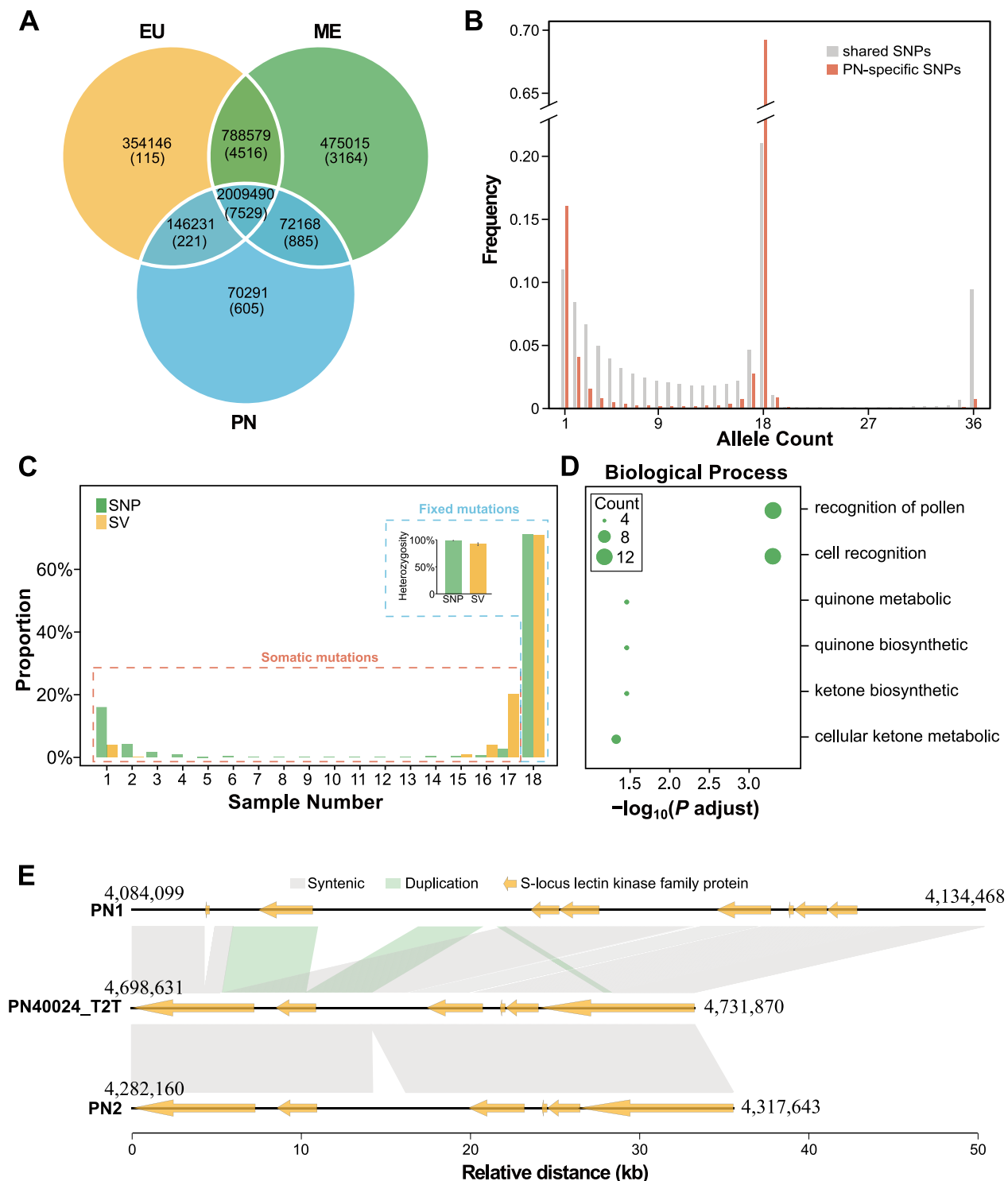


Fig. 3 | Somatic SNP and SV variants in clonal Pinot Noir. **A** A Venn diagram depicting the overlap of SNPs and SVs among the EU, ME, and PN groups. The number of SVs is indicated in parentheses. **B** Frequency of allelic variants in clonal Pinot Noir. The red segments represent the 70,291 SNPs specific to PN, as identified in the Venn diagram, while the gray segments indicate the 2,227,889 SNPs shared between PN and wild grapes. **C** The distribution of PN-specific SNPs and SVs among individuals. The bar chart displayed inside shows the average heterozygosity of

germline variants (SNPs, $n = 49,963$, and SVs, $n = 429$) in PN grapes. Data are presented as mean \pm SD. **D** GO term enrichment analysis of genes associated with PN-specific SVs. **E** A comparison of the genomes of PN and PN40024, focusing on the SVs present in PN but absent in wild grapes (EU and ME), located ~4.7 Mb on chromosome 10. The arrows on the genomes indicate the direction and location of genes belonging to the S-locus family in these regions.

been observed across clones of the Zinfandel variety³¹. GO analysis showed that genes with fixed SNPs were enriched in biological processes such as cell cycle process, negative regulation of biological process, cytoskeleton organization, microtubule cytoskeleton

organization, mRNA splicing via spliceosome, and so on (Supplementary Data 5).

Similar to SNPs, we also identified 605 SVs unique to PN by comparing SVs among the three populations (Fig. 3A). The SVs were

categorized into four types: deletion (DEL), tandem duplication (DUP), inversion (INV), and inter-chromosomal translocations (BND). DEL was the most prevalent type, accounting for 59.7%, followed by BND and DUP, while INV was the least common, constituting only 1.82% (Supplementary Table 3). We found that 87.4% of the SVs contained TEs. Similar to SNPs, among the 605 SVs unique to PN, 70.9% of them were shared by all 18 individuals, and the majority (91.9%) of these “fixed” SVs were present in a heterozygous state (Fig. 3C). Additionally, 176 variants were found in 1 to 17 individuals, with an average heterozygosity of 99.8%. For these fixed SVs that were retained in all 18 individuals, we annotated the genes on them and found they were enriched in biological processes related to recognition of pollen, cell recognition, ketone biosynthetic process, quinone metabolic process, quinone biosynthetic process, and cellular ketone metabolic process (Fig. 3D, Supplementary Data 6). We further zoomed in on the twelve genes associated with the pollen recognition process and found all of them belonging to the S-locus or related genes and hence probably related to self-incompatibility⁴⁰ (Supplementary Data 7). Nine of the twelve genes were clustered in one SV on chromosome 10 (~4.7 Mb), which was present in PN clones but absent in EU and ME *sylvestris* samples. The region containing this SV was highly heterozygous in all PN lineages (Supplementary Fig. 16). Comparative genomics between the PN haplotypes and PN40024_T2T genomes indicated that both duplication and insertion contributed to the formation of this SV on PN1 (Fig. 3E).

The impact of clonal propagation on grapevine genomes

To investigate the impact of various reproductive systems on grapevine breeding, we gathered more resequencing samples from the PN lineages with different reproductive modes for analysis, including five SG clones, four Gouais Blanc (GB) clones, ten Chardonnay (CD) clones, two Gamay Noir (GN) clones, two HE clones, and four PN40024 clones (Supplementary Data 2). The phylogenetic tree showed that the wild and domesticated grapes display reciprocal monophyly (Fig. 4A). Admixture analyses indicated HE, GN, and CD with admixture components from their parents (Fig. 4A). In contrast, the PN40024 clones were inferred to be separate, non-admixed group, perhaps reflecting distinctness evolved during selfing. Kinship analysis was conducted to verify the relationship between samples in each group and to confirm the true-to-type of these samples⁴¹. As expected, PN and GB were identified as parents of CD and GN, and PN and SG were identified as the parents of HE but not PN40024 (Fig. 4B).

We assessed genetic variation within each group to elucidate the impact of reproductive modes on genetic diversity. To begin with, we calculated the nucleotide diversity (π) value and observed heterozygosity (H_o) for each group. As expected, selfed PN40024 exhibited the lowest levels of genetic diversity ($\pi = 0.0003$) and heterozygosity (observed heterozygosity, $H_o = 0.01$) compared to all other groups. Among the groups, the wild grape ME group had the highest nucleotide diversity ($\pi = 0.0039$), while the other clonal grapes had π values ~ 0.002 (e.g., PN: 0.0024, GB: 0.0024, SG: 0.0019, CD: 0.0020, GN: 0.0028 and HE: 0.0029) (Supplementary Fig. 17), reflecting the effects of the domestication bottleneck. However, the average heterozygosity is slightly higher in clonal groups than in wild groups (EU: 0.20, and ME: 0.25, PN: 0.28, GB: 0.27, SG: 0.26, CD: 0.30, GN: 0.28, Supplementary Fig. 18).

