nature plants

ARTICLES

https://doi.org/10.1038/s41477-019-0507-8

The population genetics of structural variants in grapevine domestication

Yongfeng Zhou[1], Andrea Minio[2], Mélanie Massonnet[2], Edwin Solares[1], Yuanda Lv[1], Tengiz Beridze[3], Dario Cantu[2*] and Brandon S. Gaut[1*]

Structural variants (SVs) are a largely unexplored feature of plant genomes. Little is known about the type and size of SVs, their distribution among individuals and, especially, their population dynamics. Understanding these dynamics is critical for understanding both the contributions of SVs to phenotypes and the likelihood of identifying them as causal genetic variants in genome-wide associations. Here, we identify SVs and study their evolutionary genomics in clonally propagated grapevine cultivars and their outcrossing wild progenitors. To catalogue SVs, we assembled the highly heterozygous Chardonnay genome, for which one in seven genes is hemizygous based on SVs. Using an integrative comparison between Chardonnay and Cabernet Sauvignon genomes by whole-genome, long-read and short-read alignment, we extended SV detection to population samples. We found that strong purifying selection acts against SVs but particularly against inversion and translocation events. SVs nonetheless accrue as recessive heterozygotes in clonally propagated lineages. They also define outlier regions of genomic divergence between wild and cultivated grapevines, suggesting roles in domestication. Outlier regions include the sex-determination region and the berry colour locus, where independent large, complex inversions have driven convergent phenotypic evolution.

Most plant genomes have been assembled from homozygous source materials. In some cases—such as selfing *Arabidopsis thaliana*[1] and rice[2,3]—homozygosity is the natural form. In other species, such as maize[4], apple[5] and roses[6], homozygosity is either based on manipulated haploid tissue or inbred lines. This focus on homozygous materials is technically convenient for genome assembly but it has at least three important biological limitations. First, inbreeding can substantively modify plant genome structure and content, due to the rapid purging of deleterious alleles over even a handful of generations[7]. As a consequence, most current plant reference genomes may provide a mere snapshot of diploid genome content.

Second, homozygous genomes provide no insights into the structural variants (SVs) that distinguish heterozygous chromosomes. The result has been a pervasive, discipline-wide dearth of information about the type and size of SVs, their distribution among cultivars, their population dynamics and their phenotypic effects. This gap in knowledge is critical because studies suggest that SVs explain as much or more phenotypic variation than do single nucleotide polymorphisms (SNPs); in humans, for example, SVs are threefold more likely to associate with phenotypes than single nucleotide variants[8]. SVs are also crucial for understanding the process of adaptation; as evidence of this, SVs are the causative genetic variant for at least one-third of known domestication alleles[9]. Fortunately, comprehensive SV catalogues are beginning to appear for humans[8,10] and some critical crops, such as rice[11] and maize[12]. However, the population frequencies of crop SVs have not yet been assessed thoroughly. Such a study is a prerequisite for understanding the evolutionary forces that act on SVs and for making a pragmatic assessment as to whether SVs can be effectively tagged by SNPs in association analyses.

A third limitation of the current focus on homozygous materials is that it has restricted insights into the biology of clonally propagated crops, which exist in a state of permanent heterozygosity and accumulate somatic mutations over time. Hundreds of economically important crops are propagated clonally, including most perennial crops[13]. Here we study the population dynamics of SVs in one of these perennials, the domesticated grapevine (*Vitis vinifera* ssp. *sativa*; hereafter '*sativa*'). The grapevine is one of the most economically important horticultural crops, with ~76×10^6 t of fruit harvested globally in 2015 (refs. [14,15]). Grapevine was domesticated from its wild ancestor, the Eurasian grapevine (*Vitis vinifera* ssp. *sylvestris*; hereafter '*sylvestris*'), nearly 8,000 years ago in the Transcaucasus[16]. Domestication increased sugar content in the berry, enlarged berry and bunch size, altered seed morphology and prompted a shift from dioecy—separate male and female individuals—to hermaphroditism[17]. In theory, hermaphroditic grapevine cultivars can be selfed; in practice, selfed progeny are often non-viable, perhaps because inbreeding exposes deleterious alleles hidden in the heterozygous state[18]. Consequently, most grape cultivars represent crosses between distantly related parents; this parentage, along with the accumulation of somatic mutations, result in grape cultivars that are often highly heterozygous[19–22].

Our goal is to fill a major gap in our understanding of plant genome evolution by investigating the population genetic of SVs in wild and domesticated grapevines. To do so, we first report a genome sequence of the Chardonnay cultivar based on a combination of long- and short-read sequencing technologies. Given this genome, we assess the extent and pattern of SVs between homologous chromosomes, including SVs that cause genic hemizygosity. We then compare Chardonnay to another cultivar, Cabernet Sauvignon, and use the SVs between these genomes to help guide the inference of SVs across a population sample of grapevine cultivars and their wild progenitor. With population data, we infer the strength of selection against different types of variants, explore the effects of a shift from outcrossing in *sylvestris* to clonal propagation in cultivated *sativa*, and, finally, investigate genomic regions with particularly marked SV divergence between grapevine and its wild progenitor.

[1]Department of Ecology and Evolutionary Biology, UC Irvine, Irvine, CA, USA. [2]Department of Viticulture and Enology, UC Davis, Davis, CA, USA. [3]Institute of Molecular Genetics, Agricultural University of Georgia, Tbilisi, Georgia. *e-mail: dacantu@ucdavis.edu; bgaut@uci.edu

Table 1 | Assembly statistics of the Chardonnay genome and two comparatives used in this study

Cultivar	Abbreviation	Assembly statistics			Annotation		
		Assembly size (Mb)	Contig N50 (Mb)	Scaffold N50 (Mb)	Genes	BUSCO (%)	TE (%)
Chardonnay	Char04[a]	606	1.24	24.5	38,020	93.4	47.3
Cabernet Sauvignon	Cab08 (ref. [21])	591	2.17	-	36,687	92.5	51.1
Pinot Noir derived	PN40024 (ref. [43])	486	0.102	3.4	41,163	96.9	47.0

[a] See this paper.

Results

Rampant hemizygosity in clonal propagated grapevine genomes. We initiated our study of SVs in grapevines by generating a reference genome for the Chardonnay cultivar, choosing a clone (FPS 04) that is grown worldwide. We used a hybrid sequencing approach, based on sequence data of 58× coverage of single-molecule real-time (SMRT, Pacific Biosciences) long-reads and 162× short-read coverage. Hybrid assembly resulted in a contig N50 of 1.24 megabases (Mb); application of high-throughput chromatin conformation capture (Hi-C) improved the scaffold assembly N50 to 24.5 Mb, extending contiguity relative to other grape genomes[19,20,23] (Table 1). The resulting primary assembly was 605 Mb in length, a value 20% higher than the partially inbred Pinot Noir (PN40024) reference[20] but similar to the 590 Mb assembly of Cabernet Sauvignon (Cab08)[21]. The Chardonnay (Char04) primary assembly included 93.4% of the complete universal single-copy orthologues (BUSCO) genes, contained 37,244 annotated protein-coding genes and consisted of 47.3% transposable elements (TEs), particularly from the *gypsy* and *copia* superfamilies (Table 1 and Supplementary Table 1).

We identified heterozygous SVs (hSVs) in Char04 by remapping SMRT reads to the Char04 reference using Sniffles[24], revealing 18,998 hSVs of length >50 base pairs (bp) (Fig. 1a and Supplementary Table 2). Only 0.3% of the hSVs were detected as homozygous (Supplementary Table 2), suggesting a low rate of mis-assembly. After masking potential mis-assemblies, observed hSVs were as long as 5.3 Mb and together constituted 91.21 Mb or 15.1%, of the 605-Mb primary assembly. The hSVs were assigned to five categories relative to the reference: deletions (DELs), duplications (DUPs), inversions (INVs), translocations (TRAs) and mobile elements insertions (MEIs). DEL and MEI events were the most numerous, with 8,302 and 7,772 (Supplementary Table 2), respectively. In addition to SVs > 50 bp in length, we also detected 119,067 small (<50 bp) indels and 873,159 SNPs. After including small indels, we estimated that the two Char04 chromosomal sets may differ by as much as 15.3% in length, with 9.4% of this caused by TEs that are polymorphic between chromosomes.

We assessed the extent to which SVs affect genes. Surprisingly, we found 5,546 hemizygous genes in Char04 based on SV inferences from long-read mapping (Fig. 1b), representing 14.6% of all annotated protein-coding genes. We required that the evidence for each hemizygous gene is supported by a split read, hence each hemizygosity inference has positive support. In addition, the hemizygosity value (14.6%) is consistent with the overall proportion of chromosomal heterozygosity by length. However, the high value also raises concerns that it could be artificial due to artifacts in mapping or in the Char04 reference. To allay these concerns, we performed two additional analyses to detect hSVs. First, we repeated the analysis by mapping Char04 long reads to the PN40024 reference. We detected slightly more (6,419) hemizygous genes but they again constituted ~15% of all annotated genes in the reference. Second, we mapped SMRT reads from Cab08 to the Cab08 assembly and detected 5,702 (15.5%) hemizygous genes in this cultivar. All of these analyses are consistent and indicate that hemizygosity affects about one in seven genes in grapevine cultivars.

An integrative comparison between Chardonnay and Cabernet Sauvignon. The Char04 and Cab08 genome assemblies permitted a rare opportunity to compare highly contiguous genomes from within a single cultivated species. We detected SVs between the two genomes using three approaches. We first mapped SMRT reads from Cab08 to the Char04 primary assembly (Supplementary Fig. 1). These results yielded about threefold higher numbers of SV events between cultivars than in Char04 (Supplementary Table 2), reflecting the distinct parentage of Chardonnay and Cabernet Sauvignon[18,25–28]. Of 59,913 inferred SVs, DEL and MEI events were again the most numerous, with 24,138 and 21,722 events, respectively, between genomes. SMRT-read alignment further confirmed high hemizygosity of protein-coding genes, because the two cultivars differed in ploidy level for 9,330 genes. Of these, 2,217 showed complete presence/absence variation (PAV), a number similar to previous estimates based on less complete data[29,30]. Based on gene ontology analyses, PAV genes are biased toward functions in defence response, flower development, membrane components and transcription factors ($P < 0.001$).

We also compared Char04 and Cab08 primary assemblies by whole-genome alignment[31] (Supplementary Fig. 2), which yielded a similar numbers of SVs (52,952) but fewer MEI events (Supplementary Table 2). Finally, we mapped 25× Illumina reads from Cab08 to Char04, which detected only 62% of the number of SVs based on SMRT-read mapping (Supplementary Table 2). The length distribution of SVs varied among the three methods; SMRT-read analyses detected larger (>10 kb) events (Supplementary Fig. 3). Importantly, 75% of SVs inferred by SMRT-read alignment were confirmed either by whole-genome alignment or by short-read alignment analyses (Supplementary Fig. 4). These confirmed SVs encompassed 1,822 PAV genes and 45,403 MEIs between Char04 and Cab08 and continue to attest to remarkable SV variation among grapevine cultivars.

Strong purifying selection against SVs. To gain wider information about SVs in grapevines and their wild relatives, we amassed short-read sequencing data representing 50 grapevine cultivars and 19 wild relatives, all of which exceeded a coverage depth of 10× (Supplementary Table 3). The application of short-read alignment for detecting SVs is subject to high levels of false-negatives and false-positives[24]. To limit false-positives, we relied on our Char04 to Cab08 comparisons, specifically the subset of SVs confirmed by both long-read and short-read alignments. We examined their mapping qualities, depths and likelihoods to provide empirical cut-offs for short-read SV calls detected by Lumpy[32] and Delly[33] using Char04 as the reference. After applying the cut-offs to the population sample, we filtered overlapping and complex SVs to obtain a highly curated set of 481,096 SVs for population analyses (Supplementary Table 4). These SVs yielded relationships among accessions that were remarkably similar to those based on SNPs, providing assurance about their reliability (Supplementary Fig. 5).