The reproductive systems directly affect the efficiency of recombination, which is associated with the purging and maintenance of genetic burden^{42,43}. To detect such effects, we identified dSNPs using SIFT⁴⁴. The recessive burden for each individual was measured by the number of homozygous dSNPs; the heterozygous burden was measured by the number of heterozygous dSNPs; while the additive burden was calculated as the number of heterozygous dSNPs plus two times the number of homozygous dSNPs⁴⁵. First, we examined the dSNPs shared by all individuals within each grape group and found that

the clonal groups (PN, SG, GB, CD, GN, and HE) exhibited significantly higher genetic burden compared to the outcrossing wild groups (EU and ME) (Fig. 4C). The selfing PN40024 had the highest recessive burden and lowest heterozygous burden. In contrast, the wild grape groups (EU and ME) showed a significantly higher genetic burden of dSNPs that were not shared by all individuals within each grape group (Supplementary Fig. 19). A similar pattern was also observed for SV among the different grape groups (Fig. 4D and Supplementary Fig. 20).

Conserved heterozygous regions in grapevines

To understand the dynamics of alleles among cultivars during the breeding process, we assessed segregating sites that consistently occurred in each grape group. As shown in Fig. 4E, each column in the bar plot represents a specific genotype combination among the grape groups, with a total of 381 combinations observed. We selected 167 combinations that include more than 30 sites for subsequent analyses (Fig. 4E). The categories with the most sites were the SNPs that were heterozygous in clonally propagated PN, SG, or HE but homozygous in PN40024 due to selfing, and the SNPs that were only heterozygous in the GB sub-lineage (including GB, CD, and GN clones). A total of 97 SNPs remained heterozygous in all grape samples during domestication and diversification (wild and cultivated grapes including PN40024) and were dispersed throughout the genome. We identified 14 genes enriched in defense response and in the tricarboxylic acid cycle with SNPs located on the genes or within 2 kb up- and downstream of the genes (Supplementary Data 8). In addition, most of the homologous genes of the 14 genes were related to plant growth, and some of them are associated with recessive lethal phenotypes (Supplementary Data 9). In addition, after nine generations of selfing to generate PN40024, 99.8% of the genome was homozygous²⁵, except for 4782 conserved heterozygous SNPs (hSNPs) (Fig. 4E). GO analysis for the 577 genes that were associated with these hSNPs was enriched in only one molecular function process called chitin binding (Supplementary Data 10).

To study the effects of recombination on genome architecture and genetic burden, we identified the specific markers (see “Methods”) between PN and SG, and inspected their distribution on HE and PN40024. As shown in Fig. 5A and Supplementary Fig. 21, most of the specific markers of PN and SG could be located on the HE genome in a syntenic region. Almost all these specific markers (98.1%) were in a heterozygous state. This result supported the fact that HE is the offspring of a PN by SG cross. After continuous selfing in PN40024, 91.2% of these specific markers became homozygous, while 8.8% of them were still heterozygous (Fig. 5A, Supplementary Fig. 22). Recombination between the PN and SG haplotypes was observed on PN40024 chromosomes. For example, on chromosome 1, the end of the long arm of the SG haplotype was replaced by PN-specific markers. Most detectable recombination events occurred at the ends of chromosomes (Supplementary Fig. 22).

Close linkage of deleterious and structural variants in repulsion phases after successive selfing

Although almost the whole PN40024 genome is homozygous, some heterozygous variants (116,199 SNPs and 1125 SVs) remained in all four PN40024 clones, and most of them were clustered in blocks (Fig. 5B, Supplementary Figs. 22 and 23)²⁵. We identified six large heterozygous blocks (see “Methods”) (Supplementary Table 4). To mitigate the influence of potential assembly errors and structural deviations from the reference genome, we assessed the read coverage of each sample within the heterozygous blocks relative to the read coverage on the whole genome (Supplementary Data 11). We observed that the average read depth in the heterozygous block of chromosome 16 (coverage: 12,992,557–13,599,104) is 2.13 times higher than that outside this block, likely due to erroneous assembly or structural disparities with the reference genome. Therefore, this block was excluded from

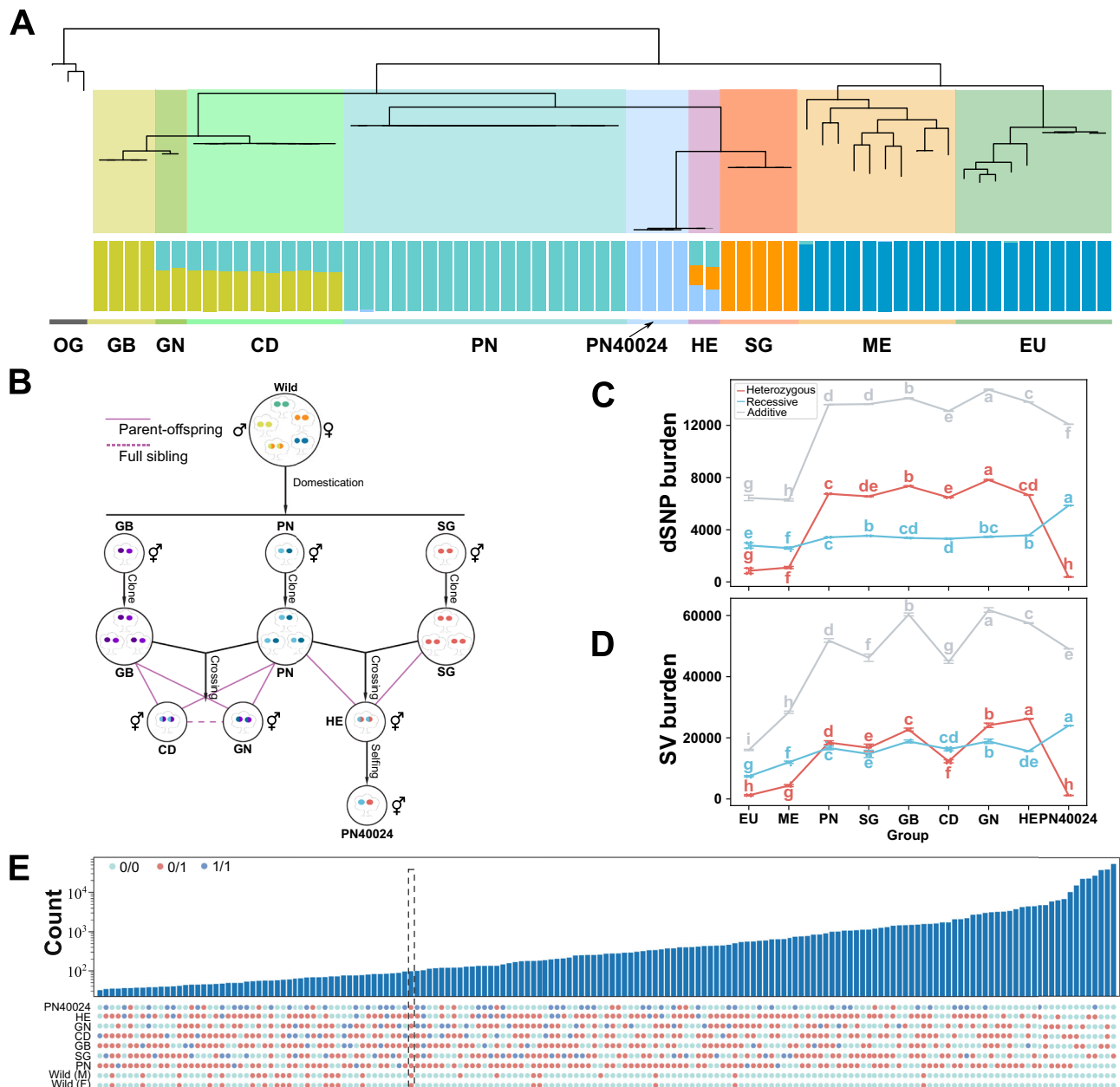


Fig. 4 | Genetic relationship among PN lineages. A Phylogenetic tree with admixture analysis. **B** Pedigree relationships among populations, highlighting three mating histories. **C** Genetic burden of dSNPs shared within each group. The sample sizes for each group, from left to right, are as follows: 10, 10, 18, 5, 4, 10, 2, 2, and 4. All data are presented as mean \pm SD for each group, with individual data points represented by dots. The two-sided Least Significant Difference (LSD) test was used for statistical analysis to compare each group; groups sharing the same letter are not significantly different ($P < 0.05$). **D** Genetic burden of SVs shared within each

group. The sample sizes for each group, from left to right, are as follows: 10, 10, 18, 5, 4, 10, 2, 2, and 4. All data are presented as mean \pm SD for each group, with dots indicating individual data points. The two-sided Least Significant Difference (LSD) test was used for statistical analysis to compare each group; groups with the same letter are not significantly different ($P < 0.05$). **E** Distribution of conserved SNPs with different genotype combinations among cultivars. The dashed black box highlights the variants that remain heterozygous across all grape populations.