Given our population set of SVs, we computed the unfolded site frequency spectrum (SFS) for 12 *sylvestris* samples and a down-sampled set of 12 *sativa* samples chosen after genetic analysis

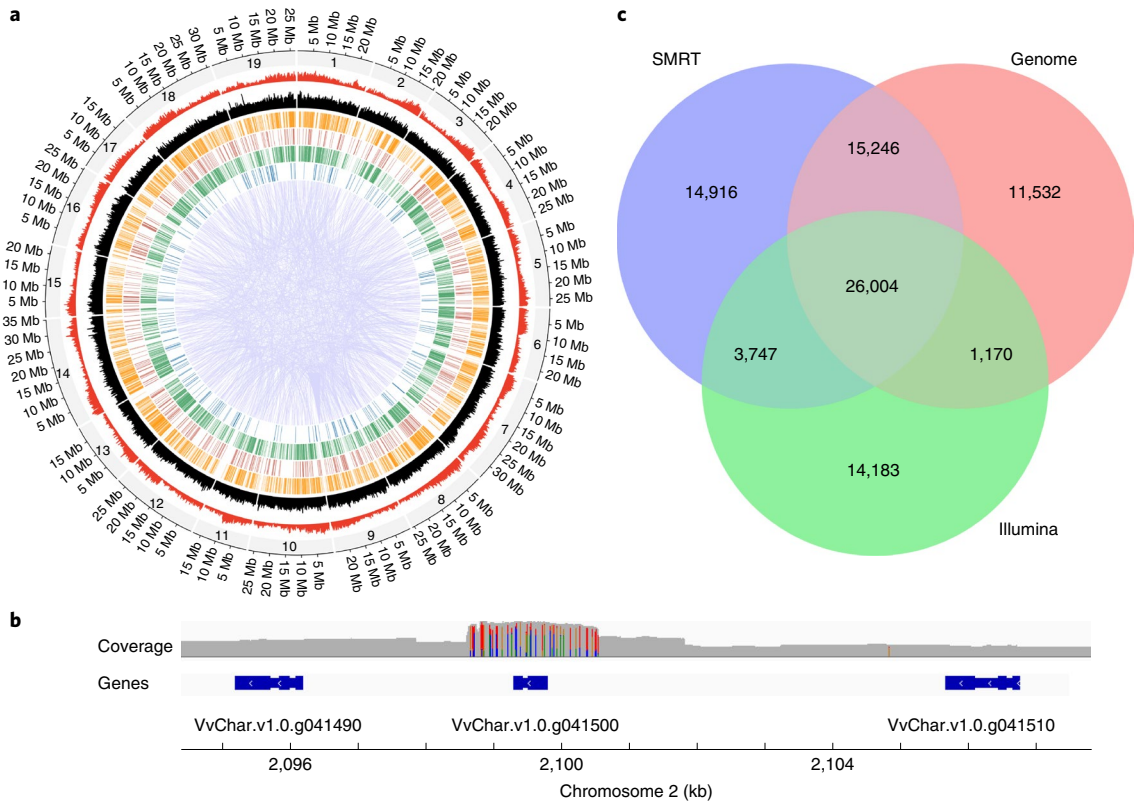

Fig. 1 | Structural heterozygosity in Char04 and comparisons of structural variation between Char04 and Cab08. a, The circle plot reports hSVs in the Char04 genome. The outermost circle denotes the number and size of chromosomes (grey), followed by gene density (red), TE density (black), deletions (orange), duplications (dark red), insertions (green), inversions (blue) and translocations represented in the middle of the circle in purple. **b**, A demonstration of hemizygous genes of Char04 supported by homozygosity, coverage and the breakpoints of SMRT reads. The vertical coloured lines in the grey coverage plot shows heterozygous sites. Both coverage and heterozygous sites support a heterozygous gene (VvChar.v1.0.g041500) and two hemizygous genes (VvChar.v1.0.g041490 and VvChar.v1.0.g041510). **c**, A Venn diagram showing the common and specific SVs detected by each method between Cab08 and Char04. The SVs shared between Illumina and SMRT calls provide the basis for criteria to identify SVs from the diversity panel.

(Supplementary Figs. 6–8). The SFS for the two taxa were similar overall (Fig. 2a), reflecting that cultivated grapevine did not undergo a severe domestication bottleneck[18,26] that can dramatically alter population frequencies. In both taxa, all SV types exhibited leftward shifts of the SFS relative to synonymous SNPs (sSNPs). The SFS of all SVs differed significantly from that of sSNPs in both taxa ($P < 0.05$, Kolmogorov–Smirnov, Bonferroni corrected), suggesting that SVs are predominantly deleterious.

To quantitate the strength of selection against SVs, we estimated the distribution of fitness effects (DFE) from population frequency data[34,35], using sSNPs as a neutral control. In both taxa, the results confirmed that non-synonymous SNPs (nSNPs) and SVs undergo strong purifying selection (Fig. 2b). They also revealed variation among SV types, because TRA events and INV events were more strongly selected against in both taxa, mirroring their more extreme SFSs. These inferences were also consistent with estimates of α, the proportion of adaptive variants, because α was estimated to be lower for INVs (<2%) and for TRAs (<7%) than for DUP ($\alpha = 25\%$ for *sylvestris*), DEL ($\alpha = 21\%$) and MEI ($\alpha = 20\%$) events (Fig. 2c). The α estimates for SVs were lower than those based on nSNPs (27% and 36% for *sylvestris* and *sativa*, respectively), which were comparable to other perennial taxa[36].

SVs accumulate in clonal propagants. SVs are deleterious, on average, but clonal propagation may allow variants to hide as heterozygous recessives[18,37]. The accumulation of recessive mutations was evident from three aspects of *sativa* genetic diversity. First, within individual heterozygosity was 11% higher, on average, in *sativa* than *sylvestris* based on SNPs (Supplementary Fig. 9). Second, sheltering of recessive mutations was evident from calculations of the additive SV burden, which is the number of heterozygous mutations plus twice the number of derived homozygous mutations per individual[38]. Individual cultivars have a 6% higher additive SV burden than their wild counterparts, on average, due to elevated heterozygosity (Fig. 3a). Enhanced burden was not evident for homozygous SVs or for presumably neutral sSNPs (Fig. 3a), suggesting that deleterious SVs accumulate and are sheltered in the heterozygous state. Finally, the SFS provided evidence of sheltering of recessive mutations in *sativa*, based on the marked reduction in frequency for any variants over 50% (12 alleles in Fig. 2a or 50 alleles in Supplementary Fig. 8, respectively). This unexpected observation may reflect features of the crossed state of heterozygous cultivars but it may also have a simple explanation: when a variant has a frequency over 50% in a clonally propagated population, then at least one individual must be homozygous, so that a recessive variant is exposed to negative selection.

The accumulation of heterozygous variants should affect linkage disequilibrium (LD), both because LD decreases as a function of population frequency[39] and because cultivated grapes tend to have more low-frequency variants than their wild counterparts (Fig. 2a). Consistent with this observation, LD decays more rapidly over physical distance for *sativa* than for *sylvestris*, despite the relative dearth of recombination via outcrossing in cultivars. LD also decays more rapidly for SVs than for SNPs in both taxa. Finally, LD between SVs and SNPs decays most rapidly of all.

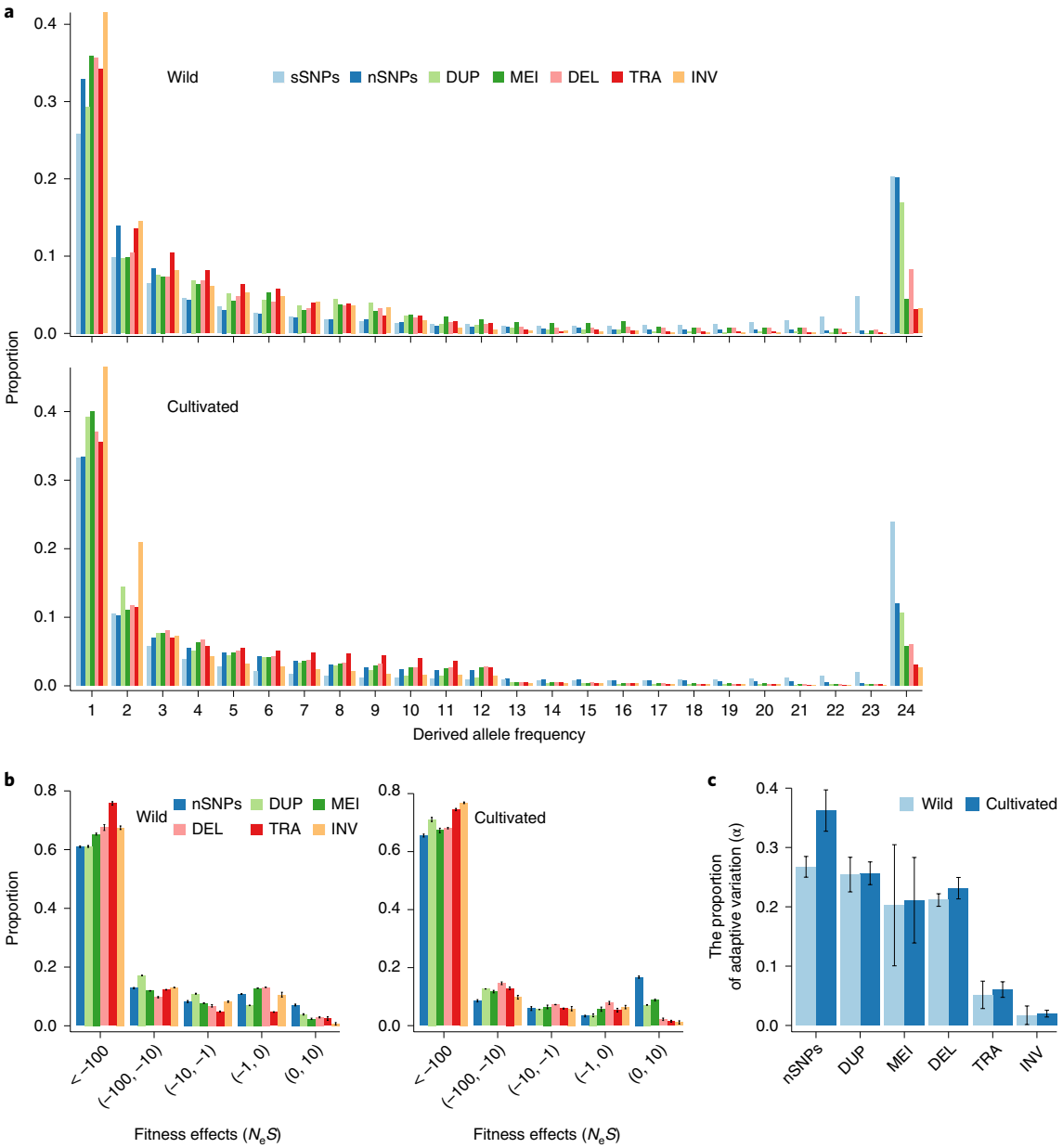

Fig. 2 | SVs are strongly deleterious and under purifying selection. a, The unfolded SFS of different types of SVs compared to presumably neutral sSNPs and nSNPs for samples of 12 wild (top) and 12 cultivated (bottom) grapevines. The types of SVs plotted include DUP, MEI, DEL, TRA and INV. **b**, The inferred distribution of fitness effects ($N_e s$) for SVs and nSNPs in wild (left) and cultivated (right) grapevines based on 100 bootstrap replicates. Error bars indicate the mean ± 95% CI. **c**, The proportion of adaptive variation (α) in wild and cultivated grapevines based on 100 bootstrap replicates. Error bars indicate the mean ± 95% CI.

SV outliers, domestication and sex determination. Cultivated grapevine differs phenotypically from its wild relatives[17]. In theory, the genes that contribute to these phenotypes can be inferred from population genetic data as regions of marked chromosomal divergence between wild and cultivated samples. We estimated both SNP and SV divergence across the genome, as measured by the fixation index (F_{ST}) in fixed windows of 20 kb (Fig. 3c). Overall, average F_{ST} estimates were substantially higher for SNPs (0.0354 ± 0.0165) than SVs (0.0135 ± 0.0066), reflecting that individual SVs are typically found at lower population frequencies (Fig. 2a).

We ranked the top 1% (or 485) F_{ST} windows for both SNPs and SVs. SNP-based windows generally conformed to a previous study[18] but SNPs and SVs both identified quantitative trait locus (QTL) regions on chromosome 2 that correspond to the sex-determination (SD) region and to the berry colour locus (Fig. 3c). An additional 410 SV-based windows were found on chromosomes 1, 2, 3, 4 and 5. Of these 410, only 81 (19.8%) overlapped with windows that also had significantly higher F_{ST} for SNP divergence. Based on gene ontology analyses, high F_{ST} windows were enriched for a few functional classes, including stilbenoid and folate biosynthesis. Stilbenes are particularly interesting because they accumulate in seeds and berry skin during berry ripening, vary in concentration between cultivars and include resveratrol[40], a component thought to have beneficial effects on human health. We also detected 78 diagnostic (or fixed) SVs between wild and cultivated samples that were associated with the gain and loss of seven and ten *sativa* genes, respectively (Supplementary Table 5). Among the ten lost, four were nucleotide-binding site leucine-rich

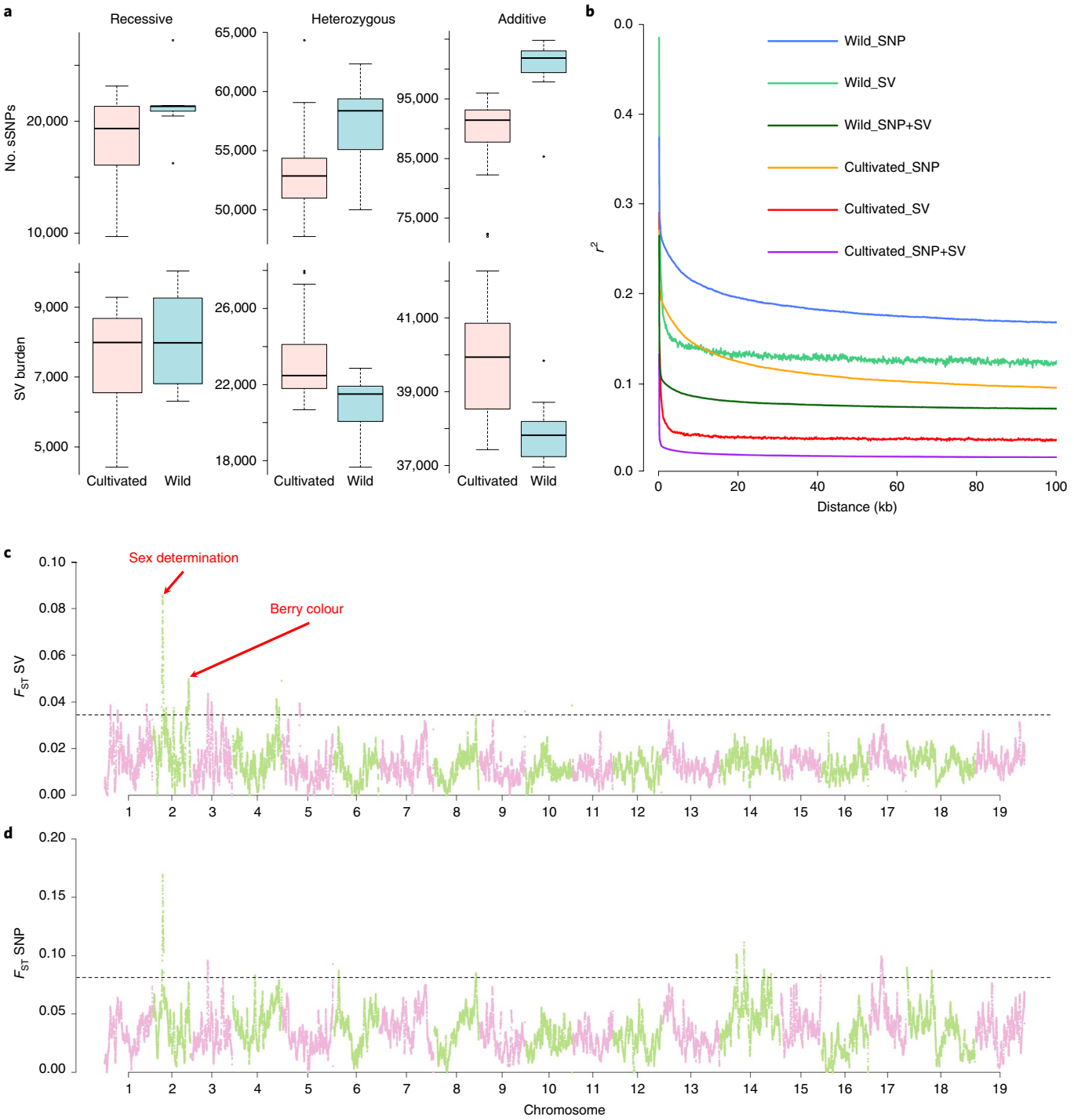

Fig. 3 | Population genetics of SVs associated with grapevine domestication. a, The recessive (number of homozygous SVs per grapevine), heterozygous and additive (number of hSVs plus two times the number of homozygous SVs per grapevine) burden in 12 wild and 12 cultivated grapevines for SVs compared to presumably neutral sSNPs. The middle bars represent the median, while the bottom and top of each box represent the 25th and 75th percentiles, respectively, and the whiskers extend to 1.5 times the interquartile range. Dots are outliers. **b**, The decay of LD, as measured by r^2, of SVs and SNPs as a function of physical distances between markers. **c**, Genetic differentiation (F_{ST}) between *sylvestris* ($n=12$) and *sativa* ($n=50$) sample across the genome based on F_{ST} of SVs in 20-kb sliding windows. The dashed horizontal line represents the cut-off for the 1% tail of the F_{ST} distribution. Peaks of divergence corresponding to the sex region and the berry colour loci are indicated. The *x* axis indicates the number and size of chromosomes across the genome. **d**, The same as **c**, except genetic differentiation is based on SNP data.

repeat disease resistance genes located between 11,053 and 11,064 Mb on chromosome 9 of PN40024.

The highest F_{ST} peak for SVs corresponded to the SD region on chromosome 2 (Fig. 3c), which also contained more SV events relative to the genomic background ($P=0.0067$; χ^2). Mutations in the SD region caused the shift in mating system during domestication. After confirming that the sex-linked region corresponds to 4.90 Mb and 5.04 Mb on PN40024 (refs. [41,42]; Supplementary Fig. 10),

we resolved complete SD haplotypes and their underlying SVs. Chardonnay is rare among cultivars because it is a homozygote for the hermaphroditic (H) haplotype[42]. These two haplotypes illustrate that the region is replete with structural variants; they differ by 37 non-genic indel difference of >50 bp in length, including the insertion/deletion of four putative TEs.

Given the complexity of the region, we focused on genic PAV variation and compared the two Char04 H haplotypes to the PN40024 primary assembly[43], which is thought to represent the female (F) haplotype[42]. Four genes exhibited PAV variation between the H and F haplotypes. One of these, Vvi*APT3*, has been proposed as a candidate SD gene[41] because it may have a role in the abortion of pistil structures[44] but it was missing from the SD region of the PN40024 reference. In our data, Vvi*APT3* was present in both the H and F haplotypes of Cab08 (Fig. 4a), suggesting that the lack of Vvi*APT3* on PN40024 was an assembly error. The remaining three PAV genes (a *DEAD DEAH* box RNA helicase gene, the TPR-containing protein and the unknown protein previously known as *ETO1*) differentiated H from F haplotypes (Fig. 4a). We also annotated two previously unrecognized genes, *Inaperturate pollen 1* (Vvi*INP1*) and a *C2H2*-type Zinc finger, in both F and H haplotypes. *INP1* expression in *Arabidopsis* alters the deposition of pollen apertures[45] and could confer pollen sterility in females.

Hermaphroditism was likely to be caused by a mutation in the dominant F sterility gene on the male (M) haplotype[42,46]. The female sterility gene is unidentified but it is probably expressed in males and knocked-down in hermaphrodites. To identify potential candidates, we performed gene expression analyses among sexes, based on expression data from two late stages of floral development (Fig. 4b). The three PAV genes were lowly expressed and thus are unlikely F sterility candidates but five genes differed significantly (adjusted $P \leq 0.05$; see Methods) in sex-specific expression. Four were more highly expressed in males, including Vvi*APT3* and the *C2H2*-type Zinc finger gene; these four constitute plausible female sterility candidates.

To investigate whether any of these candidates housed a loss-of-function SV, we built a phylogeny of the SD region, which confirmed that H haplotypes were closer to the the M haplotype from our single, confirmed *sylvestris* male than to F haplotypes (Fig. 4c). In fact, the M haplotype separated two clades of H haplotypes, providing support for more than one origin of hermaphroditism in cultivars[42]. We estimated the two clades to be 10,705 and 13,222 years old, respectively, slightly older than the accepted date of domestication. Because the *sylvestris* M haplotype was closely related to one of the Char04 haplotypes (Fig. 4c), we identified SVs in and between them. Four genes were in a hemizygous state in the wild male, including the three PAV genes, and there were also three hemizygous TEs near genes. Unfortunately, none of these SVs were obvious candidates to affect the function of the four most plausible female sterility candidates (Fig. 4b). The genetic mutation(s) that caused hermaphroditism remain unidentified but this region underscores the dynamic nature of SV events in grapevine.

Large, independent inversions drive convergent evolution of berry colour. A second region of high F_{ST} divergence between wild and cultivated grapevines encompassed the berry colour region (Fig. 3c). It, too, had more SVs than the genomic background ($P = 3.3 \times 10^{-5}$, χ^2). The region is interesting because *sylvestris* has dark berries, representing the ancestral condition[17], and because white berries originated in a subset of *sativa* cultivars like Chardonnay. SVs have been implicated in the origin of white berries, especially a 5′*Gret1* retroelement insertion that reduces the expression of a *myb* gene (Vvi*MYBA1*) that regulates anthocyanin biosynthesis[47]. Subsequently, it was shown that a frameshift mutation in a second *myb* gene (Vvi*MYBA2*) was also necessary to cause white berries[48]. Surprisingly, these two mutations (the *Gret1* insertion and the Vvi*MYBA2* frameshift) are heterozygous in most grape cultivars[49]. Somatic mutations causing white grapes delete the functional Vvi*MYBA1* and Vvi*MYBA2* alleles, leaving the plant hemizygous for null alleles[50,51].

Given the history of the *MybA* locus and the fact that it encompasses a peak of F_{ST} divergence, we investigated the region with a chromosome scale plot of Char04 reference versus Cab08, revealing a large 4.82 Mb (chr02: 12,295,113–17,118,777 bp) inversion in Char04 (Fig. 5a). This inversion was confirmed by comparison to PN40024, by the identification of discordant and split reads at the junctions (Supplementary Fig. 11) and by the lack of an inversion between Cab08 and PN40024 (Fig. 5b). The Char04 inversion was bounded by *copia* elements, suggesting they played a role in its formation. The inversion encompassed the *MybA* region but it did not affect the number of *MybA* genes because there were nine in Char04, Cab08 and PN40024. The inversion does, however, affect hemizygosity, because the entire inverted region appears to be hemizygous on the basis of read coverage and homozygosity (Supplementary Fig. 11). Thus, white berries in Chardonnay may be attributable to two related events, a large inversion on one chromosome and a simultaneous deletion on the other.

A previous study characterized the somatic mutations that led to white berries in the Tempranillo cultivar[52]. The mutations included hemizygosity at four *MybA* genes (Vvi*MybA1*, Vvi*MybA2*, Vvi*MybA3* and Vvi*MybA4*), along with a series of complex series of SVs that included a putative 4.3-Mb inversion on chromosome 2 (Supplementary Fig. 12). Given that both Chardonnay and Tempranillo have large, Mb-scale inversions associated with white berries, we investigated the generality of the association. To do so, we first built SNP-based phylogenies of white-berried cultivars and closely related dark-berried varieties. Not surprisingly, the phylogeny shows that white berry mutations occurred independently on several occasions (Fig. 5d). We then chose six pairs of closely related dark- and white-berried varieties and contrasted them using short-read analyses. For these short-read analyses, we focused on coverage and runs of homozygosity, while also carefully combing the data for evidence of split and discordant reads that span potential inversions. All five contrasts yielded evidence for a large inversion encompassing the *MybA* region (Fig. 5c). The inferred inversions ranged from 3.85 Mb to 4.82 Mb in size and included from 134 to 176 genes, with 118 genes in common (including the *MybA* genes) across all six inversions. Read coverage data, which varied across pairs, strongly suggested hemizygosity of the entire inversion in at least one contrast (Sultanina versus Kishmish vatkana) and near the *MybA* region in other contrasts (Fig. 5c).