subsequent analysis. The remaining five blocks account for ~4.3% of the genome and contain ~93% of the hSNPs and ~90% of the SVs in the entire genome.

To find out why PN40024 retained heterozygous blocks, we compared the distribution pattern of variants and found that the distribution pattern of SVs and dSNPs in heterozygous blocks of PN40024 was inherited from HE (Fig. 5B, C, Supplementary Figs. 24–29). In addition, the heterozygous blocks overlapped with regions of low recombination (Fig. 5A, Supplementary Figs. 21 and 22). According to the heterozygous blocks on PN40024, we divided the genome into two parts: the heterozygous block regions and genomic regions outside of

them. We found that there were more hSNPs located on genes and the flanking region of genes (2 kb upstream and downstream) in the heterozygous block regions than in regions outside of them in HE (Supplementary Fig. 30). Previous work has shown that heterozygous blocks in selfing lineages are caused by dSNPs in repulsion in low recombination regions^{3,46}. We assessed this phenomenon in PN40024 by assessing GERP values for dSNPs. Remarkably, dSNPs in heterozygous regions exhibited significantly rightward shifts in predicted deleterious effects relative to dSNPs outside of heterozygous regions, suggesting the presence of numerous large-effect deleterious mutations within heterozygous blocks ($P < 0.001$, Wilcoxon rank sum

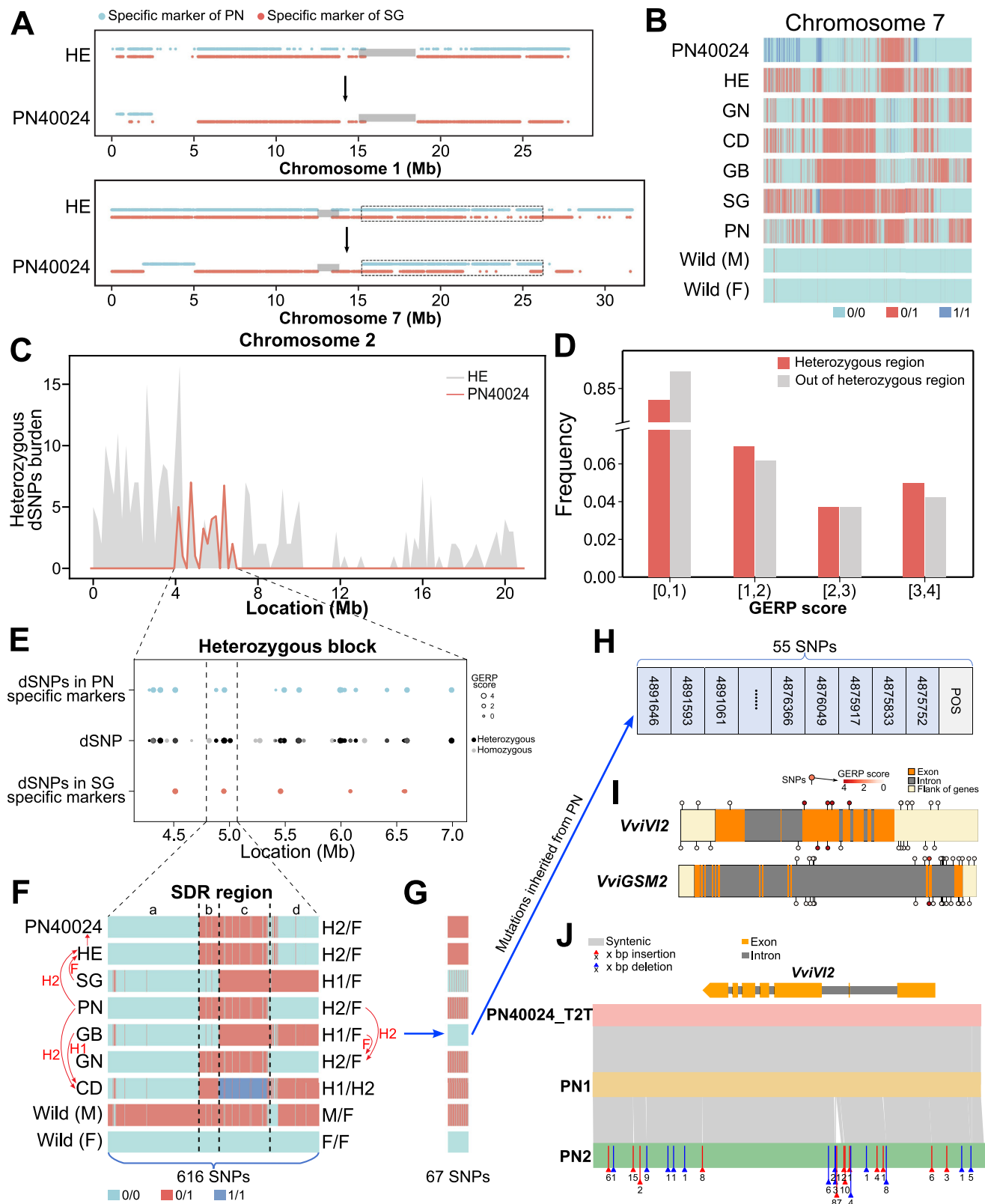


Fig. 5 | Heterozygous regions on the PN40024 genome and their genetic burden. **A** Distribution of PN- or SG-specific markers on HE and PN40024 across chromosome 1 and chromosome 7. **B** Genotypes of conserved sites in each cultivar on chromosome 7. **C** Density of heterozygous dSNPs in HE and PN40024 on chromosome 2. **D** Frequency spectrum of GERP values in heterozygous blocks compared to regions outside of these blocks. **E** Distribution of dSNPs and dSNPs originating from PN or SG based on the specific markers within the heterozygous

block on chromosome 2 of PN40024. The size of the circles represents GERP scores. **F** Genotypes of conserved sites in the SDR region. **G** The 67 heterozygous sites in PN40024 that do not exhibit both 0/0 and 1/1 genotypes simultaneously in other grape varieties. **H** Out of the 67 SNPs, 55 are identified as originating from PN. **I** Distribution of the 55 SNPs across two genes. **J** Comparison between PN40024 and Pinot Noir assemblies around the *VviV12* gene.

test, Fig. 5D). We then identified dSNPs in PN40024 that were inherited from both PN and SG in the heterozygous blocks, which suggested that dSNPs are probably located on both haplotypes of these blocks (Fig. 5E and Supplementary Fig. 31). Therefore, it was presumed that, due to the low recombination rate in these heterozygous regions and large-effect dSNPs in repulsion phases, the genome could not purge these variants, leading to the retention of heterozygosity in these blocks.