Discussion

Our analyses of SVs from genomes and population samples of grapevines provide insights into several features of plant genome evolution, domestication and phenotypic change. First, we find that SVs are common enough that ~1 in seven genes are hemizygous in a single individual, that two distinct cultivars (Chardonnay and Cabernet Sauvignon) differ in PAV for roughly 5% (2,217 of 38,020) genes and that ~25% (9,330 of 38,020) of genes varied in hemizygosity between these same two cultivars. All of these values were based on long-read sequencing data and supported by split-read analyses. To date, there have been few explicit comparisons between individual genomes based on long-read data and there have been even fewer to analyse hemizygosity in heterozygotes. Nonetheless, the high number of PAV and hemizygous genes is not particularly surprising, based on two pieces of evidence. The first is that previous studies have hinted at high PAV in grapevine. For example, transcriptomic sequencing of the Tannat cultivar revealed >1,800 genes that were not present in the Pinot Noir reference[30] and a recent study has demonstrated high hemizygosity in one grapevine cultivar[22]. The second is that recent studies in other plant species

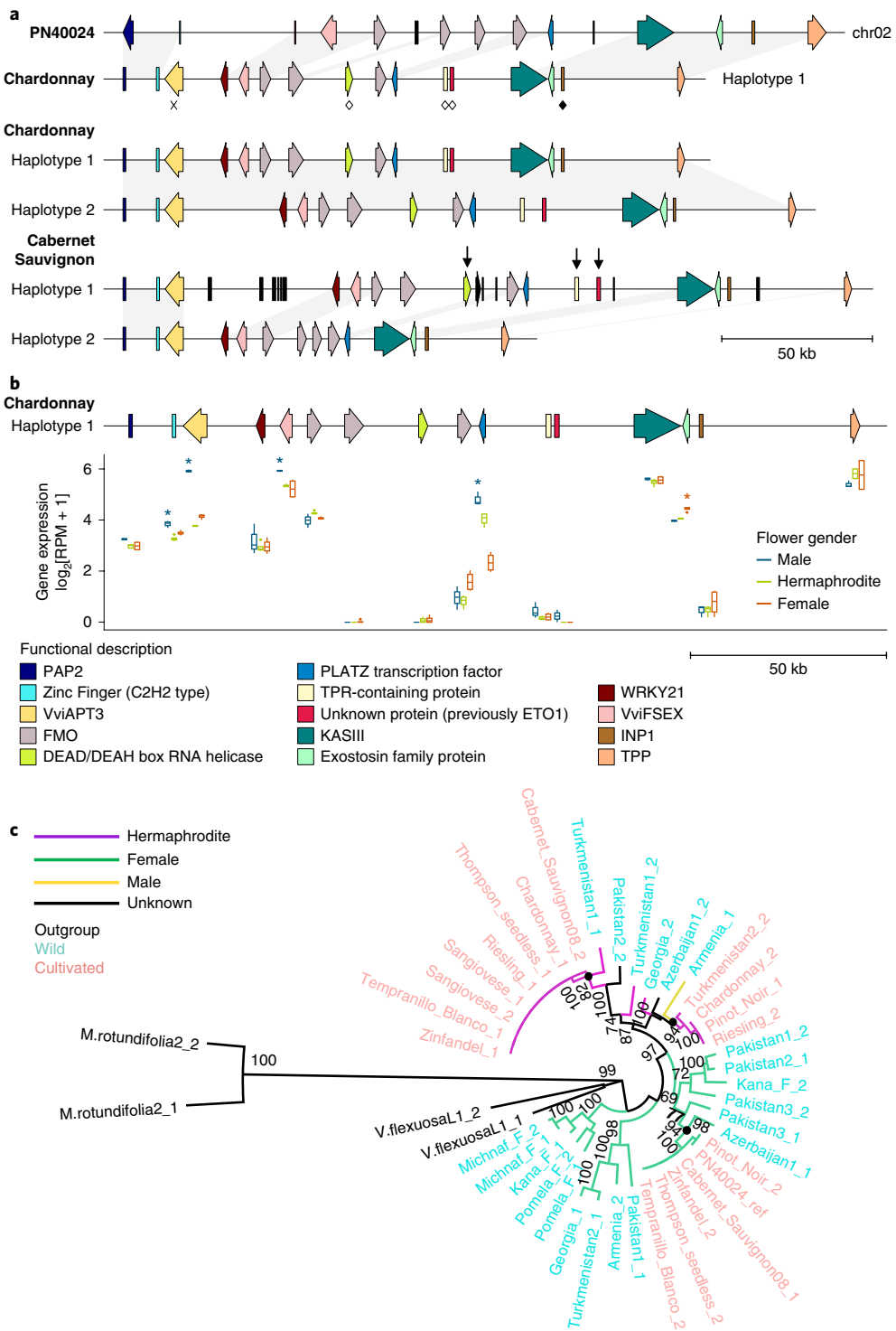

Fig. 4 | Haplotypes of the sex region and the evolution of sex in grapevine. a, Comparison of the SD region among cultivars. The PN40024 (V2) haplotype represents the primary assembly. Chardonnay is homozygous hermaphroditic (HH) and both haplotypes from Char04 are shown. Cabernet Sauvignon is heterozygous (HF), with Haplotype 1 of Cab08 representing the presumed H haplotype. A cross (X) denotes the gene Vvi*APT3* that is absent from PN40024 assembly but found in both F and H haplotypes; open diamonds denote the genes located on chromosome 0 in the PN40024 assembly and the filled diamond denotes a new functional annotation in Char04 (*INP1*). Protein-coding genes are coloured according to their functional annotation. Genes that are not shared among genome assemblies are coloured in black. Black arrows highlight genes that are found on inferred H haplotypes in Chardonnay and Cabernet Sauvignon. **b**, Gene expression values of each flower sex type projected on the Chardonnay protein-coding genes are shown at both G (flowers closely pressed together) and H (flowers separating, just before blooming) stages as $\log_2$[RPM + 1]. Asterisks denote a significant difference ($P \leq 0.05$) based on DESeq2 using the Benjamini–Hochberg adjustment. For box plots, the middle bar represents the median, the bottom and top of each box represent the 25th and 75th percentiles, respectively, and the whiskers extend to 1.5 times the interquartile range. Dots are outliers. **c**, A phylogeny of the SD region recapitulates known sex types for cultivars and detects two H clades split by the single known male in the wild sample, suggesting more than one origin of the H type.

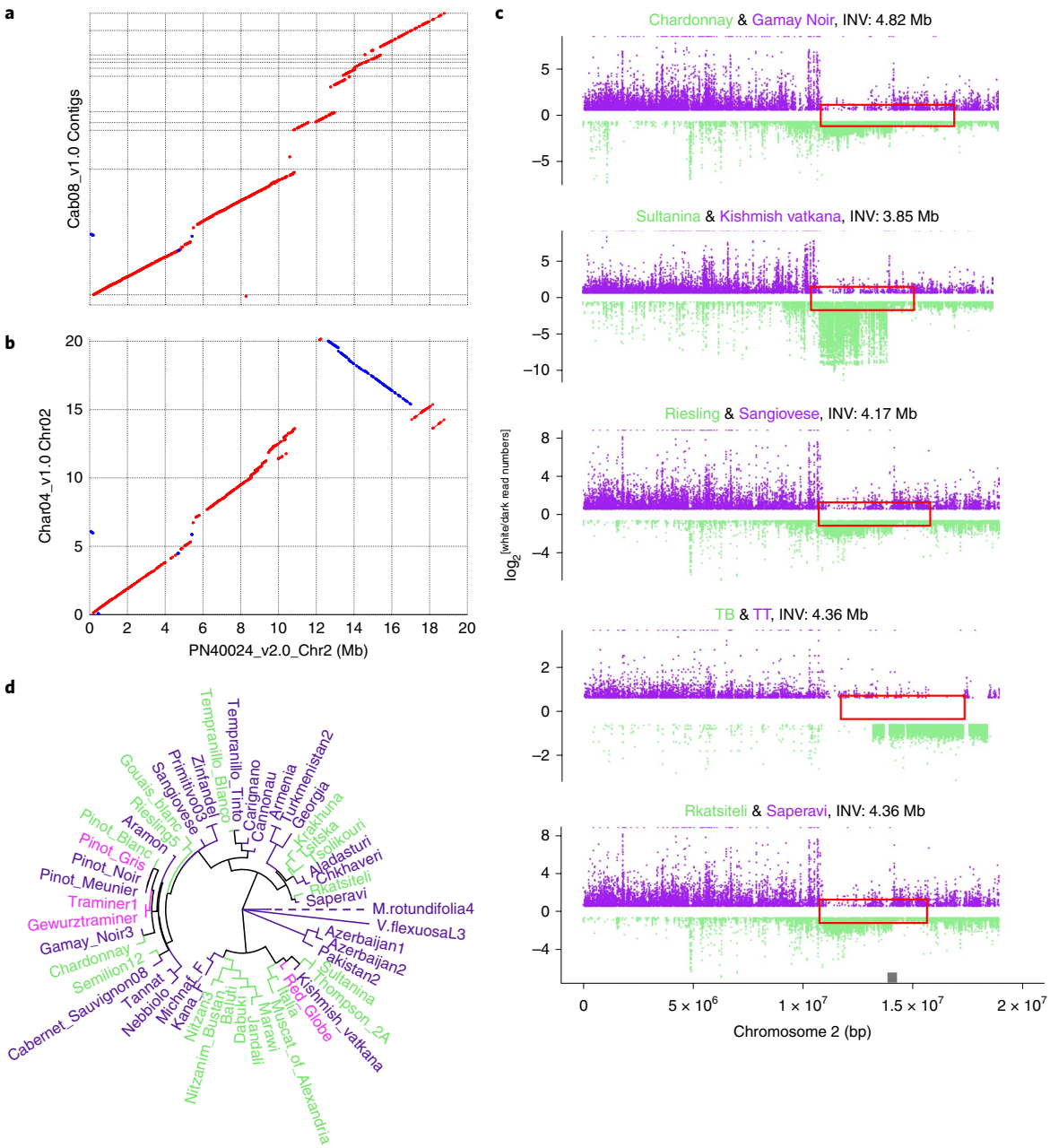

Fig. 5 | Convergent evolution of inversions associated with white berries. a, A dot plot between PN40024 chromosome 2 and Cab08 contigs. **b**, A dot plot between PN40024 chromosome 2 and Char04 chromosome 2 that reveals a 4.82-Mb inversion overlapping with the major berry colour QTL in grapevines. **c**, These plots contrast coverage across chromosome 2 for pairs of white berry and dark berry grapevines. In each contrast, the white berry grape is labelled in green. The *y* axis is the $\log_2$[white/dark read numbers] so that, for example, regions of very low values indicate relatively few reads in the white berry grape. For each contrast, the size of the inferred inversion is provided, based on the presence of split reads. The red open box indicates the approximate locations of inferred inversions and the black filled box at the bottom shows the location of the *MybA* genes on PN40024. TB, Tempranillo Blanco; TT, Tempranillo Tinto. **d**, A phylogeny based on genome-wide SNPs from a selection of grape varieties, with the colour of text labels reflecting the colour of the berry.

have documented high PAV between individuals. For example, a comparison of two maize inbred lines (B73 and Mo17) has revealed that 10% of genes are mutually non-syntenic and that ~17% of B73 genes contain large-effect mutations relative to Mo17, including the loss of exons[12].

Thus, the emerging picture is that PAV is rampant in plant genomes but two important questions remain. The first is whether grapes are extreme in their levels of hemizygosity and PAV. We suspect, but do not yet know, that PAV is a more prominent feature of clonally propagated perennials than selfed and inbred lines[12,53–55]. A test of this conjecture requires wider species sampling and more explicit comparisons. One potentially unique feature of grapevines is the exceptionally long history of some cultivars[56], which provides an extended opportunity for hemizygosity to accumulate. The second question is about the functional consequences of genomic hemizygosity and PAV variation, which remains unexplored on a genome-wide scale for any plant taxon. Our data nonetheless hint at the potential for these genes to contribute to phenotypic differences

between cultivars because they are enriched for functional classes, such as 'defence response', that probably contribute to agronomic properties. Identifying the functional effects of hemizygous and PAV genes is an important emerging frontier[12].

We used the comparison between Chardonnay and Cabernet Sauvignon to help filter SV calls on an expanded sample of cultivated *sativa* and wild *sylvestris* accessions. We showed that 75% of long-read based SV calls between Char04 and Cab08 were verified by at least one of two other comparative methods—genome alignment and short-read data (Fig. 1c). Focusing on the SVs that were detected by both short- and long-read analyses, we then applied the criteria of these 'gold standard' SV calls to a short-read resequencing dataset of 69 accessions. The result was a set of 481,096 SVs of >50 bp from throughout the genome that recapitulated relationships among accessions based on SNPs (Supplementary Fig. 5). Given these population genetic data, we inferred the strength of selection against SV types and also contrasted SV frequencies between the domesticate and its wild progenitor, a contrast that has not been performed previously in other crops[9].

Overall, we found that selection acts against SVs but with some variation among SV types. For example, we infer that MEI, DEL and DUP SVs are selected against with selection coefficients similar to, but slightly larger than, nSNPs (Fig. 2b). In comparison, selection is stronger against TRAs and INVs than other SV types, with correspondingly lower rates of presumed adaptive events (Fig. 2c). Surprisingly, negative selection appears to be stronger in *sativa* than *sylvestris*, based on DFE and α estimates (Fig. 2). This comparison between taxa must be interpreted with caution, however, because the inferential models were designed to analyse outcrossing species like *sylvestris* and not clonally propagated crops. Nonetheless, the results consistently suggest that SV events are more deleterious than nSNPs, on average, and that INV and TRA events are especially deleterious.

Previous work has shown that domesticates accumulate deleterious variants[9,18,57–59]; this is particularly true of clonally propagated crops that can hide deleterious recessive variants in a heterozygous state[9,18,37]. Given our inference that SVs are generally deleterious, we therefore expected that they accumulate in the heterozygous state in *sativa* relative to outcrossing *sylvestris*. The evidence supports this expectation: cultivars had 11.31% more heterozygous variants than wild samples (Supplementary Fig. 9), on average, and as a result accumulated ~5.3% more SVs than *sylvestris* acessions (Fig. 3a), a difference nearly identical to that found for putatively deleterious SNPs[18]. Hence, grapevine cultivars contain a higher SV burden than wild accessions, just as they do for nSNPs[18]. The SFS of SVs (Fig. 2a) may contain an interesting clue about the potential recessive nature of these mutations, because the *sativa* sample demonstrates an abrupt decline of SV proportions when the population frequency is above 50% (Fig. 2a). This abrupt decline is more noticeable for SVs and nSNPs than for sSNPs, which are presumably predominantly neutral. This decline is consistent with the exposure of recessive SVs becoming visible to selection in the homozygous state but it may also simply reflect different aspects of genetic history. Further work on grapevine and on clonal plants more generally, will help to elucidate whether this is a common property of clonally propagated crops.

Although several studies have reported the number of SVs in plant resequencing datasets, fewer papers have measured SV population frequencies[9,60,61]. These frequencies are important if one hopes to use association-based analyses to infer causative SVs that affect phenotypes. Similarly, it is critical to know if SNP-based genome-wide association (GWAS) assays often tag causative SVs. We took a preliminary look at this issue by measuring the decline of LD over physical distance for SVs, SNPs and a combined dataset (SVs + SNPs). For grapevines, at least, the results are discouraging, because: (1) LD declines more rapidly for cultivars than for wild plants[62]; (2) LD declines more rapidly for SVs than SNPs in both *sativa* and *sylvestris*; and (3) the fastest rates of decline are for the combined SV + SNP datasets (Fig. 3b). These rapid LD declines reflect the fact that SVs are typically at lower population frequencies than SNPs, owing to their deleterious effects. We do not know yet if LD patterns for SVs are similar for other crops. If we assume that SVs are generally deleterious in plant populations, then the rapid decline of LD over physical distance for SVs is expected to be a general phenomenon.

Finally, we used the 481,096 SVs to investigate regions of high F_{ST} divergence between domesticated grapevine and our sample of putatively wild accessions. Many of the SV-based peaks of F_{ST} divergence corresponded to those found previously with SNPs[18], especially the SD region on chromosome 2 (Fig. 3c,d). Careful investigation of this region revealed genic PAV between M, H and F haplotypes and helped to narrow the search for candidate SD genes (Fig. 4). However, some peaks of divergence differed between SVs and SNPs, including peaks that were enriched for stilbenoid and folate biosynthesis genes, which may contribute to nutritional value of the berries. Another surprising SV peak centred around the berry colour locus. Further investigation of this region revealed a 4.82-Mb inversion in Chardonnay (Fig. 5b) and evidence to support that the independent origin of white berries (Fig. 5d) is commonly mediated by such inversions (Fig. 5d). These large inversions may explain why one GWAS analysis found associations to berry colour over a 10-Mb region of chromosome 2 (ref. [63]). As documented previously for a somatic mutation to white berries in the Tempranillo cultivar[52], these inversions result in hemizygosity of the *MybA1* and *MybA2* null alleles (Supplementary Fig. 12).

We have established that somatic mutations to white berries are associated with hemizygosity of *MybA* genes and with large, Mb-scale inversions. But the bigger question is why large inversions mediate these somatic shifts in berry colour. We can think of three explanations. The first is that the inversion contains non-*MybA* genes that also affect phenotype. The inversions in our datasets contain 118 genes in common across five independent contrasts. To investigate this hypothesis, we mapped gene expression data from dark and white berries collected over four stages of berry development[64] and counted the proportion of differentially expressed genes between colour morphs. The proportion of differentially expressed genes in the Char04 inversion was no higher than the genomic background ($P = 0.82$, χ^2), suggesting that the inversion is not enriched for genes that contribute to berry colour. The second explanation is that inversions are common because of underlying properties of the chromosome 2 sequence, such as the enhanced fragility recently documented in some chromosomal regions of stickleback fish[65]. The berry colour region does not contain any obvious differences in TE distribution or other gross features suggesting it is particularly labile (Fig. 1a) but this explanation remains a possibility given that *copia* elements flank the inversion in Char04. Finally, it is possible that similar somatic inversions occur commonly in *Vitis* but most are lost because even small inversion events are strongly selected against (Fig. 2b). However, a few such inversions may affect an obvious phenotype—like berry colour—that are then prone to human selection. Whatever the underlying cause(s) for these large inversions, they represent a stunning example of convergent evolution via independent, complex SV events.

Methods

Genome sequencing, assembly and polishing. The Chardonnay clone chosen for sequencing was FPS 04, a clone commonly grown in California and throughout the world. The reference plant is located at Foundation Plant Services, University of California, Davis. DNA isolation and the preparation of SMRTbell libraries followed ref. [21]. The preparation of paired-end (PE) Illumina libraries followed ref. [18]. SMRTbell libraries were sequenced on a Pacific Biosciences (PacBio) RSII system, generating a total of 31.51 Gb (52×). Illumina sequencing was conducted on a HiSeq4000 sequencing platform in 150 PE mode (54×) and 100 PE mode

(124×). Both SMRTbell and Illumina libraries were sequenced at the UC Irvine High Throughput Genomics Center. Raw reads were deposited to the Short Read Archive (SRA) at the NCBI under the BioProject ID PRJNA550461.

Genome assembly was based on a hybrid strategy, that used both long and short sequencing reads, and that merged three separate assemblies. The first assembly used Canu v.1.5 (ref. [66]) to assemble SMRT reads, based on default parameters and with a genome size of 600 Mb. A second, hybrid assembly was generated with DBG2OLC[67] based on contigs from the Platanus assembly and the longest 30× PacBio reads. The Platanus assembly was based on ref. [68] v.1.2.4 with default settings, using trimmed 178× Illumina PE reads. The DBG2OLC settings (options: k 31 AdaptiveTh 0.01 KmerCovTh 2 MinOverlap 30 RemoveChimera 1) were similar to those used for previous hybrid assemblies[69,70], except that the *k*-mer size was increased to 31. The *k*-mer size was increased to minimize the number of mis-assemblies by including 90% of all *k*-mers reported by the meryl programme in the Canu package[66]. The consensus stage for the DBG2OLC assembly was performed with PBDAGCON[71] and BLASR[72]. Third, PacBio genomic reads were assembled using FALCON-Unzip v1.7.7 (ref. [21]). Multiple assembly parameters (length_cutoff_pr) were tested; the least fragmented assembly was obtained with a minimum length cut-off of 9 kb. The final FALCON-Unzip parameters can be found in Supplementary Text 1. Unzip phasing and haplotype separation were performed with default parameters.

To integrate information obtained from the different assembly methods—Canu, DBG2OLC and FALCON-Unzip—we opted for an iterative approach of assembly merging using quickmerge[69], following a broader application of assembly merging based on ref. [70]. Quickmerge merges assemblies to increase the contiguity of the most complete (query) genome by taking advantage of the contiguity of the second reference sequence. To merge the assemblies, we followed a series of steps. First, the DBG2OLC and Canu assemblies were merged into a single assembly, QM1, using DBG2OLC assembly as the query, the Canu assembly as the reference and run options (options: hco 5.0 c 1.5 l 260000 ml 20000). Contigs that were unique to the Canu assembly were incorporated in the subsequent assembly, QM2, by a second round of quickmerge (options: hco 5.0 c 1.5 l 260000 ml 20000). In this second quickmerge run, the merged assembly from the previous step, QM1, was used as the reference assembly, and the Canu assembly was used as the query. A third round of merging (options: hco 5.0 c 1.5 l 345000 ml 20000) was performed using primary contigs of FALCON-Unzip as the reference assembly and the previous resultant assembly, QM2, as the query, generating the QM3 assembly. The final assembly, QM4, was generated by a fourth run of quickmerge (options: hco 5.0 c 1.5 l 345000ml 20000), using QM3 as the reference and the FALCON-unzip assembly as the query.

All the assemblies described above, including the preliminary assemblies (Canu, DBG2OLC and Falcon-Unzip), temporary assemblies (QM1–QM3) and the final assembly (QM4), were polished twice with long reads using Quiver (Pacific Biosciences) from SMRT Analysis v.2.3 (using parameter: -j 80). Long reads (>1,000 bp), consisting of ~43× coverage, were used for polishing. The assemblies were also polished twice using Pilon v.1.16 (ref. [73]) run using default settings. For this purpose, Illumina reads were aligned to the assembly using Bowtie2 v.2.32 (ref. [74]) and sorted using samtools v.1.3 (ref. [75]).

BUSCO v.2.0 was used to measure gene space completeness and conserved gene model reconstruction of all generated assemblies[76]. The embryophyta database, which contained 1,440 highly conserved genes, was used to measure gene model reconstruction and estimate assembly completeness. Quast v.2.3 (ref. [77]) was run to calculate assembly length and N50 on each assembly. Dot plots were generated using nucmer and mumplot from MUMmer v.3.23 (ref. [31]) with the options: -l 100 -c 1000 -d 10 -banded -D 5. The BUSCO v.3 (ref. [76]) pipeline was applied to the final genome assembly, using the embryophyta_odb9 database.

The final assembly included both primary haplotype sequences and alternative contigs (otherwise known as haplotigs). To remove some of the alternative contigs and minimize redundancies, we performed a contig reduction. Contig reduction was executed by first aligning the final assembly to itself using Blat v.36 (ref. [78]). A python script was generated for filtering contigs that did not meet one minimum and two maximum thresholds: contig length, %query alignment and %alignment overlap. In practice, the three thresholds were investigated over ranges—for example, minimum contig length ranged from 0 to 100,000 bp; percentage query alignment (PctQA) was examined over 18 randomly chosen values between 90 and 99.9999%, and percentage aligned overlap (PctAO) (80 and 90%), as well as maximum PctQA (100%) and PctAO (110 and 120%). New filtered genome assemblies were generated after filtering contigs based on a combinatorial of these five parameters. A gradient descent was performed on three additional parameters generated for each new filtered assembly; assembly size, contig N50 and BUSCO scores. Two formulae were generated to calculate PctQA and PctAO. $\text{PctQA} = \frac{\text{Aligned Query Length}}{\text{Total Query Length}}$ and $\text{PctAO} = \frac{\text{Aligned Query Length}}{\text{Aligned Reference Length}}$. Alignments generated from contigs aligning to themselves were not considered.

Scaffolding and gap closing. A Dovetail Hi-C library was prepared in a similar manner as described previously in ref. [79]. The library was sequenced on an Illumina platform to produce 211×10^6 2 × 100 bp PE reads, which provided 1,624× physical coverage of the genome (1–50 kb pairs). The input de novo assembly, shotgun reads and Dovetail Hi-C library reads were used as input data for HiRise[80]. Shotgun and Dovetail Hi-C library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Hi-C read pairs mapped in draft scaffolds were analysed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins and make joins above a threshold. After scaffolding, shotgun sequences were used to close gaps between contigs.

MUMmer v.4.0 (ref. [31]) was used to identify and to sever erroneous junctions between contigs. The resulting scaffolds underwent a second scaffolding procedure using SSPACE-longreads v.1.1 (ref. [81]) with default parameters and a minimum coverage of ten reads (options: -l 10). Gaps were closed using PBjelly (PBSuite v.15.8.24)[82] with default parameters for all the gap-closing steps and assembled with options: '-x -w 1000000 -k -n 10'. Scaffolds were again manually curated as described above.