As an example, the ~200 kb SDR was located on the heterozygous block on chromosome 2. There are four SDR haplotypes: Male-like haplotype (M), female-like haplotype (F), and hermaphrodite-haplotype (H1 and H2) with dominance effects $M > H > F$. H1 and H2 originated from the recombination of F and M to become hermaphroditic^{47,48}. We found that PN, GN, HE, and PN40024 are H2/F genotypes, SG and GB are H1/F genotype, and CD is H1/H2 genotype, which indicated that HE inherited the H2 haplotype from PN and the F haplotype from SG. Interestingly, after nine generations, PN40024 maintained the heterozygous genotype (H2/F) at SDR locus, enabling it to remain hermaphroditic (Fig. 5F). The F/F genotype could not be selfing by PN40024, because the flowers would be female. However, why did the H2 haplotype not become homozygous? One possibility is the presence of strongly recessive deleterious (even recessive lethal) SNPs on the H2 haplotype. To investigate this, we screened the sites that are heterozygous (0/1 genotype) in PN40024 and cannot simultaneously exist in a homozygous state for both the reference allele (0/0 genotype) and the alternative allele (1/1 genotype) in other grape accessions. Finally, 67 SNPs were identified. To focus on dSNPs located on the H2 haplotype inherited from PN, we excluded SNPs inherited from SG, resulting in 55 SNPs found in both PN and PN40024 accession (Fig. 5G, H). All these SNPs are located on two genes (*VviV12* and *VviGSM2*) and their flanking regions. Six of the SNPs had a GERP score reaching the maximum value of four, indicating a strong detrimental effect on grapes (Fig. 5I). Five of the six SNPs were in the CDS region of *VviV12*, along with 24 indels and an SV (87 bp) (Fig. 5J). Knocking out of the *AtV12* gene in *Arabidopsis* resulted in a severe reduction in pollen germination (Supplementary Fig. 32)^{49–52}. Therefore, the *VviV12* gene, associated with recessive lethality, located on the H2 haplotype likely maintained the SDR region in a heterozygous state on chromosome 2.

Discussion

In this study, we integrated comparative genomic and population genetic analyses to investigate the impacts of different reproductive systems on grapevine genomes and breeding. We collected grapevine lineages that represent different reproductive systems, from the wild ancestors of domesticated grapes to a diverse collection of accessions related to PN, including clonal variants, hybrid materials, and a selfed lineage. The results revealed the effects of different reproductive systems on population genetic characteristics, including heterogeneity, SFS, and the structural and deleterious burdens. The clonal propagation dramatically masked the harmful effects of recessive deleterious variants in heterozygous states, while selfing purged most of the deleterious and structural burdens. These discoveries underscore the genetic basis of reproductive systems, enhance our understanding of complex evolutionary genomic processes and provide a theoretical guideline for future grapevine genomic breeding by combining beneficial variations of agronomic and resistant traits while purging deleterious variations.

PN has long served as a pivotal genetic resource for breeding. We assembled a haplotype-resolved chromosomal genome of PN to capture the haplotypic diversity within this genome, including a ~4 Mb inversion between haplotypes PN1 and PN2 (Fig. 1A and Supplementary Fig. 2). We identified 1794 and 1819 gene families that were exclusive to PN1 and PN2 (Fig. 1B, C), respectively, suggesting that >10% of genes were in a hemizygous state in PN, which is also observed in other grapevine cultivars^{36,45,53–55}.

We conducted further investigations into somatic and fixed variants within the PN population to understand the origin of this cultivar. After domesticated grapevines were introduced to Europe, they exchanged genes with European wild grape²⁸. The signals of similar introgression events are evident in the PN genome and likely shaped agronomic traits. For instance, flowering-related loci *CONSTANS-like 5* (*COL5*) in the PN genome show evidence of both positive selection and introgression from the EU wild grape population to PN or its ancestral lineages (Fig. 2D). One gene belonging to the same family, *CONSTANS-LIKE 9*, has been reported previously to be under both introgression and positive selection in wine grapes in the Iberian Peninsula region^{10,56}. In addition, the SVs could play an important role in PN, which was also reported in other studies^{35,45}. For example, an SV on chromosome 10 harbored a gene cluster of S-locus family (Fig. 3E), which is probably related to the switch to self-fertility in domesticated grapes.

The reproductive systems have varying impacts on rates of linkage and recombination, the levels of genetic drift effects, the efficacy of selection in plants, which could have pronounced effects on genomic variations^{42,43}. In order to uncover the impact of clonal propagation in grapevine, we conducted an investigation into both somatic mutations and fixed mutations in PN using comparative population genetics. Almost all fixed mutations in PN were heterozygous (98.6% for SNPs and 91.9% for SVs, Fig. 3C). Clonal propagation is widely applied in grapevine production, which have a great advantage in ensuring the stability of vigorous phenotypes and genotypes^{5,57}. In this context, our results suggest that rapid selection with the aid of clonal propagation by breeders could occur once favorable mutations are identified during grapevine breeding²⁸. The highly heterozygous fixed mutations highlight the advantageous role of clonal reproduction in fixation of vigorous phenotypes and genotypes during the cultivation of grapes like PN.

On other hand, clonal propagation inhibits the recombination of variants and leads to the accumulation of somatic mutations⁵¹. However, its effect on fitness is variable and depends on somatic mutation rates and patterns, which can differ among various plants or even among different organs of a single individual^{58,59}. In this study, we identified 20,328 SNPs and 176 SVs that were specific to the PN group and that varied among PN clones. We deemed this group of mutations to be somatic mutations, since they represent new mutations since the formation of the PN varietal. We recognize, however, that some of these mutations may be chimeric—i.e., fixed throughout the tissues of the plant—but our sampling strategy is unable to detect such chimerisms. Additionally, the misidentification of somatic variations due to errors in SNP calling could occur. In this study, we decreased the occurrence of false positive results by using the most up-to-date sequencing and bioinformatic methods, including the complete T2T reference genome PN40024_T2T for read mapping, high-coverage whole genome sequencing data, and stringent criteria to remove low-quality variants. Therefore, these putative errors are limited and are unlikely to have significant impacts on the results of the population genetics analyses in this study. We found that most somatic mutations were exclusive to one individual and tended to be located in intergenic regions (Fig. 3C, Supplementary Fig. 15), suggesting that somatic variations tend to be rare which is consistent with that observed in cultivar Zinfandel and maize^{3,31}. In theory, deleterious somatic mutations can decrease the fitness of clonal plants^{10–12}. However, previous studies have used forward simulation to show that clonal propagation does not decrease fitness under recessive selection model^{20,60}. In addition to deleterious mutations, somatic mutation can give rise to beneficial variants that contribute to new traits^{36,38,39}, and even lead to the development of new cultivars, for example, the three bud sports varieties (Pinot Grigio, Pinot Meunier, and Pinot Blanc) from PN²².

Selfing, predominantly adopted by 10–15% of flowering plants, is another common strategy for plant reproduction⁶¹ and is applied for

breeding due to its many advantages, including the fixation of desirable variants and protection from potential performance downsides of hybridization⁶². However, there is a long-running debate as to whether selfing is “dead end”^{61,63,64}. Inbreeding depression is widely observed in the plant kingdom^{64,65}, including grapevines^{20,45,66}. One reason for this phenomenon could be the uncovering of recessive deleterious mutations, which has been observed in other selfing crops, such as potato and maize^{3,67,68}. In this study, we observed a significant decrease in genetic diversity in PN40024 group compared to other clonal and crossing reproduction cultivars (PN, SG, HE, GB, CD, GN, and the wild grapes). Successive selfing dramatically purged heterozygous dSNPs (62% compared to HE) and SVs (65% compared to HE) in PN40024, but 32% and 31% of the heterozygous burdens (dSNPs and SVs) were shifted to a homozygous state in PN40024 (Fig. 4C, D, Supplementary Figs. 19 and 20).