Gene annotation. Repetitive sequences were identified with RepeatMasker[83] using the repeat library previously developed for *V. vinifera* cv. Cabernet Sauvignon[84]. Ab initio prediction of protein-coding genes was carried out with SNAP v.2006-07-28 (ref. [85]), Augustus v.3.0.3 (ref. [86]) and GeneMark-ES v.4.32 (ref. [87]). Ab initio predictions were combined with the predictions of Augustus trained with BUSCO genes, as well as the gene models annotated with PASA v.2.1.0 (ref. [88]), using the experimental data reported in Supplementary Text 2. RNA-seq data obtained from public databases (Supplementary Text 2) were: (1) assembled using both an on-genome strategy, with Stringtie v.1.3.3 (ref. [89]), and a de novo transcriptome procedure, with Trinity v.2.4.0 in genome-guided mode setting a maximum intron length of 10 kb (option:–genome_guided_max_intron 10000); (2) clustered with CD-HIT-EST v.4.6 (ref. [90]), with coverage threshold 90% (option: -c 0.9); and (3) filtered with Transdecoder v.3.0.1 (ref. [91]), which retained only genes with a full-length open reading frame (ORF). Experimental evidences (transcripts and proteins) were mapped on the genome using Exonerate v.2.2.0 (ref. [92]) and PASA v.2.1.0 (ref. [88]), and together with all the predictions used as input to EVidenceModeler v.1.1.1 (ref. [93]). Weights used in EVidenceModeler are reported in Supplementary Text 3. The annotation was refined and enhanced with alternative transcripts using PASA v.2.1.0 (ref. [93]) and assembled experimental evidences; parameters used for refining the gene structures are described in Supplementary Text 4. Models not showing a full-length ORF from start codon to stop codon or showing in-frame stop codons were removed. Transcripts were blast-searched for homologue proteins in the RefSeq plant protein database (ftp://ftp.ncbi.nlm.nih.gov/refseq, retrieved 17 January 2017). Functional domains were identified using InterProScan v.5 (ref. [94]) using the databases provided in Supplementary Text 5. Gene models with no significant blast hit against RefSeq plant protein database (high-scoring segment pair length < 50 amino acids) and lacking any functional domain were discarded. Gene ontology obtained from InterPro domains and RefSeq homologues with at least 50% of reciprocal coverage and identity were combined using Blast2GO v.4 (ref. [95]) to assign a functional annotation, gene ontology and enzyme commission descriptions to each predicted transcript.

Chromosome assignment and heterozygosity in the Chardonnay genome. The Char04 primary assembly consisted of 684 scaffolds, that summed to 606 Mb with an N50 close to that of an average grape chromosome size (25.4 Mb). We aligned the Char04 primary assembly to the PN40024 genome using the nucmer function in MUMmer4 (ref. [31]). The top 23 scaffolds covered 82% (492 Mb) of the Char04 primary assembly and aligned to the PN40024 chromosomes (Supplementary Fig. 1), except two long scaffolds with lengths of 20 Mb (Char04v1.0_683) and 11 Mb (Char04v1.0_682). These two scaffolds did not align to PN40024 genome assembly but did align to Cab08 contigs. At the same time, chromosome 13 of the PN40024 genome aligned to only a few small Char04 scaffolds for the purposes of presentation (Fig. 1).

The largest 22 scaffolds of Char04 were collinear with PN40024 and summed to 481 Mb. Each chromosome was represented by one scaffold, except chromosomes 7 and 11, which consisted of two and three scaffolds, respectively. For all ensuing analyses, we treated these 22 scaffolds as the Char04 reference genome. We evaluated heterozygosity in this reference for both small variants (SNPs + indels < 50 bp) and large structural variants (SVs ≥ 50 bp). SNPs and indels were called on the basis of remapping 124X Illumina 100-bp PE reads to the reference. The Illumina reads for this application and for diversity analyses (see below) were trimmed using Trimmomatic-0.36 to remove adaptor sequences and bases for which average quality per base dropped below 20 in 4-bp windows. Filtered reads were then mapped to the Char04 reference with default parameters implemented in bwa-0.7.12 using the BWA-MEM algorithm[96]. The bam files were filtered (unique mapping with a minimum mapping quality of 20) and sorted using samtools v.1.9 (ref. [55]). PCR duplicates introduced during library construction were removed with MarkDuplicates in picard-tools v.1.119 (https://github.com/broadinstitute/picard). SNPs and small indels were called with the HaplotypeCaller in GATK v.4.0 pipeline and then filtered following ref. [18].

To identify SVs in the Char04 genome (that is between the two haplotypes), we called SVs using the Sniffles pipeline[24]. First, PacBio reads longer than 500 bp were mapped onto Char04 primary assembly using the two aligners Minimap2 v.2.14

with the MD flag[97] and NGMLR v.0.2.7 (ref. [24]), separately. Variant calling was then performed with Sniffles. SV analysis outputs (VCF files) were filtered based on the following four steps: (1) we removed SVs that had ambiguous breakpoints (flag: IMPRECISE) and also low-quality SVs that did not pass quality requirements of Sniffles (flag: UNRESOLVED); (2) we removed SV calls shorter than 50 bp; (3) we removed SVs with less than four supporting reads; and (4) we removed duplicate SV calls from Sniffles (Sniffles frequently called multiple SVs at the same position for multiple pairs of breakpoints. In these cases, we kept the SV with the most supporting reads). The same filtering steps were applied in downstream analyses when we called SVs between Cab08 and Char04 primary assemblies (see below). In general, using the aligner Minimap2 from the Sniffles pipeline led to detecting more SVs (for example, 37,169 in total in Char04) than long-read mapping with NGMLR v.0.2.7 (23,972 in total in Char04). Given the differences from the two mapping protocols, we built consensus SVs calls using SURVIVOR v.1.0.3 (ref. [98]). Using the merged SV set, we called genotypes and combined them into a single VCF using the population calling steps of the Sniffles pipeline[24]. The genotypes of SV calls from both programmes (NGMLR and Minimap2) were intersected using bedtools v.2.25 (ref. [99]) to get the final PacBio SV calls. False-positives associated with assembly errors were identified when homozygous no-reference (1/1) SVs were called. For downstream analyses, we masked those regions when we used Char04 primary genome assembly as the reference.

Comparing SVs between Chardonnay and Cabernet Sauvignon. Char04 and Cab08 genomes were compared using three different alignment approaches: whole-genome alignment, long-read alignment and short-read alignment. First, we compared primary contigs of Cab08 (N50 = 2.2 Mb) and Char04. Cab08 primary contigs were aligned to the Char04 reference using nucmer (nucmer -maxmatch -noextend) in MUMmer4 (ref. [31]). After filtering one-to-one alignments with a minimum alignment length of 1,000 bp (delta-filter -1 -l 1000), the show-diff function and NucDiff[100] were used to extract the features and coordinates of SVs.

The second comparison was based on alignment of SMRT reads from Cab08 onto the Char04 reference. SMRT reads from Cab08, representing ~140× coverage, were mapped onto Char04 genome using Minimap2 and NGMLR, as described above. SVs were genotyped on the basis of merged SV calls from both mappers, using the population calling steps of Sniffles pipeline[24]. The SV calls were filtered and duplicates were removed following the four steps listed in the previous section. The genotypes of SV calls from both programmes were intersected using bedtools v.2.25 (ref. [99]) to get the final SMRT-based SV calls. These SMRT-based SV calls were used as the 'gold standard' for downstream analyses.

Finally, we mapped Cab08 Illumina PE reads corresponding to ~15× of raw coverage, which mimics the coverage of population data (see below). These reads were filtered, mapped onto the Char04 reference, and then the bam files were cleaned, sorted with PCR duplicates and masked following ref. [18]. SVs were called with all the population samples (69 in total, see below) using both LUMPY v.0.2.13 (ref. [32]) and DELLY2 v.0.7.7 (ref. [33]). For LUMPY, the read and insert lengths were extracted from mapping files (bam files) for each sample using samtools v.1.9 (ref. [75]) and the SVs were genotyped using SVTyper[32]. The SV calls from DELLY and LUMPY were merged using SURVIVOR v.1.0.3 (ref. [98]). SVs for all 69 population samples presenting the following five criteria were retained: (1) a minimum of three PE reads or split reads (SR) supporting the given SV event across all samples; (2) SV calls with precise breakpoints (flag PRECISE); (3) SVs passing the quality filters suggested by DELLY and LUMPY (flag PASS); (4) SV length ≥ 50 bp; (5) complex SVs consisting of, or overlapping, SVs were excluded. SV calls for Cab08 and Char04 were extracted using vcftools v.0.1.13 (ref. [101]) to permit the comparison of the three detecting methods.

The coordinates and SV features for all SV calls of Cab08 and Char04 based on whole-genome alignment, SMRT reads and Illumina short-read alignments were extracted and saved as bed files. SV calls of the three methods were compared using bedtools v.2.25 (ref. [99]) with a minimum reciprocal overlap of 80%. We took the intersect of the DELLY and LUMPY calls to separate SVs into three categories: (1) shared between methods, which was roughly 74.6% of the SV calls; (2) DELLY-specific SVs; (3) LUMPY-specific SVs. We then combined the three sets using SURVIVOR[98] and intersected it with SMRT-based SV calls to get a shared variant call format (VCF) file. Finally, we extracted mapping and quality statistics from the short-read SV calls that corresponded to the 'gold standard' long-read calls. These statistics were used in the population mappings as cut-offs to filter short-read SV calls (see below).

SNP and SV calling for population samples. Illumina whole-genome resequencing data were gathered from 69 accessions (Supplementary Table 4), each with coverage >10×. The mean mapping depth across accessions was 21.6×. The sample of accessions included 12 wild (ssp. *sylvestris*) samples from the Near East, where grape was domesticated, along with 50 *vinifera* cultivars that represent major lineages. These *sylvestris* samples were carefully filtered for provenance and authenticity[26] and were shown to be distinct from cultivars[18]. The sample also included three *V. flexuosa* and four *Muscadinia rotundifolia* accessions from North America, which were used as outgroups for downstream population genetic analyses.

SNPs and indels were called for this population sample using the HaplotypeCaller in the GATK v.4.0 pipeline, following ref. [18]. SNPs and indels were filtered and annotated using SnpEff v.4.0 (ref. [102]), following ref. [18]. SVs were called from short-read alignment using the LUMPY & DELLY pipelines, as described above. The merged SV genotypes were filtered following the six steps enumerated in the previous section, with the added proviso that SV calls missing in 30% of all individuals were excluded for population genetic analyses. In addition, we used statistics from the intersected set of SVs called from Cab08 to Char04 comparisons to filter 'real' SVs (see previous section). That is, we used statistics from the set of SVs detected by short-read alignment that were confirmed by corresponding to 'gold standard' SV calls by long-read alignment. These cut-off statistics included: (1) a minimum number of supporting four reads in LUMPY calls (flag SU refers to the number of supporting reads, which equals to SP + PE); (2) a minimum number of three SR or PE reads supporting each of the reference and variant alleles in DELLY calls (the flag DR/RR: number of PE/SR reads supporting the reference allele; and the flag DV/RV: number of PE/SR reads supporting the variant allele); (3) a mapping quality ≥ 20 in DELLY calls (flag MAPQ); (4) a genotype quality score ≤ −5 (flag GQ) in DELLY calls. SV calls that did not pass these criteria were treated as missing data.

Mobile element insertions. We used the filtered BAM files with PCR duplicates masked for each sample as input for detecting polymorphic TEs with the Mobile Element Locator Tool (MELT) v.2.1.4 (ref. [103]). MELT uses unaligned and split reads from BWA alignments, a reference genome and consensus TE sequences to identify polymorphic TEs. Because MELT relies on sequence similarity for identifying TEs, we used a Hidden Markov Model method to build consensus sequences for the TE families that represented >4% of the Char04 reference (LINES: L1; LTR retrotransposons: *Copia* and *Gypsy*; and DNA transposons: MuDR and MULE-MuDR; Supplementary Table 2). We preprocessed BAM and TE consensus files with the Preprocess and BuildTransposonZIP utilities of MELT, respectively.