After nine generations of selfing, there were 116,199 SNPs and 1,125 SVs that persisted in heterozygosity in all four PN40024 clones (Fig. 5B, Supplementary Figs. 22, 23, 25, and 28)²⁵. A similar phenomenon – higher than expected heterozygosity – has also been reported for genomic regions of other selfed lineages, such as maize, potato, *Brachypodium*, and *Arabidopsis*^{3,67,69–71}. The hSNPs were found predominantly in six heterozygous blocks in PN40024 (Supplementary Table 4). Some of these heterozygous sites could reflect read misalignment^{69,71}. We examined read coverage in the heterozygous blocks and found that one such block on chromosome 16 could be false (Supplementary Data 11). To explore the genomic origins of these retained variants in selfing offspring, we compared them to the HE, PN, SG populations. We found that the variants in heterozygous states of HE have a higher proportion anchoring to genes in the regions corresponding to the heterozygous blocks in PN40024 than those outside of these regions (Supplementary fig. 30). The genes that are heterozygous in PN40024 were involved in chitin binding (Supplementary Data 10). In addition, rare recombination events were observed in these heterozygous block regions in PN40024 (Fig. 5A, Supplementary Figs. 21 and 22), indicating that low recombination rates prevent the genome from becoming homozygous during selfing. Similar results were also found in maize and potato^{3,68}. For example, Roessler et al. observed a significantly higher proportion of nonsynonymous SNPs in heterozygous blocks compared to homozygous blocks, and noted low recombination rates at these regions. Similarly, Zhang et al. showed that a highly heterozygous region of selfed potatoes contained two large-effect deleterious mutations (*led1* and *yl1*) in repulsion. Interestingly, the sex-determination region remains highly heterozygous in hermaphroditic domesticated grapes including PN40024 (Fig. 5F)^{45,47,72,73}. One reason, of course, is that plants with the F/F genotype are no longer hermaphroditic^{47,72}, and there are (to our knowledge) very few cases of homozygous H2/H2 genotypes. One potential reason is the presence of homozygous lethal alleles. We provide some evidence and speculate that one candidate gene, *VviV2*, on H2 haplotype of SDR region contains a recessive lethal mutation that prevents the propagation of the homozygous H2/H2 genotype (Figs. 5F, I and J). Thus, the deleterious variants located in low-recombining regions, where Hill-Robertson interference decreases the efficacy of purging deleterious variants in repulsion phases, are a primary driver of the retained heterozygous blocks in PN40024 and other selfed plant lineages^{65,67}.

One goal of breeders is to combine known beneficial alleles while purging deleterious alleles^{68,74–77}. The contribution of phased genomes, such as PN40024_T2T²⁵, the PN genome introduced here, along with several additional grapevine varieties^{53–55,78–80} and wild *Vitis* genomes^{81–84}, can help identify comprehensive variants in grapevine and develop efficient molecular markers for grapevine improvement. This study also presented a new understanding of applying different reproductive systems for breeding. Outcrossing is a good strategy for combining and introducing of new traits in domesticated grapes, and

subsequent clonal reproduction could retain these preferred alleles for production. However, the clonal reproduction leads to the accumulation of deleterious variants, which is not suitable for future breeding. Inbreeding depression is one of the main obstacles to fixing these preferred alleles through selfing while exposing heterozygous deleterious variants accumulated during clonal propagation^{20,75}. In this study, we revealed that deleterious mutations in repulsion phase are the cause of several heterozygous blocks in the PN40024 genome. Although this phenomenon has only been observed in one grapevine progeny PN40024, similar occurrences have been reported in crops such as potato and maize^{3,67}. Breaking the linkage of these deleterious mutations in repulsion phase would be a practical method, as has been done in potatoes⁶⁸. In addition, the influence of structural variations on grapevine phenotypes highlights the necessity of integrating SVs into breeding programs^{55,85}.

Our study illuminates the diverse impacts of different breeding methodologies, including cloning, crossing, and selfing, on the genetic heterozygosity, beneficial variants, and dSNP and SV burdens in the grapevine genome. Armed with this knowledge, breeders can make informed decisions regarding the selection of breeding methods and combinations to strategically pursue their specific breeding objectives.

Methods

Plant materials and sample collection

Fresh and healthy leaf tissue from plants of *Vitis vinifera* cultivar “PN” clone AGIS_01 was collected from the grapevine germplasm collection at the Agricultural Genomics Institute at Shenzhen, Shenzhen, Guangdong Province, China, and immediately frozen in liquid nitrogen. These materials were packaged for PacBio HiFi and Hi-C sequencing, respectively, and subsequently submitted to the company for genomic DNA extraction and library preparation.

Library preparation and DNA sequencing

Isolation of high-molecular-weight genomic DNAs using the DNeasy Plant Mini kit according to the manufacturer’s instructions. For PacBio HiFi sequencing, single-molecule real-time cells were sequenced on the PacBio Sequel II platform using Circular Consensus Sequencing (<https://github.com/PacificBiosciences/ccs>) with default parameters. For the Hi-C library, samples were digested with the restriction enzyme DpnII and constructed following a standard Hi-C protocol as described previously. These Hi-C libraries were sequenced on the Illumina HiSeq X Ten platform.

De novo haplotype-resolved genome assembly and quality assessment

Using the Hi-C Integrated Assembly mode of HiFiasm, we first generated two contig-level haplotype genomes using the HiFi and Hi-C data of PN⁸⁶. Genomic heterozygosity was assessed by GenomeScope (v2.0)⁸⁷, employing a *k*-mer-based methodology applied to raw HiFi reads. Then, we used RagTag to determine the approximate order of contigs on chromosomes by using PN40024_T2T as a reference genome²⁵. Subsequently, the Hi-C sequencing data were harnessed to anchor all contigs via Juicer (v1.5)⁸⁸. This was succeeded by employing a 3D-DNA scaffolding pipeline to further refine the structure⁸⁹. Manual adjustments were conducted on the acquired outcomes using Juicebox (v1.11.08, <https://github.com/aidenlab/Juicebox>), followed by a secondary application of the 3D-DNA approach to achieve the genome at the scaffold level. Employing Minimap2 (v2.24)⁹⁰, a comprehensive comparison was conducted between scaffold-level genomes from different *Vitis* species and the raw HiFi data. The outcomes were then imported into IGV (v2.12.3)⁹¹ to pinpoint the precise sequence positions of gaps. To validate the accuracy of gap filling, the identified sequences were reintegrated into the genome utilizing Minimap2⁹⁰.

For genome quality, we used QAST to count the basic information of the contig level genome and the final genome (<https://github.com>).

[com/ablab/quast](#)). The genome completeness was evaluated by BUSCO using the embryophyta_odb10 database⁹². The genome continuity was evaluated by calculating the contig N50 length. For genome accuracy, we mapped the genome with HiFi reads using Minimap2 and calculated mapping rates⁹⁰.

Annotation of genes and transposable elements

We primarily used this genome-wide annotation pipeline for genome annotation (<https://github.com/unavailable-2374/Genome-Wide-Annotation-Pipeline>). RNA-seq datasets were collected covering a variety of tissues, including flowers, leaves, and other tissues. RNA sequences were aligned to repeat mask assemblies using Hisat2 (v2.10.2) and subsequently assembled into transcripts using StringTie (v1.3.0)^{93,94}. Genes were first searched by using transcripts and UniProt (<https://www.uniprot.org/>) as evidence. In this process, we used Exonerate, Genewise, and Transfrag. In short, an initial gene model was created for the genes and further searches were performed using AUGUSTUS (v3.4.0)⁹⁵. Genes involving duplicated regions, CDS regions shorter than 90 nucleotides, or without any evidence to support them were filtered out. Finally, all results were checked with a hidden Markov model downloaded from the Pfam database to obtain the final gene model.

RepeatModeler (open-2.0.3) was used to build the TE library, with the -LTRStruct⁹⁶. Genome-wide TE annotation was performed using RepeatMasker (open-4.1.2, <https://github.com/rmhubble/RepeatMasker>), with -e rmbblast -lcambig and slow model.

Identification of telomeres and centromeres

For telomere identification, plant telomeric sequences (CCCATTT at the 5' end and TTTAGGG at the 3' end) were identified, and 70 out of the expected 76 telomeres (spanning 38 chromosomes of 2 haplotypes, 35/38 for each genome) were identified using the telomere pipeline developed by the TIDK (v0.2.0)⁹⁷.

TRF v4.09 was used to finish tandem repeat annotation, and then we merged the results of annotation using TRF2GFF⁹⁸. To complete data visualization, we analyzed the results in IGV (v2.12.3)⁹¹. The results were compared with TE annotation and TRF using IGV to identify the centromeres.

Genome comparison between PN1, PN2 and PN40024_T2T

We aligned the PN1, PN2, PN40024_T2T genomes using Minimap2, and indexed the alignment BAM file using SAMtools (v1.4)⁹⁹. Next, SyRI¹⁰⁰ was used to find structural variants between the genomes, and the results were visualized using plots¹⁰¹.

MUMmer (v4.0) was used to compare the genome with the reference genome PN40024_T2T using whole-genome alignments¹⁰². First, we aligned the two genome sequences using Nucmer and then filtered one-to-one alignments with a minimum alignment length of 10,000 bp (delta-filter -i 95 -l 10000).

To identify gene families in PN40024_T2T, PN1, and PN2, Orthofinder v2.5.2 was utilized¹⁰³ according to a previous study¹⁰⁴, comparing their protein-coding gene sequences. Basically, the genes that similar to each other and distinct from genes in other groups were clustered together based on the sequence similarity. If no gene was identified in a gene family for a specific genome, this gene family was counted as absent in that genome. The gene abundance in one gene family on one genome was calculated by dividing the number of the genes in this gene family on that genome by the maximum number of the genes in this gene family among the three genomes.

Comparing SVs on the Pinot Noir genomes

To identify SVs on the PN genome (between the two haplotypes), we called SVs using the Sniffles pipeline¹⁰⁵. First, PacBio reads longer than 500 bp were mapped onto PN1 PN2 genomes using the aligner Minimap2⁹⁰. Variant calling was then performed with Sniffles. SV

analysis outputs (VCF files) were filtered by VCFtools (v0.1.16)¹⁰⁶ to find the heterozygous SVs between the two haplotypes in PN. To further validate the existence of this inversion, we designed primers at the break points on both haplotypes, following with PCR and gel electrophoresis analyses (Supplementary Table 2 and Supplementary Fig. 4).

SNP calling and filtering

We used 68 grapevine resequencing samples, including ten wild grapes (*V. vinifera* subsp. *sylvestris*) from Europe (EU), ten wild grapes from the Middle East (ME), 18 PN, five SG, four GB, ten CD, two GN, two HE, and four PN40024, along with three muscadine grapes used as outgroup. Among these samples, nine of the clonal PN collected from Foundation Plant Services, University of California, Davis were sequenced in this study, while the remaining data was downloaded from the database on the website (Supplementary Table 4).

Using default parameters, fastp (v0.21) was used to regulate the quality of resulting raw reads¹⁰⁷. The PN40024_T2T assembly (PRJNA882193 in NCBI)²⁵ served as the reference genome. Quality-controlled reads were mapped to the genome using bwa (v 0.7.15) with default parameters¹⁰⁸. SAMtools (v1.4) and GATK (v4.1.8) were used for sorting and indexing the bam file with no duplicates^{109,110}. GTX, which is based on the Haplotype Caller of GATK, was used for SNP calling across all samples.

To reduce false positives, filtering was conducted using VCFtools (v0.1.16)¹⁰⁶. We removed genotypes with a genotype quality <20 (option -minGQ 20). SNPs with more than two alleles were excluded (options -min-alleles 2 --max-alleles 2), as well as SNPs with more than 20% missing genotypes (option -max-missing 0.8).

The filtered SNPs were phased by Beagle (v5.4) genotype imputation method¹¹¹. The results were then reversed using Model 1 in the superSFS (<https://github.com/xhchauvet/superSFS>) script with parameter 3 to infer ancestral alleles according to outgroup species.

Phylogeny and population structure

For the analysis related to introgression in PN, the phylogenetic tree was constructed by IQ-TREE (v2.1.4)¹¹² based on general time reversible (GTR) model using variants from three outgroup samples, 20 wild grapes and 42 wine cultivars in the 20 kb region where *COL5* is located. For the analysis related to the impact of clonal propagation on grapevine genomes, the whole genome SNPs of the 68 previously mentioned grapevines were first thinned using VCFtools (--thin 1000). Phylogenetic tree was then constructed using IQ-TREE based on the GTR model. To construct the phylogenetic tree of the *COL5* and *V12* genes among different species, the DNA sequences of these genes were analyzed using MEGA X¹¹³ with the method neighbor-joining and p-distance.

The population structure based on nuclear SNPs was estimated using Admixture (v1.3.0) with *K* varying from 2 to 10¹¹⁴. Kinship and IBSO analyses were performed using King⁴¹. The relationship between samples were assessed according to the following criterion: identical clones ($K \geq 0.49$ and $IBSO \leq 0.001$), parent-offspring ($0.177 < K < 0.354$ and $IBSO \leq 0.001$), highly related/sibling ($0.177 < 0.354$ and $IBSO \leq 0.25$)²¹.

Population genetic analyses

The Nucleotide diversity (π) value for each sample was calculated using VCFtools with a 100 kb window size. Sequence similarity (D_{xy}) and *f*-statistic (f_d) were calculated using the Python script: popgenWindows.py (https://github.com/simonhmartin/genomics_general) with every 50 kb nonoverlapping windows. For f_d analysis, ME were used as the sister population to PN, and the gene flow between EU and PN grapes was evaluated. The heterozygous sites were counted using VCFtools. The number of heterozygous sites in each sample was divided by the total number of SNPs to calculate the heterozygosity of each sample. These statistics were evaluated within or between all six groups (EU, $n = 10$; ME,

$n=10$; PN, $n=18$; SG, $n=3$; HE, $n=2$; and PN40024, $n=4$). The heterozygosity was calculated by dividing the number of heterozygous sites in each sample by the total number of SNPs.

Selective sweep detection

PBScan (<https://github.com/thamala/PBScan>) was used for PBS analysis²⁶. The PN group was designated as POP1, the ME group as POP2, and the EU group as POP3. The analysis was conducted every 50 SNPs with a step size of 50 SNPs based on sequence similarity (D_{xy}).

Detection of deleterious mutations

We used Sorting Intolerant From Tolerant 4G (SIFT 4G) (https://github.com/pauline-ng/SIFT4G_Create_Genomic_DB) to annotate the SNP dataset in order to estimate the functional effects of mutations⁴⁴. Swiss-Prot (<https://www.uniprot.org/help/downloads>) was utilized as a reference protein set to construct a grapevine database. Meanwhile, we used the genome and annotation files of PN40024_T2T. The GFF3 format was converted to the Ensemble GTF format. The generated database was then used to annotate the flipped SNPs. SIFT values range from 0 to 1, and any nonsynonymous position with a SIFT score of less than 0.05 was considered putatively deleterious.

ANGSD analysis

We used ANGSD software to calculate the locus spectrum (<http://www.popgen.dk/angsd/index.php/ANGSD>), which shows the frequencies of individual alleles at specific loci in the PN population¹¹⁵. First, we ran the script using the reversed VCF file, combining the reference genome to extract the sequence file of the ancestral genotype state (anc.fa). We then used the -doSaf 1 parameter to generate.saf (site allele frequency likelihood), and finally, the realSFS command was used to generate the site spectrum result file.sfs text format.

SV calling and filtering

Delly (v1.0.3)¹¹⁶ was used to call SVs with PN40024_T2T as the reference genome. The “call” and “merge” functions were first used to get the.bcf files, and then the Delly “call” function was used again with “-v” parameter. Subsequently, the BCFtools (v1.13) “merge” function was used to merge the.bcf files. We removed low-quality SVs from the merged SVs that did not pass quality filters. After that, we used VCFtools to remove SV genotypes with more than 20% missing genotypes. The number of SVs containing TEs was counted based on the TE annotation of PN40024_T2T using Bedtools¹¹⁷.

Gene set enrichment analysis and gene ontology enrichment analysis

We downloaded the protein database (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/swissprot.gz>) from NCBI. Protein sequence alignment was then performed using Diamond¹¹⁸. The database was first established using the Diamond “makedb” command, and then sequence alignment was performed using the Diamond “blastp” command. Next, the GSEA and GO enrichment analyses were performed using clusterProfiler (v4.0)²⁷ with the GO terms of the whole grapevine genome proteins as the background. First, we used “AnnotationForge” package in R to build “OrgDb” for gene to GO mapping. Then, for the GSEA analysis, the genes were ordered according to their PBS values, and the “gseGO” function of clusterProfiler was used for these genes with parameters: pAdjustMethod = “BH”, qvalueCutoff = 0.05. For the GO analysis, “enrichGO” function of clusterProfiler were used for candidate genes with the parameters: pAdjustMethod = “BH”, qvalueCutoff = 0.05.

Analysis of heterozygous variants

The heterozygous blocks were detected by following the methods: if the distance between two hSNPs was less than 150,000 bp, the two SNPs were counted as continuous variation sites; if the span of the

continuous variation sites was more than 500,000 bp, this region was counted as heterozygous blocks.

For analysis of different genotype combinations among cultivars, the conserved sites, which were consistent in each cultivar group, were used for counting. The combination categories with site numbers less than 30 were not included.

For the distribution analysis of SNPs on genes, 2 kb upstream or downstream of the genes were counted as promoters or terminators. The density of heterozygous SVs and dSNPs was calculated in every 200 kb non-overlapping window.

Species specific marker

To select the specific marker belonging to PN or SG, we used the following criteria: the variants must exist in only one grapevine cultivar with the frequency greater than 50%, and must not be observed in all individuals of another grapevine cultivar.

The GERP score

The longest transcripts were extracted from the genomes of *Arabidopsis thaliana*, *Populus trichocarpa*, *Oryza sativa*, *Vitis rotundifolia*, *Vitis amurensis*, *Cissus rotundifolia*, *Malus domestica*, *Ficus macrocarpa* and PN40024, then Orthofinder (v2.5.4)¹⁰³ was used to generate single-copy protein sequences, which were used to construct the phylogenetic tree using IQ-TREE.

The genome of PN40024 was split into individual chromosomes, a multigene linear comparison was performed using last (<https://gitlab.com/mcfrith/last>), multiple .maf files were merged and converted to .fa files. Each SNP on each of PN40024's chromosomes was assessed using GERP¹¹⁹.

Reporting summary

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

Data availability

All data sequenced in this study have been deposited in the NCBI Sequence Read Archive under Bioproject accession PRJNA951461 and the National Genomics Data Center (NGDC) Genome Sequence Archive (GSA) under Bioproject accession PRJCA016741. The assembly and annotation have been deposited in Zenodo [<https://zenodo.org/record/8080252>]. Source data are provided with this paper.

References

- Labroo, M. R., Studer, A. J. & Rutkoski, J. E. Heterosis and hybrid crop breeding: a multidisciplinary review. *Front. Genet.* **12**, 643761 (2021).
- Nordborg, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929 (2000).
- Roessler, K. et al. The genome-wide dynamics of purging during selfing in maize. *Nat. Plants* **5**, 980–990 (2019).
- Dieterich Mabin, M. E., Brunet, J., Riday, H. & Lehmann, L. Self-fertilization, inbreeding, and yield in alfalfa seed production. *Front. Plant Sci.* **12**, 700708 (2021).
- Miller, A. J. & Gross, B. L. From forest to field: perennial fruit crop domestication. *Am. J. Bot.* **98**, 1389–1414 (2011).
- Yu, L. et al. Somatic genetic drift and multilevel selection in a clonal seagrass. *Nat. Ecol. Evol.* **4**, 952–962 (2020).
- Padovan, A., Lanfear, R., Keszei, A., Foley, W. J. & Külheim, C. Differences in gene expression within a striking phenotypic mosaic Eucalyptus tree that varies in susceptibility to herbivory. *BMC Plant Biol.* **13**, 29 (2013).
- López, E. H. & Palumbi, S. R. Somatic mutations and genome stability maintenance in clonal coral colonies. *Mol. Biol. Evol.* **37**, 828–838 (2020).

9. Foster, T. M. & Aranzana, M. J. Attention sports fans! The far-reaching contributions of bud sport mutants to horticulture and plant biology. *Hortic. Res.* **5**, 44 (2018).
10. Ally, D., Ritland, K. & Otto, S. P. Aging in a long-lived clonal tree. *PLoS Biol.* **8**. <https://doi.org/10.1371/journal.pbio.1000454> (2010).
11. Barrett, S. C. H. Influences of clonality on plant sexual reproduction. *Proc. Natl. Acad. Sci. USA* **112**, 8859–8866 (2015).
12. Bobiwash, K., Schultz, S. T. & Schoen, D. J. Somatic deleterious mutation rate in a woody plant: estimation from phenotypic data. *Heredity* **111**, 338–344 (2013).
13. Charlesworth, D. & Charlesworth, B. Quantitative genetics in plants: the effect of the breeding system on genetic variability. *Evolution* **49**, 911–920 (1995).
14. 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).
15. Cheng, L. et al. Leveraging a phased pangenome for haplotype design of hybrid potato. *Nature*; <https://doi.org/10.1038/s41586-024-08476-9> (2025).
16. Emanuelli, F. et al. Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biol.* **13**, 39 (2013).
17. Janick, J. & Paull, R. E. *The Encyclopedia of Fruit & Nuts* (CABI North American Office, 2008).
18. This, P., Lacombe, T. & Thomas, M. R. Historical origins and genetic diversity of wine grapes. *Trends Genet.* **22**, 511–519 (2006).
19. Coito, J. L. et al. Vitis flower types: from the wild to crop plants. *PeerJ* **7**, e7879 (2019).
20. 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).
21. Ramos-Madriral, J. et al. Palaeogenomic insights into the origins of French grapevine diversity. *Nat. Plants* **5**, 595–603 (2019).
22. Vezzulli, S. et al. Pinot blanc and Pinot gris arose as independent somatic mutations of Pinot noir. *J. Exp. Bot.* **63**, 6359–6369 (2012).
23. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
24. Velt, A. et al. An improved reference of the grapevine genome reasserts the origin of the PN40024 highly homozygous genotype. *G3* **13**. <https://doi.org/10.1093/g3journal/jkad067> (2023).
25. Shi, X. et al. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Hortic. Res.* **10**, uhad061 (2023).
26. Hämälä, T. & Savolainen, O. Genomic patterns of local adaptation under gene flow in *Arabidopsis lyrata*. *Mol. Biol. Evol.* **36**, 2557–2571 (2019).
27. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
28. Xiao, H. et al. Adaptive and maladaptive introgression in grapevine domestication. *Proc. Natl. Acad. Sci. USA* **120**, e2222041120 (2023).
29. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).
30. Hassidim, M., Harir, Y., Yakir, E., Kron, I. & Green, R. M. Overexpression of CONSTANS-LIKE 5 can induce flowering in short-day grown *Arabidopsis*. *Planta* **230**, 481–491 (2009).
31. Vondras, A. M. et al. The genomic diversification of grapevine clones. *BMC Genom.* **20**, 972 (2019).
32. Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. Retrotransposon-induced mutations in grape skin color. *Science* **304**, 982 (2004).
33. Walker, A. R., Lee, E. & Robinson, S. P. Two new grape cultivars, bud sports of Cabernet Sauvignon bearing pale-coloured berries, are the result of deletion of two regulatory genes of the berry colour locus. *Plant Mol. Biol.* **62**, 623–635 (2006).
34. Leng, F. et al. Comparative transcriptomic analysis between ‘Summer Black’ and its bud sport ‘Nantaihutezao’ during developmental stages. *Planta* **253**, 23 (2021).
35. Aversano, R. et al. Distinct structural variants and repeat landscape shape the genomes of the ancient grapes Aglianico and Falanghina. *BMC Plant Biol.* **24**, 88 (2024).
36. Carbonell-Bejerano, P. et al. Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiol.* **175**, 786–801 (2017).
37. Urrea, C. et al. Identification of grapevine clones via high-throughput amplicon sequencing: a proof-of-concept study. *G3* **13**. <https://doi.org/10.1093/g3journal/jkad145> (2023).
38. Roach, M. J. et al. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLoS Genet.* **14**, e1007807 (2018).
39. Ban, S. & Jung, J. H. Somatic mutations in fruit trees: causes, detection methods, and molecular mechanisms. *Plants* **12**, <https://doi.org/10.3390/plants12061316> (2023).
40. Liang, M. et al. Evolution of self-compatibility by a mutant Sm-RNase in citrus. *Nat. Plants* **6**, 131–142 (2020).
41. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
42. Hough, J., Williamson, R. J. & Wright, S. I. Patterns of selection in plant genomes. *Annu. Rev. Ecol. Evol. Syst.* **44**, 31–49 (2013).
43. Slotte, T. et al. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**, 831–835 (2013).
44. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
45. Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
46. Hedrick, P. W., Hellsten, U. & Grattapaglia, D. Examining the cause of high inbreeding depression: analysis of whole-genome sequence data in 28 selfed progeny of *Eucalyptus grandis*. *N. Phytol.* **209**, 600–611 (2016).
47. Zou, C. et al. Multiple independent recombinations led to hermaphroditism in grapevine. *Proc. Natl. Acad. Sci. USA* **118**, <https://doi.org/10.1073/pnas.2023548118> (2021).
48. Zhou, Y., Muyle, A. & Gaut, B. S. Evolutionary genomics and the domestication of grapes. in *The Grape Genome* (eds Cantu, D. & Walker, M. A.) 39–55 (Springer International Publishing, Cham, 2019).
49. Leskow, C. C. et al. Allelic differences in a vacuolar invertase affect *Arabidopsis* growth at early plant development. *J. Exp. Bot.* **67**, 4091–4103 (2016).
50. Vu, D. P. et al. Vacuolar sucrose homeostasis is critical for plant development, seed properties, and night-time survival in *Arabidopsis*. *J. Exp. Bot.* **71**, 4930–4943 (2020).
51. Wang, J. et al. The peptidyl-prolyl isomerases FKBP15-1 and FKBP15-2 negatively affect lateral root development by repressing the vacuolar invertase VIN2 in *Arabidopsis*. *Planta* **252**, 52 (2020).
52. Seitz, J. et al. How pollen tubes fight for food: the impact of sucrose carriers and invertases of *Arabidopsis thaliana* on pollen development and pollen tube growth. *Front. Plant Sci.* **14**, 1063765 (2023).
53. 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.* <https://doi.org/10.1093/hr/uhae080> (2024).
54. Zhang, K. et al. The haplotype-resolved T2T genome of teinturier cultivar Yan73 reveals the genetic basis of anthocyanin biosynthesis in grapes. *Hortic. Res.* **10**, uhad205 (2023).
55. Wang, X. et al. Integrative genomics reveals the polygenic basis of seedlessness in grapevine. *Curr. Biol.* **34**, 3763–3777.e5 (2024).
56. Freitas, S. et al. Pervasive hybridization with local wild relatives in Western European grapevine varieties. *Sci. Adv.* **7**, eabi8584 (2021).

57. Forneck, A., Benjak, A. & Rühl, E. Grapevine (*Vitis* ssp.): example of clonal reproduction in agricultural important plants. in *Lost Sex*, (eds Schön, I., Martens, K. & Dijk, P.) 581–598 (Springer Netherlands, 2009).
58. Schoen, D. J. & Schultz, S. T. Somatic mutation and evolution in plants. *Annu. Rev. Ecol. Syst.* **50**, 49–73 (2019).
59. Zheng, Z. et al. Somatic mutations during rapid clonal domestication of *Populus alba* var. *pyramidalis*. *Evol. Appl.* **15**, 1875–1887 (2022).
60. 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).
61. Wright, S. I., Kalisz, S. & Slotte, T. Evolutionary consequences of self-fertilization in plants. *Proc. Biol. Sci.* **280**, 20130133 (2013).
62. Levin, S. A. *Encyclopedia of Biodiversity* (Academic Press, 2001).
63. Stebbins, G. L. Self fertilization and population variability in the higher plants. *Am. Nat.* **91**, 337–354 (1957).
64. Takebayashi, N. & Morrell, P. L. Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. *Am. J. Bot.* **88**, 1143–1150 (2001).
65. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nat. Rev. Genet.* **10**, 783–796 (2009).
66. Holtgräwe, D. et al. A partially phase-separated genome sequence assembly of the *Vitis* Rootstock ‘Börner’ (*Vitis riparia* × *Vitis cinerea*) and its exploitation for marker development and targeted mapping. *Front. Plant Sci.* **11**, 156 (2020).
67. Zhang, C. et al. The genetic basis of inbreeding depression in potato. *Nat. Genet.* **51**, 374–378 (2019).
68. Zhang, C. et al. Genome design of hybrid potato. *Cell* **184**, 3873–3883.e12 (2021).
69. Bukowski, R. et al. Construction of the third-generation *Zea mays* haplotype map. *GigaScience* **7**, 1–12 (2018).
70. Stritt, C. et al. Migration without interbreeding: evolutionary history of a highly selfing Mediterranean grass inferred from whole genomes. *Mol. Ecol.* **31**, 70–85 (2022).
71. Jaegle, B. et al. Extensive sequence duplication in *Arabidopsis* revealed by pseudo-heterozygosity. *Genome Biol.* **24**, 44 (2023).
72. Massonnet, M. et al. The genetic basis of sex determination in grapes. *Nat. Commun.* **11**, 2902 (2020).
73. 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).
74. Wallace, J. G., Rodgers-Melnick, E. & Buckler, E. S. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu. Rev. Genet.* **52**, 421–444 (2018).
75. Wu, Y. et al. Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313–2328.e15 (2023).
76. Huang, X., Huang, S., Han, B. & Li, J. The integrated genomics of crop domestication and breeding. *Cell* **185**, 2828–2839 (2022).
77. Long, Q. et al. Population comparative genomics discovers gene gain and loss during grapevine domestication. *Plant Physiol.* **195**, 1401–1413 (2024).
78. Shirasawa, K. et al. De novo whole-genome assembly in an inter-specific hybrid table grape, ‘Shine Muscat’. *DNA Res.* **29**, <https://doi.org/10.1093/dnares/dsac040> (2022).
79. Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A. & Cantu, D. Diploid genome assembly of the wine grape carménère. *G3* **9**, 1331–1337 (2019).
80. Djari, A. et al. Haplotype-resolved genome assembly and implementation of VitExpress, an open interactive transcriptomic platform for grapevine. *Proc. Natl. Acad. Sci. USA* **121**, e2403750121 (2024).
81. Girollet, N. et al. De novo phased assembly of the *Vitis riparia* grape genome. *Sci. Data* **6**, 127 (2019).
82. Cochetel, N. et al. A super-pangenome of the North American wild grape species. *Genome Biol.* **24**, 290 (2023).
83. Zhang, T. 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).
84. Minio, A., Cochetel, N., Massonnet, M., Figueroa-Balderas, R. & Cantu, D. HiFi chromosome-scale diploid assemblies of the grape rootstocks 110R, Kober 5BB, and 101-14 Mgt. *Sci. Data* **9**, 660 (2022).
85. Liu, Z. et al. Grapevine pangenome facilitates trait genetics and genomic breeding. *Nat. Genet.* <https://doi.org/10.1038/s41588-024-01967-5> (2024).
86. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
87. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. Genome-Scope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
88. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (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. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
91. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
92. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
93. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
94. 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).
95. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
96. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
97. Brown, M., González De la, R., Pablo, M. & Mark, B. A *Telomere Identification Toolkit* (Zenodo, 2023).
98. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
99. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, <https://doi.org/10.1093/gigascience/giab008> (2021).
100. 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).
101. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).
102. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
103. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
104. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
105. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
106. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

107. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, <https://doi.org/10.1002/imt2.107> (2023).
108. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
109. Li, H. et al. The sequence alignment/map format and SAM tools. *Bioinformatics* **25**, 2078–2079 (2009).
110. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
111. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
112. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
113. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
114. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
115. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinforma.* **15**, 356 (2014).
116. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
117. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
118. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
119. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).

Acknowledgements

We thank all members in the Zhou lab for their useful discussion on the project. We thank Becky Gaut for the PN grape sequencing. This work was supported by the National Natural Science Foundation of China (No. 32372662) to Y.Z., the Science Fund Program for Distinguished Young Scholars of the National Natural Science Foundation of China (Overseas) to Y.Z., the National Key Research and Development Program of China (No. 2023YFD2200700), and National Science Foundation (USA) grant #1741627 to B.S.G.

Author contributions

Conceptualization: Y.Z. and H.X. Assembly generation: Siyang H., S.C., and X.W. Sample acquisition: Z.L., X.X., H.X., W.L., S.R., A.M.W., and B.S.G. Methodology and investigation: H.X., Y.W., Siyang H., W.L., X.S., Q.L., X.X., Y.P., P.W., Z.J., and Sanwen H. Writing—original draft: H.X., Y.W., W.L., X.S., Siyang H., and Y.Z. Writing—review & editing: All authors.

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/s41467-025-56817-7>.

Correspondence and requests for materials should be addressed to Yongfeng Zhou.

Peer review information *Nature Communications* thanks Crispin Jordan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025