MEIs were detected across the population by using the following four steps from the MELT pipeline: (1) TE variants compared to Char04 genome were detected for each accession individually using IndivAnalysis; (2) all polymorphic TE calls from all samples were merged to detect breakpoints of insertions in the reference genome using GroupAnalysis; (3) the resulting variants file was then used to call genotypes of all insertions for each sample using the Genotype utility; (4) a consensus VCF file was creating after filtering the detected MEIs using the MakeVCF utility. We again used only the first 22 longest scaffolds to represent the reference genome in these analyses because fragmented scaffolds affect the performance of the programme[104]. These four steps were performed for each TE family, separately. To set a threshold of maximum divergence, we used both short- and long-read alignments of Cab08 onto Char04 for calling MEI. Then, the four analysis steps were performed for each TE family, separately, with two different thresholds of maximum divergence, 5% and 10%, between putative polymorphic TEs and the consensus sequence. Comparison of the MEIs detected using short- and long-read alignments showed a higher overlap of MEIs between the two kinds of sequencing when applying a maximum divergence threshold (that is, divergence from an inferred consensus TE) of 5% rather than 10% (58% and 33%, respectively). Accordingly, we used MEI calls based on 5% divergence for downstream analyses after filtering. MEI calls were discarded that did not pass the MELT quality filters, with imprecise breakpoints, that were missing in 30% of the population sample and that were shorter than 50 bp.

Population genetic analyses. Our analyses of the Illumina population data resulted in SV calls for a wide variety of events, including INS, DEL, DUP, INV and TRA. In general, variant calling using short-read alignment allowed us to detect only short insertions (INS, Supplementary Fig. 2) and we therefore excluded INS variants from further analyses. Complex variants, which were defined as composite variants of different types (for example, a reverse tandem duplicate: INVDUP), were also excluded. We also removed any DELLY & LUMPY SV calls in the remaining categories (DEL, DUP, INV and TRA) that overlapped with MEI calls or genomic regions annotated as TEs. Finally, we only retained SV calls that shared the same breakpoints across the population samples. Altogether, we considered five distinct SV categories—DEL, DUP, INV, TRA and MEI—in our population genetic analyses. We also conducted principal component analyses for SNP and SV calls using PLINK v.1.9 (ref. [104]; Supplementary Fig. 6).

SNPs and SVs with a minor allele frequency >0.1 were used for analyses of LD in the wild and the cultivated grapevine samples, respectively. LD decay along physical distance were measured by the squared correlation coefficients (r^2) between all pairs of SNPs in a physical distance of 300 kb, using PLINK v.1.9 (ref. [104]). The decay of LD against physical distance was estimated using nonlinear regression of pairwise r^2 versus the physical distance between SNPs or SVs mid-positions[39].

Since LD decayed in 20 kb in both the wild and the cultivated samples, we divided the Char04 genome into 24,056 non-overlapping windows of 20 kb in size to calculate genomic differentiation of SVs between wild and cultivated samples and to compare SV differentiation to SNPs. For a window to be included in downstream analyses, we required at least 1,000 bases after filtering. Levels

of genetic differentiation between species at each site were estimated using the method-of-moments F_{ST} estimators based on vcftools v.0.1.13 (ref. [101]), which calculates indices of the expected genetic variance between and within species allele frequencies. We then averaged F_{ST} values of all sites in each 20-kb non-overlapping window.

We calculated the unfolded SFS using the *V. flexuosa* and *Muscadinia rotundifolia* samples as outgroup. To derive the SFS, we counted the number of sites at which k of n haplotypes carry the derived variant for SNPs (synonymous: fourfold sites; and non-synonymous: zerofold sites) and SVs (DEL, DUP, INV, TRA and MEI). To exclude direct effects of selection on synonymous sites, we detected selective sweeps based on the composite likelihood ratio test implemented in SweeD v.3.2.1 (ref. [105]). Synonymous sites at genomic windows with top 5% composite likelihood ratio values were excluded in SFS and downstream analysis.

We calculated the SFS for the sample of 12 putatively wild *sylvestris* samples, a down-sampled set of 12 cultivars and the full set of 50 cultivars (Supplementary Fig. 8). To identify a set of 12 cultivars to down sample, we inferred population structure across samples for all wild *sylvestris* and grapevine cultivars using the NGSadmix utility of ANGSD v.0.912 (ref. [106]) based on SNP sites with <20% missing data, a minimal base quality of 20 and a minimal mapping quality of 30. We predefined the number of genetic clusters K from 2 to 8 and the maximum iteration of the expectation maximization algorithm was set to 10,000. On the basis of these population structure results (Supplementary Fig. 7), the down-sampled set of 12 cultivars was chosen to represent major genetic clusters, being sure not to include clonal accessions and also representing accessions with the least missing data (Supplementary Table 4).

Distribution of fitness effects of SVs. We applied the programme DFE-α v.2.15 to estimate the DFE and the proportion of adaptive variants (α) for nSNPs, DELs, DUPs, INVs, TRAs and MEIs[34,35]. In these analyses, we used information from sSNPs as the neutral reference, based on the unfolded SFS. First, we fitted a demographic model to the SFS for neutral sites using maximum likelihood. We chose a two-epoch demographic model that allows a single step change in population size from N_1 to N_2 t_2 generations in the past[34]. We performed many maximum likelihood searches, each with a different starting point and treated the parameter values that produced the highest log-likelihood as the maximum likelihood estimates of the demographic parameters. Next, given the estimated parameters of the demographic model, we inferred the DFE by fitting a γ distribution to the SFS for the selected sites. As above, we carried out multiple searches with different starting values for β and s, where β is the shape parameter of the γ distribution and s is the mean fitness effect of variants. The maximum likelihood estimates of the DFE parameters and the observed divergence at the selected and neutral sites were then used to estimate the proportion of substitutions (α) that have been fixed by positive selection[35]. We obtained 95% confidence intervals (CIs) for the parameter estimates by analysing 100 bootstrap replicates of SFS and divergence datasets that were generated by randomly sampling genes. Following the findings of ref. [107], we used high-quality data from two North American wild *Vitis* species as outgroup to infer the ancestral state of variants. We note, however, that the inference of the ancestral state of SVs is likely to be inaccurate because the genetic divergence between the wild *Vitis* species and Char04 complicated the mapping process. We therefore also used the folded SFS to estimate the DFE and α, using polyDFE v.2.0 (ref. [108]). The results were similar, so we presented the polyDFE results with 95% CIs obtained from the inferred discretized DFEs from 100 bootstrap datasets.

SVs and sex determination. F_{ST} values for both SNPs and SVs showed clear outlier peaks in the SD region (Fig. 3). The SNPs of the SD region were phased and imputed based on a genetic map[108] using Shapeit v.2.12 (ref. [109]), following ref. [18]. To examine relationships among different sex haplotypes, we built maximum likelihood trees from SNPs in the region. Maximum likelihood trees were based on 10,000 bootstrap replicates, as implemented in ref. [110]. We built trees for the two regions, corresponding to the peaks of SNP divergence[18]. We reasoned that the true SD region should cluster by sex, which was true for the first peak of the SD region but not the second (Supplementary Fig. 10). We therefore concluded that the first peak, defined as the region between 4.90 Mb and 5.04 Mb on chromosome 2 of the PN40024 assembly, represents the SD region. BEAST v.1.8.0 (ref. [111]) was applied to calculate genetic divergence on the basis of a tree with a relaxed molecular clock. After a burn-in of 100,000 steps, data were collected once every 1,000 steps from 10×10^6 Markov chain Monte–Carlo cycles. The divergence time between haplotypes was based on a genome-wide divergence time of 46.9×10^6 years ago between *M. rotundifolia* and *Vitis* species[112].

The boundaries of the SD region were determined by mapping the coding sequences of the chr02:4840000–4980000 region from PN40024 12× v.2 (ref. [43]) to the Char04 and Cab08 references. For both Chardonnay and Cabernet Sauvignon haplotypes, gene models were refined by mapping all the coding sequences identified in the four haplotypes onto Char04 and Cab08 genome assemblies, separately, using GMAP v.2015-11-20 with default parameters[113].

We analysed gene expression data from the three grape flower sexes. Raw sequencing data were obtained from the short-read archive (SRP041212). Reads were first trimmed using Trimmomatic v.0.36 (ref. [114]) with the options: LEADING:3 TRAILING:3 SLIDINGWINDOW:10:20 MINLEN:20. High-quality reads were mapped onto the primary and haplotig genome assemblies of Char04 and Cab08 (ref. [17]) separately, using HISAT2 v.2.0.5 (ref. [115]) with the following options: –end-to-end–sensitive–no-unal. The Bioconductor package GenomicAlignments v.1.12.1 (ref. [116]) was used to extract counts of uniquely mapped reads ($Q > 20$). Mapped reads were then normalized by millions of mapped reads per library (RPM). Differential expression analysis across flower sexes (male versus female, male versus hermaphrodite, female versus hermaphrodite) was performed using the Bioconductor package DESeq2 v.1.16.1 (ref. [117]) using samples of the last two flower growth stages as replicates to allow enough statistical power. *P* values were adjusted using the method of Benjamini and Hochberg[118]. These same data were analysed previously using the same methods, based on mapping to the PN40024 reference[18]. The previous work found a tendency toward female-biased expression of genes in the sex region. However, in the current analyses, the genes that differ in expression in the SD tend to show male-biased expression. The differences between studies reflect mapping biases between the presumed F haplotype in the PN40024 (ref. [42]) and the H haplotype in the Char04 reference. For these reasons, we consider the gene expression analyses to be a tool to help identify interesting candidate loci but caution that additional studies of sex-biased expression are merited.

SVs and berry colour. We compared genomes of two cultivars with dark berries (PN40024 and Cab08) with two cultivars with white berries (Char04 and Sultanina) using pairwise whole-genome alignments and called SVs using the MUMmer4 pipeline. Dot plots were generated using mumplot (mumplot -l 100 -c 1000 -d 10 -banded -D 5) for chromosome 2 where the berry colour QTL is located. For Char04 and Cab08, we verified the SV calls using the Sniffles pipeline[24] after mapping SMRT reads onto the PN40024 reference genome using both the Minimap2 (ref. [97]) and NGMLR[24]. We also zoomed in on this region for SV calls for the population samples to investigate the potential association of SVs, gene expression and the berry colour in different cultivars.

To identify whether other white berry accessions housed large inversions that include the berry-colour genes, we determined the orientation of the rearranged chromosome fragments and putative breakpoints from bam files of discordant PE reads and SR after mapping short reads to the PN40024 genome v.2.0 (ref. [43]). Reads were mapped using the BWA-MEM algorithm in bwa-0.7.12 (ref. [96]). The discordant reads and split reads were extracted using samtools v.1.9 (ref. [75]) and LUMPY v.0.2.13 (ref. [32]). To select breakpoints distinguishing genomes of dark and white berry cultivars, the discordant, the splitter and the original bam files were inspected visually using IGV v.2.2 (ref. [119]).

To detect potentially hemizygous regions on chromosome 2, we calculated runs of homozygosity for each sample using the software PLINK v.1.9 (ref. [104]) with the following options: –homozyg-window-het 0–homozyg-snp 41–homozyg-window-snp 41–homozyg-window-missing 0–homozyg-window-threshold 0.05–homozyg-kb 500–homozyg-density 5000–homozyg-gap 1000. Copy number variation analyses were conducted in cnv-seq[120] using the neighbouring grapevines with white and dark berry colours with bam file of the former as a test and bam file of the latter as a reference. The $\log_2$ values of the adjusted copy number ratio were plotted in *R*.

Gene expression analyses of the berry colour region used the raw RNA-seq data from SRA: SRP049306-SRP049307 (ref. [64]). The data were generated from berries sampled during berry development at four stages, including two before and two after veraison (onset of ripening), from ten Italian varieties (five dark and five white). RNA-seq data were mapped onto the Char04 reference and analysed as described in the previous section. Differential gene expression analysis was performed for each berry growth stage, separately, by comparing samples from dark cultivars with berries from white varieties. Genes presenting an adjusted *P* value ≤ 0.05 between dark and white cultivars were considered as significantly expressed. Gene expression analyses focused on the 173 genes in the Char04 chromosome 2 inversion.

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

Data availability

Raw SMRT reads for were deposited to the SRA at the NCBI under the BioProject ID PRJNA550461. Genome assembly and annotation of genes and transposable elements are available at https://zenodo.org/record/3337377#.XS0i9ZOpG_M. VCFs and custom scripts are available on request.

Received: 13 May 2019; Accepted: 26 July 2019;
Published online: 9 September 2019

References

1. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
2. Goff, S. A. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100 (2002).

3. Yu, J. A Draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
4. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
5. Daccord, N. et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
6. Raymond, O. et al. The *Rosa* genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
7. Roessler, K. et al. The genomics of selfing in maize (*Zea mays* ssp. *mays*): catching purging in the act. *Nat. Plants* https://doi.org/10.1038/s41477-019-0508-7 (2019).
8. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
9. 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).
10. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675 (2019).
11. Fuentes, R. R. et al. Structural variants in 3000 rice genomes. *Genome Res.* **29**, 870–880 (2019).
12. Sun, S. et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* **50**, 1289–1295 (2018).
13. Miller, A. J. & Gross, B. L. From forest to field: perennial fruit crop domestication. *Am. J. Bot.* **98**, 1389–1414 (2011).
14. *Report on the World Vitivinicultural Situation* (The International Organisation of Vine and Wine, 2016); http://www.oiv.int/public/medias/4906/press-release-2016-bilan-en.pdf
15. Migicovsky, Z. et al. Patterns of genomic and phenomic diversity in wine and table grapes. *Hortic. Res.* **4**, 17035 (2017).
16. McGovern, P. et al. Early neolithic wine of Georgia in the South Caucasus. *Proc. Natl Acad. Sci. USA* **114**, E10309–E10318 (2017).
17. This, P., Lacombe, T. & Thomas, M. R. Historical origins and genetic diversity of wine grapes. *Trends Genet.* **22**, 511–519 (2006).
18. 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).
19. Velasco, R. et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).
20. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
21. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
22. Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A. & Cantu, D. Diploid genome assembly of the wine grape Carménère. *G3* **9**, 1331–1337 (2019).
23. Roach, M. J. et al. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLoS Genet.* **14**, e1007807 (2018).
24. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
25. Bowers, J. et al. Historical genetics: the parentage of Chardonnay, Gamay, and other wine grapes of northeastern France. *Science* **285**, 1562–1565 (1999).
26. Myles, S. et al. Genetic structure and domestication history of the grape. *Proc. Natl Acad. Sci. USA* **108**, 3530–3535 (2011).
27. 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).
28. Beridze, T. et al. Plastid DNA sequence diversity in a worldwide set of grapevine cultivars (*Vitis vinifera* L. subsp. *vinifera*). *Bull. Georgian Nat. Acad. Sci.* **5**, 91–96 (2011).
29. Minio, A., Lin, J., Gaut, B. S. & Cantu, D. How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front. Plant Sci.* **8**, 826 (2017).
30. Silva, C. D. et al. The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome. *Plant Cell* **25**, 4777–4788 (2013).
31. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
32. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
33. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
34. Keightley, P. D. & Eyre-Walker, A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251–2261 (2007).
35. Eyre-Walker, A. & Keightley, P. D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26**, 2097–2108 (2009).
36. Lin, Y.-C. et al. Functional and evolutionary genomic inferences in populus through genome and population sequencing of American and European aspen. *Proc. Natl Acad. Sci. USA* **115**, E10970–E10978 (2018).
37. Ramu, P. et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
38. Henn, B. M. et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc. Natl Acad. Sci. USA* **113**, E440–E449 (2016).
39. Hill, W. G. & Robertson, A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).
40. Parage, C. et al. Structural, functional, and evolutionary analysis of the unusually large stilbene synthase gene family in grapevine. *Plant Physiol.* **160**, 1407–1419 (2012).
41. Fechter, I. et al. Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. *Mol. Genet. Genom.* **287**, 247–259 (2012).
42. Picq, S. et al. A small XY chromosomal region explains sex determination in wild dioecious *V. vinifera* and the reversal to hermaphroditism in domesticated grapevines. *BMC Plant Biol.* **14**, 229 (2014).
43. Canaguier, A. et al. A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). *Genom. Data* **14**, 56–62 (2017).
44. Coito, J. L. et al. VviAPRT3 and VviFSEX: two genes involved in sex specification able to distinguish different flower types in *Vitis*. *Front. Plant Sci.* **8**, 98 (2017).
45. Dobritsa, A. A. & Coerper, D. The novel plant protein INAPERTURATE POLLEN1 marks distinct cellular domains and controls formation of apertures in the *Arabidopsis* pollen exine. *Plant Cell* **24**, 4452–4464 (2012).
46. VanBuren, R. et al. Origin and domestication of papaya Yh chromosome. *Genome Res.* **25**, 524–533 (2015).
47. Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. Retrotransposon-induced mutations in grape skin color. *Science* **304**, 982 (2004).
48. Walker, A. R. et al. White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J.* **49**, 772–785 (2007).
49. Fournier-Level, A. et al. Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. ssp. *sativa*) berry: a quantitative trait locus to quantitative trait nucleotide integrated study. *Genetics* **183**, 1127–1139 (2009).
50. 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).
51. Yakushiji, H. et al. A skin color mutation of grapevine, from black-skinned Pinot Noir to white-skinned Pinot Blanc, is caused by deletion of the functional VvmybA1 allele. *Biosci. Biotechnol. Biochem.* **70**, 1506–1508 (2006).
52. Carbonell-Bejerano, P. et al. Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiol.* **175**, 786–801 (2017).
53. Springer, N. M. et al. The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nat. Genet.* **50**, 1282–1288 (2018).
54. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
55. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
56. Ramos-Madrigal, J. et al. Palaeogenomic insights into the origins of French grapevine diversity. *Nat. Plants* **5**, 595–603 (2019).
57. Liu, Q., Zhou, Y., Morrell, P. L. & Gaut, B. S. Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.* **34**, 908–924 (2017).
58. Renaut, S. & Rieseberg, L. H. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Mol. Biol. Evol.* **32**, 2273–2283 (2015).
59. Wang, L. et al. The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* **18**, 215 (2017).

60. Flagel, L. E., Willis, J. H. & Vision, T. J. The standing pool of genomic structural variation in a natural population of *Mimulus guttatus*. *Genome Biol. Evol.* **6**, 53–64 (2014).
61. Uzunović, J., Josephs, E. B., Stinchcombe, J. R. & Wright, S. I. Transposable elements are important contributors to standing variation in gene expression in *Capsella grandiflora*. *Mol. Biol. Evol.* **36**, 1734–1745 (2019).
62. Liang, Z. et al. Whole-genome resequencing of 472 *Vitis* accessions for grapevine diversity and demographic history analyses. *Nature Commun.* **10**, 1190 (2019).
63. Laucou, V. et al. Extended diversity analysis of cultivated grapevine *Vitis vinifera* with 10K genome-wide SNPs. *PLoS ONE* **13**, e0192540 (2018).
64. Massonnet, M. et al. Ripening transcriptomic program in red and white grapevine varieties correlates with berry skin anthocyanin accumulation. *Plant Physiol.* **174**, 2376–2396 (2017).
65. Xie, K. T. et al. DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* **363**, 81–84 (2019).
66. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
67. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. S. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6**, 31900 (2016).
68. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
69. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **44**, e147 (2016).
70. Solares, E. A. et al. Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/267401v2 (2018).
71. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
72. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
73. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
74. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
75. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
76. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
77. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
78. Kent, W. J. BLAT—the BLAST-Like alignment tool. *Genome Res.* **12**, 656–664 (2002).
79. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
80. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
81. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211 (2014).
82. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
83. Hancock, J. M. & Zvelebil, M. J. *Dictionary of Bioinformatics and Computational Biology* (John Wiley & Sons, Ltd., 2004).
84. Minio, A. et al. Iso-Seq allows genome-independent transcriptome profiling of grape berry development. *G3* **9**, g3.201008.2018 (2019).
85. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
86. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
87. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
88. Haas, B. J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
89. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
90. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
91. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
92. Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
93. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
94. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
95. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
96. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
97. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
98. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
99. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–11.12.34 (2014).
100. Khelik, K., Lagesen, K., Sandve, G. K., Rognes, T. & Nederbragt, A. J. NucDiff: in-depth characterization and annotation of differences between two sets of DNA sequences. *BMC Bioinformatics* **18**, 338 (2017).
101. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
102. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
103. Gardner, E. J. et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
104. 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).
105. Pavlidis, P., Živković, D., Stamatakis, A. & Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol. Biol. Evol.* **30**, 2224–2234 (2013).
106. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
107. Keightley, P. D., Campos, J., Booker, T. & Charlesworth, B. Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of Drosophila melanogaster. *Genetics* **203**, 975–984 (2016).
108. Hyma, K. E. et al. Heterozygous mapping strategy (HetMappS) for high resolution genotyping-by-sequencing markers: a case study in grapevine. *PLoS ONE* **10**, e0134880 (2015).
109. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
110. 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).
111. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
112. Ma, Z.-Y. et al. Phylogenomics, biogeography, and adaptive radiation of grapes. *Mol. Phylogenet. Evol.* **129**, 258–267 (2018).
113. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
114. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
115. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
116. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
117. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/002832v3 (2014).
118. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).
119. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics* **14**, 178–192 (2013).

120. Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).

Acknowledgements

We are grateful for the technical assistance of R. Gaut and R. Figueroa-Balderas, the services of the High Performance Computing Cluster and the Genomics High Throughput Facility at UC Irvine, and the comments of A. Muyle, D. Seymour, D. Koenig, T. Batarseh, G. Martin, P. Morrell and J. Ross-Ibarra. This work was supported by seed funding from UC Irvine, NSF grant no. 1542703 to B.S.G., NSF grant no. 1741627 to B.S.G. and D.C. and support to D.C. by J. Lohr Vineyards and Wines, E. & J. Gallo Winery and the Louis P. Martini Endowment in Viticulture.

Author contributions

Y.Z., D.C. and B.S.G. designed the research. Y.Z., D.C. and B.S.G. wrote the manuscript. Y.Z., A.M., M.M., E.S. and Y.L. performed the analyses. T.B. provided data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-019-0507-8.

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

Correspondence and requests for materials should be addressed to D.C. or B.S.G.

Peer review information: *Nature Plants* thanks Briana Gross and other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

nature research

Corresponding author(s): Brandon Gaut, Dario Cantu

Last updated by author(s): July 16, 2019

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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	No software had been used for data collection
Data analysis	FALCON-Unzip v1.7.7; GATK v4.0; SnpEff v4.0; Canu v1.5; DBG2OLC; PBDAGCON; BLASR; quickmerge; Pilon v1.16; MUMmer4 v3.23, v4.0; BUSCO v3; Blat v. 36; HiRise; SSPACE-longreads v1.1; PBjelly; RepeatMasker; Augustus v3.0.3; GeneMark-ES v4.32; CD-HIT-EST v4.6; Blast2GO v4; bwa-0.7.12; NGMLR v0.2.7; SURVIVOR v1.0.3; edtools v2.25; Minimap2; Sniffles; NucDiff; LUMPY v0.2.13; ELLY2 v0.7.7; SVTyper; vcftools v0.1.13; MELTv2.1.4; PLINK v1.9; SweeD v3.2.1; ANGSD v0.912; DFE-α v2.15; polyDFE v2.0; BEAST v1.8.0; MEGAX; Shapeit v2.12; GMAP v.2015-11-20; Trimmomatic v.0.36; HISAT2 v.2.0.5; GenomicAlignments v.1.12.1; DESeq2 v1.16.1; IGV v2.2; samtools v1.9;

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Raw reads were deposited to the Short Read Archive (SRA) at the NCBI under the BioProject ID: PRJNA550461. VCF files for SNPs and SVs are available from the authors on request.

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

Ecological, evolutionary & environmental sciences study design

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

Study description	We sequenced and assembled the 605Mb genome of the Chardonnay grape (Vitis vinifera ssp. sativa), which we used to catalog structural variants within and between genomes. We then extended our work to a sample of 69 cultivars and wild accessions to conduct population and evolutionary genomic studies.
Research sample	We chose to sequence Chardonnay (FBS04) because it is a common clone that is homozygous for the hermaphrodite haplotype (HH). The other genome samples were based on availability in public repositories at the time of analysis. Representative grapevine cultivars were taken from public databases. Wild samples (Vitis vinifera ssp. sylvestris) were selected to be from the Near East, where grapevine had been domesticated, and for their provenance in previous publications (Zhou et al. PNAS 2017, Myles et al. PNAS 2011)
Sampling strategy	The wild samples were sampled from several countries in the Near East where grapevine had been domesticated and the cultivars had been sampled to cover most of common grapevine varieties across the world.
Data collection	SMRT long reads and Illumina short reads and Hi-C data were sequenced for Chardonnay 04 in the UCI sequence center. Illumina short reads for other wild or cultivated samples were collected from NCBI SRA database.
Timing and spatial scale	Chardonnay was sampled in 2014. World-wild cultivated grapevine samples and wild samples from the Near East came from public sources, most of which were generated in the last two to four years.
Data exclusions	No data were excluded.
Reproducibility	All attempts to repeat the analyses were successful.
Randomization	Not applicable.
Blinding	The generation of sequencing data was blind. Sequence analysis was performed as if blinded - e.g., wild and domesticated germplasm were evaluated by the same pipelines.

Did the study involve field work? ☐ Yes ☒ No

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging