Structural variants, clonal propagation, and genome evolution in grapevine (*Vitis vinifera*)

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

[1] Dept. Ecology and Evolutionary Biology, UC Irvine, Irvine, CA 92617

[2] Dept. Viticulture and Enology, UC Davis, Davis, CA 95616

[3] Institute of Molecular Genetics, Agricultural University of Georgia, Tbilisi, Georgia

* Co-corresponding authors:

Dario Cantu	Brandon Gaut
595 Hilgard Lane	321 Steinhaus Hall
Dept. Viticulture and Enology	Dept. Ecology and Evolutionary Biology
UC Davis	UC Irvine
Davis, CA 95616	Irvine, CA 92617
dacantu@ucdavis.edu	bgaut@uci.edu

ABSTRACT:

Little is known about the type and numbers of structural variants (SVs) in plant genomes, the evolutionary processes that shape SV variation, or their effects on phenotypes. Here we assembled the 605Mb genome of the Chardonnay grape (*Vitis vinifera* ssp. *sativa*), which we use to catalog SVs within and between genomes and also across 69 cultivars and wild accessions. SVs underlie widespread hemizygosity of protein coding genes within Chardonnay and genic presence-absence variation between cultivars. Negative selection acts against all SV types, but they also accrue as heterozygous, recessive mutations in clonally propagated grapevines. SV hotspots in the sex determination region and the berry color locus illustrate their impact on phenotype and the role of SVs as drivers of convergent phenotypic evolution.

INTRODUCTION:

Most plant reference genomes have been produced from homozygous source materials. In some cases -- such as selfing *Arabidopsis thaliana* -- homozygosity is the natural form. In other species, such as maize (Sun et al., 2018), apple (Daccord et al., 2017) and roses (Raymond et al., 2018), homozygosity is either based on haploid tissue or a manipulated state. The focus on homozygous materials is technically convenient, but it has at least two important biological limitations. First, 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 of knowledge is critical, because some studies suggest that SVs explain as much or more phenotypic variation than SNPs (Chia et al., 2012, Yao et al., 2015) and because SVs can play an important role in adaptation (Fang et al., 2012). For example, SVs are the causative genetic variant for at least one-third of known domestication alleles (Gaut et al., 2018). Second, the focus on homozygous genomes has restricted insights into the biology of clonally propagated crops, which exist in a state permanent heterozygosity and accrue somatic mutations over time (McKey et al., 2010).

These observations argue that the evolution and function of plant genomes cannot be understood without an explicit focus on the number, types, and population dynamics of SVs. Here we study the evolutionary history and potential phenotypic effects of SVs in domesticated grapevine (*Vitis vinifera* ssp *sativa;* hereafter *'sativa'*). Grapevines are clonally propagated and arguably the most important horticultural crop in the world (Myles et al., 2011), with ~76 million tons of fruit harvested globally in 2015 (OIV, 2015, Migicovsky et al., 2017). The products of grape cultivation – which include table grapes, raisins, juice, wine and oil - contribute an estimated $220B annually to the American economy alone (2018). Grapevines were domesticated from their wild ancestor, the wild Eurasian grapevine (*Vitis vinifera* ssp. *sylvestris;* hereafter *'sylvestris'*), nearly ~8,000 years ago in the Transcaucasus (McGovern et al., 2017). Domestication

facilitated higher sugar content in the berry, increased berry and bunch size, altered seed morphology, and prompted a shift from a dioecious to an hermaphroditic mating system (This et al., 2006). The shift in mating system was especially dramatic, because all extant wild *Vitis* species have separate male and female plants, indicating that dioecy has been maintained since the origin of the genus > 35Mya (Liu et al., 2016, Ma et al., 2018).

In theory, hermaphroditic grape cultivars can be selfed; in practice, selfed progeny are often non-viable, perhaps because inbreeding exposes deleterious alleles hidden in the heterozygous state (Zhou et al., 2017). Consequently, most grape cultivars represent crosses between distantly related parents, resulting in high levels of heterozygosity within cultivars (Jaillon et al., 2007, Velasco et al., 2007, Minio et al., 2017). Many of these highly heterozygous cultivars have been maintained by clonal propagation for hundreds and even thousands of years (This et al., 2006), providing an opportunity for the accumulation of somatic mutations that further elevate heterozygosity. At present, however, we have only a nascent understanding of the rate and pattern of somatic mutations in any plant taxon (Schmid-Siegert et al., 2017, Roach et al., 2018), and still fewer insights into the population dynamics of SVs within crops (Gaut et al., 2018).

RESULTS

Rampant Hemizygosity within Grapevines: 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 employed a hybrid sequencing approach, based on sequence data of 58X coverage of Single Molecule Real Time (SMRT, Pacific Biosciences) long reads and 162X short-read coverage (Online.). Hybrid assembly resulted in a contig N50 of 1.24Mb (Online.); application of Hi-C improved the scaffold assembly N50 to 24.5Mb, vastly extending contiguity relative to other grape genomes (Jaillon et al., 2007, Velasco et al., 2007, Chin et al., 2016, Roach et al., 2018) (**Table 1**). The resulting primary assembly was 605Mb in length, a value 20% higher than the partially inbred Pinot Noir PN40024 reference (Jaillon et al., 2007) but similar to the 590Mb assembly of Cabernet Sauvignon (Cab08) (Chin et al., 2016). The Char04 primary assembly had a BUSCO score of 93.4%, contained 38,020 annotated protein-coding genes, and consisted of

47.3% transposable elements (TEs), particularly from the *gypsy* and *copia* superfamilies (**Tables 1 & S1**).

We identified heterozygous SVs (hSVs) within Char04 by remapping SMRT reads to the Char04 reference (Sedlazeck et al., 2018b), revealing 18,998 hSVs of length > 50 bp (**Figure 1A** & **Table S2**). Only 0.3% of the hSVs were detected as homozygous (**Table S2**), suggesting a low rate of misassembly. After masking these regions, observed hSVs were as long as 5.3 Mb and together constituted 91.21 Mb, or 15.1%, of the 605Mb primary assembly. 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 (**Table S2**), respectively. In addition to SVs >50 bp in length, we also detected 119,067 small (< 50bp) indels and 873,159 SNPs. After including small indels, we estimate that the two Char04 chromosomal sets may differ by as much as 15.3% in length, with 7.4% of this caused by TEs that are polymorphic between chromosomes.

Surprisingly, we found 5,546 hemizygous genes in Char04 based on inferences from long-read-mapping (**Figure 1B**), representing 14.6% of all annotated protein-coding genes. This value is consistent with the overall proportion of chromosomal heterozygosity by length, but it also raises concerns that it could be artificially high 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 within this cultivar. All of these analyses are consistent in indicating that hemizygosity affects ~1 in 7 genes in grapevine cultivars.

SVs between grapevine genomes: The Char04 and Cab08 genome assemblies permitted a rare opportunity to compare highly contiguous genomes from within a single cultivated species (Sun et al., 2018). We detected SVs between genomes using three approaches.

We first mapped SMRT reads from Cab08 to the Char04 primary assembly (**Figure S1**). These results yielded ~3-fold higher numbers of SV events between cultivars than within Char04 (**Table S2**), reflecting the distinct parentage of Chardonnay and Cabernet Sauvignon (Bowers et al., 1999, Arroyo-García et al., 2006, Beridze et al., 2011, Myles et al., 2011). Of 59,913 inferred SVs, DEL and MEI events were again 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 exhibited complete presence/absence variation (PAV), a number similar to previous estimates based on less complete data (Da Silva et al., 2013, Minio et al., 2017). Based on GO analyses, PAV genes are biased toward functions in defense response, flower development, membrane components and transcription factors ($P < 0.001$).

We also compared Char04 and Cab08 primary assemblies by whole genome alignment (Marçais et al., 2018) (**Figure S2**), which yielded a similar numbers of SVs (52,952) but fewer MEI events (**Table S2**). Finally, we mapped 25x Illumina reads from Cab08 to Char04, which detected only 62% of the number of SVs based on SMRT reads (**Table S2**). The length distribution of SVs varied among the three methods; SMRT-read analyses detected larger (>10kb) events (**Figure S3**). Importantly, 75% of SVs inferred by SMRT-read alignment were confirmed either by genome alignment or by short-read alignment analyses (**Figure 1D; Figure S4**). These confirmed SVs still encompassed 1,822 PAV genes and 45,403 MEIs between Char04 and Cab08 and continue to attest to remarkable SV variation among grapevine cultivars.

Selection pressures on SVs: There have been few population genetic analyses of SVs in wild (Flagel et al., 2014) or domesticated plants (Gaut et al., 2018). 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 >10X (**Table S3**). The application of short-read alignment for detecting SVs is subject to high levels of false-negatives and false-positives (Sedlazeck et al., 2018a). To limit false-positives, we relied on our Char04 to Cab08 comparisons,

specifically the subset of SVs called by both long-read and short-read alignments (**Figure 1D**). We examined the mapping quality, mapping depth and likelihood of these calls to provide empirical cut-offs for short-read SV calls (Online.). After applying these 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 (**Table S4**). These SVs yielded relationships among accessions that were remarkably similar to those based on SNPs, providing reassurance about their reliability (**Figure S5**).

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 (**Figures S6-S8**). The SFS for the two taxa were similar overall (**Figure 2A**), reflecting the fact that cultivated grapevine did not undergo a severe domestication bottleneck (Myles et al., 2011, Zhou et al., 2017) that can dramatically alters population frequencies. In both taxa, all SV types exhibited leftward shifts of the SFS relative to synonymous SNPs (sSNPs), and their SFS differed significantly from that of sSNPs in both taxa ($p < 0.05$, Kolmogorov-Smirnov, Bonferroni corrected). These leftward shifts suggest 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, using sSNPs as a neutral control. In both taxa, the results confirmed that non-synonymous SNPs (nSNPs) and SVs undergo strong purifying selection (**Figure 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 (α=25% for *sylvestris*), DEL (α=21%) and MEI (α=20%) events (**Figure 2C**). 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 (Lin et al., 2018). Based on DFE and α estimates, negative selection appears to be stronger in *sativa* than *sylvestris* (**Figure 2**). However, the comparison between taxa must be interpreted with caution because the inferential models

were designed to analyze outcrossing species like *sylvestris* and not clonally propagated crops. Nonetheless, the results strongly suggest that SV events are more deleterious than nSNPs, on average, and that INV and TRA events are especially deleterious.

SVs and clonal propagation: SVs are deleterious, on average, but clonal propagation may allow variants to hide as heterozygous recessives (Ramu et al., 2017, Zhou et al., 2017). The accumulation of recessive mutations was evident from three aspects of *sativa* genetic diversity. First, heterozygosity based on all SNPs was 11% higher, on average, within *sativa* than *sylvestris* (**Figure S9**). Second, sheltering of recessive mutations was evident from calculations of the additive SV load, which is the number of number of heterozygous mutations plus twice the number of derived homozygous mutations per individual (Henn et al., 2016). Individual cultivars have a 6% higher additive SV load than their wild counterparts, on average, due to elevated heterozygosity (**Figure 3A**). Enhanced load was not evident for homozygous SVs or for presumably neutral sSNPs (**Figure 3A**), suggesting that deleterious SVs accrue and are sheltered in the heterozygous state. These patterns of SV load are consistent with forward simulations showing that clonal propagation can lead to the accumulation of deleterious recessive mutations without a notable fitness decrease (Zhou et al., 2017). Finally, the SFS provided evidence of sheltering of recessive mutations within *sativa*, based on the marked reduction in frequency for any variants over 50% (**Figures 2A&S8**). This unexpected observation has 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 the 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 (Hill and Robertson, 1968) and because cultivated grapes tend to have more low frequency variants than their wild counterparts **(Figure 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. This last finding is important because SVs have been implicated to

affect phenotypes and explain more phenotypic variation than SNPs in some systems (Chia et al., 2012, Yao et al., 2015). The rapid decline of LD suggests that it may be difficult to identify causative SVs by relying on linkage to SNPs.

SVs and domestication: Cultivated grapevine differs phenotypically from its wild relatives (This et al., 2006). 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 F_{ST} in fixed windows of 20 kb (**Figure 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 (**Figure 2A**).

We ranked the top 1% (or 485) F_{ST} windows for both SNPs and SVs. SNP-based windows based generally conformed to a previous study (Zhou et al., 2017), but SNPs and SVs both identified QTL regions on chromosome 2 that corresponded to the sex-determination region and to the berry color locus (**Figure 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 GO 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 (Parage et al., 2012), 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 10 *sativa* genes, respectively (**Table S5**). Among the 10 lost, four were NBS-LRR disease resistance genes located between 11.053 to 11.064 Mb on chromosome 9 of Char04.

SVs and the sex determination (SD) region: The highest F_{ST} peak for SVs corresponded to the sex determination region on chromosome 2 (**Figure 3C**). This region also

contained more SV events relative to the genomic background ($p < 0.0067$; χ^2). In theory, mutations within this region caused a shift in mating system from dioecy in the wild to hermaphroditism under cultivation. These mutations are of substantial practical and theoretical interest (Muyle et al., 2017), but causative genes have been identified rarely (Wang et al., 2012, Boualem et al., 2015, Harkess et al., 2017, Muyle et al., 2017).

In *V. vinifera*, the sex-determining (SD) region maps to chromosome 2, but the boundaries of the region have been unclear. One study mapped the SD region to a 152kb region between chromosomal position 4.90Mb and 5.04Mb on PN40024 (Fechter et al., 2012, Picq et al., 2014), but other studies have implicated a larger region extending from 4.90Mb to 5.33Mb (Hyma et al., 2015, Zhou et al., 2017), including two distinct peaks of SNP divergence (Zhou et al., 2017). We tested linkage between gender and each peak by building a separate phylogeny for each region. Only the first region clustered with the known mating genotypes of cultivars (**Figure S10)**, confirming that the narrower region segregates with sex type. We therefore defined the SD region as corresponding to between 4.90Mb and 5.04Mb on chromosome 2 of Char04.

Chardonnay is rare among cultivars because it is a homozygote for the hermaphroditic (H) haplotype (Picq et al., 2014). As a result, we were able to resolve and annotate two *H* haplotypes from the SD region. In contrast, Pinot Noir and Cabernet Sauvignon are heterozygous for the *H* and female (*F*) haplotype. Comparing the *H* haplotype on the Char04 reference to the PN40024 primary assembly (Canaguier et al., 2017), which is thought to represent the *F* haplotype (Picq et al., 2014), we identified PAV for four genes. One of these, *VviAPT3*, has been proposed as a candidate SD gene (Fechter et al., 2012), because it may have a role in the abortion of pistil structures (Coito et al., 2017). Both the *H* and *F* haplotypes of Cab08 contain *VviAPT3* **(Figure 4A)**, suggesting that the lack of *VviAPT3* on PN40024 was an assembly error, which is not surprising given the prevalence of SVs in the region.

The remaining three genes that differed between Char04 and PN40024 appear to differentiate *H* from *F* haplotypes, because they were found on both Char04 *H* haplotypes

and differed between Cab08 haplotypes (**Figure 4A**). Unfortunately, detailed functions were not available for these three genes, but all three were lowly expressed, based on expression data from two late stages of floral development (Online.) (**Figure 4B**). We also identified a previously unrecognized gene, *Inaperturate pollen 1* (*VviINP1*), in both *F* and *H* haplotypes. *INP1* expression in *Arabidopsis* alters the deposition of apertures, which are thought to facilitate the exit of pollen tubes during germination (Dobritsa and Coerper, 2012). Hence, *INP1* and could be involved in pollen sterility in females, but this gene was again lowly expressed in our data (**Figure 4B**). Finally, we also identified and annotated a previously unrecognized *C2H2*-type Zinc finger gene. Altogether, our resolution and annotation of *H* and *F* haplotypes in Char04 and Cab08 reveal several genes that may function in SD.

It is thought that dioecy (like that found in wild *Vitis* species) originated via a two-step process (Oberle, 1938, Charlesworth and Charlesworth, 1978). The first step was the evolution of a recessive male sterility mutation that drives the origin of females. The second was the formation of a closely-linked, dominant female-sterility mutation, which causes males (*M*). When cultivated grapes reverted to hermaphroditism, it was likely to be caused by a mutation in the dominant *F* sterility gene on the *M* haplotype (VanBuren et al., 2015). In grapevine, this model is consistent with two observations: *i*) both the *H* and *F* haplotypes, but not the dominant *M* haplotype, are found within cultivated varieties and *ii*) *H* haplotypes are more similar to *M* than *F* haplotypes (Picq et al., 2014).

The female sterility gene has yet to be identified, but it is likely to be a gene expressed in males and knocked-down in hermaphrodites. To identify such genes, we performed gene expression analyses among sexes after mapping of RNAseq reads to the Char04 reference (Online.). Five genes differed significantly (adj. $p \leq 0.05$) in sex-specific expression. Four were more highly expressed in males, including *VviAPT3* and the *C2H2*-type Zinc finger gene; these four constitute the most plausible female sterility candidates.

Reasoning that the true female sterility gene must have a knock-out mutation in *H* relative to *M* haplotypes, we built a phylogeny of the SD region to investigate the

relationship of the *M* haplotype from our single, confirmed *sylvestris* male to inferred *H* and *F* haplotypes (**Figure 4C**). The *M* haplotype separated two clades of *H* haplotypes, thus providing further support for more than one origin of hermaphroditism (Picq et al., 2014). Based on SNP variants among haplotypes and using a previously determined rate of SNP divergence (Ma et al., 2018), we estimated that one of the two *H* clades dated to 10,705 years before present (ybp), with a 95% confidence interval ranging from 8,468 to 12,805 ybp. The second clade appeared to be older, at 13, 222 ybp (95% CI: 10,459 to 15,816 ybp). These dates are older than the accepted ~8,000 ybp date of domestication (McGovern et al., 2017) but represent an upper bound that is likely to decrease with the inclusion of more *sylvestris* males.

Another prominent feature of the phylogeny is that the single *sylvestris M* haplotype was closely related to one Char04 *H* (**Figure 4C**). We compared SVs between these two haplotypes in the hope of identifying candidates for the causative *M->H* mutation. Several genes were in a hemizygous state in the wild male, including *PAP2*, the *DEAD DEAH* box RNA helicase gene, the TPR-containing protein and the unknown protein previously known as *ETO1* (**Figure 4A**). There were also three hemizygous TEs near genes, but unfortunately none of these MEIs were obvious candidates to affect the function of the four female sterility candidates identified via expression analyses (**Figure 4B**). Hence, both the genetic mutation that causes hermaphroditism and the identity of the dominant female sterility gene remain elusive. Nonetheless, our resolution and annotation of three *H* and one *F* haplotype narrows the list of candidate genes and underscores the dynamic nature of genome evolution in this important region.

Hemizygosity via Large Inversions Drive Shifts in Berry Color: A second region of high F_{ST} divergence between wild and cultivated grapevines encompassed the berry color region (**Figure 3C**). It, too, had more SVs than the genomic background ($p = 3.3x10^{-5}$, $\square^2$). The region is interesting because *sylvestris* has dark berries, representing the ancestral condition (This et al., 2006), and because white berries have originated in a subset of *sativa* cultivars. SVs have been implicated in the origin of white berries, especially a 5'*Gret1* retroelement insertion that reduces the expression of a *myb* gene

(*VviMYBA1)* that regulates anthocyanin biosynthesis (Kobayashi et al., 2004). Subsequently, it was shown that a frameshift mutation in a second *myb* gene (*VviMYBA2)* was also necessary to cause white berries (Walker et al., 2007). Surprisingly, these two mutations (the *Gret1* insertion and the *VviMYBA2* frameshift) are heterozygous in most grape cultivars (Fournier-Level et al., 2009). Somatic mutations causing white grapes delete the functional *VviMYBA1* and *VviMYBA2* alleles, leaving the plant hemizygous for null alleles (Walker et al., 2006, Yakushiji et al., 2006).

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 vs. Cab08, revealing a large 4.82Mb (chr02: 12,295,113bp-17,118,777bp) inversion in Char04 (**Figure 5A**). This inversion was confirmed by comparison to PN40024, by the identification of discordant and split reads at the junctions (**Figure S11**), and by the lack of an inversion between Cab08 and PN40024 (**Figure 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 (**Figure S11**). 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.

Another study has recently characterized the somatic mutations that led to white berries in the Tempranillo cultivar (Carbonell-Bejerano et al., 2017). The mutations included hemizygosity at both *VviMybA1* and *VviMybA2*, along with a series of complex series of SVs that included a putative 4.3Mb inversion on chromosome 2 (Carbonell-Bejerano et al., 2017). 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 (Online.). The phylogeny shows that white-berry mutations occurred independently on several occasions (**Figure 5B**). We then chose six pairs of closely

related red 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 (Online.). All six contrasts yielded evidence for a large inversion encompassing the *MybA* region (**Figure 5C**). The inferred inversions ranged from 3.85Mb to 4.82Mb 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 vs. Kishmish vatkana) and near the *MybA* region in other contrasts (**Figure 5C**).

Somatic mutations to white berries are associated with hemizygosity of *MybA* genes and with large, Mb scale inversions. But why are large inversions associated with the white berry phenotype? We can think of three explanations. The first is that the inversion contains non-*MybA* genes that also affect phenotype. To investigate this hypothesis, we mapped gene expression data from red and white berries collected over four stages of berry development (Massonnet et al., 2017) and counted the proportion of differentially expressed genes between color morphs. The proportion of differentially expressed genes within the Char04 inversion was no higher than the genomic background ($p = 0.82$; $\square^2$), suggesting that the inversion is not enriched for genes that contribute to berry color. The second explanation is that inversions are common because of underlying properties of the chromosome 2 sequence. The region does not contain any obvious differences in TE distribution or other gross features (**Figure 1A**), but this explanation remains a possibility, particularly given flanking *copia* elements in Char04. Finally, it is possible that similar inversions have occurred commonly throughout *Vitis* genome evolution, that most are lost because they are selected against (**Figure 2B**), but that only a few affect an obvious phenotype - like berry color - that is prone to human intervention. Whatever the underlying cause(s) for these large inversions, they represent a stunning example of convergent evolution via independent SV events.

CONCLUSIONS:

Altogether, our sequencing of the Chardonnay genome, coupled with comparisons to the genomes of Cabernet Sauvignon and Pinot Noir, have provided insights into the evolution of clonally propagated genomes and into plant genomes more broadly. One insight is that grapevine genomes are riddled with heterozygous SVs, to the extent that they comprise up to 15% of the chromosome by length and cause 1 in 7 genes to the be hemizygous. Although negative selection acts against SVs and is particularly strong against inversions and translocations, SVs nonetheless accumulate in cultivars due to clonal propagation and the sheltering of recessive somatic mutations. Only a small proportion of these SV events are estimated to be adaptive, but some clearly contribute to agronomically important phenotypes, such as hermaphroditism and white berry color. Although we cannot yet pinpoint the mutations that led to the former, we have shown that the latter originated on multiple, independent occasions via complex and large Mb-scale inversions.

MATERIALS AND 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 (Chin et al., 2016). The preparation of paired-end Illumina libraries followed (Zhou et al., 2017). SMRTbell libraries were sequenced on a PacBio RSII system, generating a total of 24.2 Gbp (~55X). Illumina sequencing was conducted on a HiSeq4000 sequencing platform in 150 paired-end (PE) mode (54X) and 100 PE mode (124X). 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: PRJNAXXX.

Genome assembly was based on a hybrid strategy, that utilized both long and short sequencing reads, and that merged three separate assemblies. The first assembly utilized Canu v1.5 (Koren et al., 2017) to assemble SMRT reads, based on default parameters and with a genome size of 600 Mb. A second, hybrid assembly was generated

with DBG2OLC (Ye et al., 2016) based on contigs from the Platanus assembly and the longest 30X Pacbio reads. The Platanus assembly was based on (Kajitani et al., 2014) v1.2.4 with default settings, using trimmed 178X Illumina paired-end 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 (Chakraborty et al., 2018, Solares et al., 2018), except that the k-mer size was increased to 31. The k-mer size was increased to minimize the number of misassemblies by including 90% of all k-mers reported by the meryl program within the Canu package (Koren et al., 2017). The consensus stage for the DBG2OLC assembly was performed with PBDAGCON (Chin et al., 2013) and BLASR (Chaisson and Tesler, 2012). Third, PacBio genomic reads were assembled using FALCON-Unzip v1.7.7 (Chin et al., 2016). 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 ***Supplemental 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* (Chakraborty et al., 2016), following a broader application of assembly merging based on (Solares et al., 2018). 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 345000 ml 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 v2.3 (using parameter: -j 80). Long reads (> 1,000bp), consisting of ~43X coverage, were used for polishing. The assemblies were also polished twice using *Pilon* v1.16 (Walker et al., 2014) run using default settings. For this purpose, Illumina reads were aligned to the assembly using Bowtie2 v2.32 (Langmead and Salzberg, 2012) and sorted using samtools v1.3 (Li et al., 2009).

BUSCO v2.0 was used to measure gene space completeness and conserved gene model reconstruction of all generated assemblies (Simão et al., 2015). The embryophyta database, which contained 1,440 highly conserved genes, was used to measure gene model reconstruction and estimate assembly completeness. Quast v2.3 (Gurevich et al., 2013) was run to calculate assembly length and N50 on each assembly. Dot plots were generated using nucmer and mumplot from MUMmer4 v3.23(Marçais et al., 2018) with the options: -l 100 -c 1000 -d 10 -banded -D 5. The BUSCO v3 (Simão et al., 2015) pipeline was applied to the final genome assembly, using the embryophyta_odb9 database.

The final assembly included both primary haplotype sequences and alternative contigs (aka 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(Kent, 2002). 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 – e.g., minimum contig length ranged from 0, 10000, 50000, 100000 bp; % query alignment was examined over 18 randomly chosen values between 90% to 99.9999%, and % 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 formulas were generated to calculate PctQA and PctAO. $PctQA = \frac{Aligned\ Query\ Length}{Total\ Query\ Length}$ and $PctAO = \frac{Aligned\ Query\ Length}{Aligned\ Reference\ Length}$. Alignments generated from contigs aligning to themselves were not considered. The scripts and code used for assembly and alternate haplotig reduction are available on GitHub: https://github.com/esolares/CAP

Scaffolding and GapClosing

A Dovetail HiC library was prepared in a similar manner as described previously (Lieberman-Aiden et al., 2009). The library was sequenced on an Illumina platform to produce 211 million 2x100bp paired end reads, which provided 1,624x physical coverage of the genome (1-50kb pairs). The input *de novo* assembly, shotgun reads, and Dovetail HiC library reads were used as input data for HiRise (Putnam et al., 2016). Shotgun and Dovetail HiC library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of HiC read pairs mapped within draft scaffolds were analyzed 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 v4.0 (Marçais et al., 2018) was used to identify and to sever erroneous junctions between contigs. The resulting scaffolds underwent a second scaffolding procedure using SSPACE-longreads v1.1 (Boetzer and Pirovano, 2014) with default parameters and a minimum coverage of 10 reads (options: -l 10). Gaps were closed using PBjelly (PBSuite v15.8.24; (English et al., 2012)) 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 (A.FA. and P., 2015) using the repeat library previously developed for *V. vinifera* cv. Cabernet Sauvignon (Minio et al., 2018). *Ab initio* prediction of protein-coding genes was carried out with

SNAP (ver. 2006-07-28) (Korf, 2004), Augustus v3.0.3 (Stanke et al., 2006), and GeneMark-ES v4.32 (Lomsadze et al., 2005). *Ab initio* predictions were combined with the predictions of Augustus trained with BUSCO genes, as well as the gene models annotated with PASA v2.1.0 (Haas et al., 2003), using the experimental data reported in ***Supplemental text 2***. RNA-seq data obtained from public databases (***Supplemental text 2*)** were *i*) assembled using both an on-genome strategy, with Stringtie v1.3.3 (Pertea et al., 2015), and a *de novo* transcriptome procedure, with Trinity v2.4.0 in genome-guided mode setting a maximum intron length of 10Kb (option: --genome_guided_max_intron 10000); *ii*) clustered with CD-HIT-EST v4.6 (Li and Godzik, 2006), with coverage threshold 90% (option: -c 0.9); and *iii*) filtered with Transdecoder v3.0.1 (Haas and Papanicolaou, 2016), which retained only genes with a full-length open reading frame (ORF). Experimental evidences (transcripts and proteins) were mapped on the genome using Exonerate v2.2.0 (Slater and Birney, 2005)and PASA v2.1.0 (Haas et al., 2003), and together with all the predictions used as input to EVidenceModeler v1.1.1 (Haas et al., 2008). Weights used in EVidenceModeler are reported in ***Supplemental text 3***. The annotation was refined and enhanced with alternative transcripts using PASA v2.1.0 (Haas et al., 2008) and assembled experimental evidences; parameters used for refining the gene structures are described in ***Supplemental 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 homolog proteins in the RefSeq plant protein database (ftp://ftp.ncbi.nlm.nih.gov/refseq, retrieved January 17th, 2017). Functional domains were identified using InterProScan v5 (Jones et al., 2014) using the databases provided in ***Supplemental text 5.*** Gene models with no significant blast hit against RefSeq plant protein database (HSP<50 amino acids) and lacking any functional domain were discarded. Gene ontology (GO) obtained from InterPro domains and RefSeq homologs with at least 50% of reciprocal coverage and identity were combined using Blast2GO v4 ((Conesa et al., 2005) to assign a functional annotation, gene ontology (GO), and enzyme commission (EC) 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(Marçais et al., 2018). The top 23 scaffolds covered 82% (492 Mb) of the Char04 primary assembly and aligned to the PN40024 chromosomes (Fig. S1), except two long scaffolds with lengths of 20Mb (Char04v1.0_683) and 11Mb (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 (**Figure 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 2 and 3 scaffolds, respectively. For all ensuing analyses, we treated these 22 scaffolds as the Char04 reference genome. We evaluated heterozygosity within this reference for both small variants (SNPs + indels < 50 bp) and large structural variants (SVs ≥ 50 bp). SNPs and indels were called based on 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 adapter 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 (Li, 2014). The bam files were filtered (unique mapping with a minimum mapping quality of 20) and sorted using samtools v1.9 (Li et al., 2009). PCR duplicates introduced during library construction were removed with MarkDuplicates in picard-tools v1.119 (https://github.com/broadinstitute/picard). SNPs and small indels were called with the HaplotypeCaller in GATK v4.0 pipeline, and then filtered following (Zhou et al., 2017).

To identify SVs within the Char04 genome (i.e. between the two haplotypes), we called SVs using the Sniffles pipeline (Sedlazeck et al., 2018b). First, Pacbio reads longer than 500bp were mapped onto Char04 primary assembly using the two aligners Minimap2 v2.14 with the MD flag (Li, 2018) and NGMLR v0.2.7 (Sedlazeck et al., 2018b), separately. Variant calling was then performed with Sniffles. SV analysis outputs (VCF files) were filtered based on the following four steps: *i*) we removed SVs that had

ambiguous breakpoints (flag: IMPRECISE) and also low quality SVs that did not pass quality requirements of Sniffles (flag: UNRESOLVED); *ii*) we removed SV calls shorter than 50 bp; *iii*) we removed SVs with less than 4 supporting reads; and *iv*) 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 lead to detecting more SVs (e.g., 37,169 in total within Char04) than long-read mapping with NGMLR v0.2.7 (23,972 in total within Char04). Given the differences from the two mapping protocols, we built consensus SVs calls using SURVIVOR v1.0.3 (Jeffares et al., 2017). Using the merged SV set, we called genotypes and combined them into a single VCF using the population calling steps of the Sniffles pipeline (Sedlazeck et al., 2018b). The genotypes of SV calls from both programs (NGMLR and Minimap2) were intersected using bedtools v2.25 (Quinlan, 2014) 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. The first consisted then to compare 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 (Marçais et al., 2018). After filtering 1-to-1 alignments with a minimum alignment length of 1,000 bp (delta-filter -1 -l 1000), the show-diff function and NucDiff (Khelik et al., 2017) 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 ~140X coverage, were mapped onto Char04 genome using Minimap2 and NGMLR, as described above. SVs

were genotyped based on merged SV calls from both mappers, using the population calling steps of Sniffles pipeline (Sedlazeck et al., 2018b). 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 programs were intersected using bedtools v2.25 (Quinlan, 2014) 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 ~15X 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 (Zhou et al., 2017). SVs were called with all the population samples (69 in total, see below) using both LUMPY v0.2.13 (Layer et al., 2014) and DELLY2 v0.7.7 (Rausch et al., 2012). For LUMPY, the read and insert lengths were extracted from mapping files (bam files) for each sample using samtools v1.9 (Li et al., 2009), and the SVs were genotyped using SVTyper (Layer et al., 2014). The SV calls from DELLY and LUMPY were merged using SURVIVOR v1.0.3 (Jeffares et al., 2017). SVs for all 69 population samples presenting the following five criteria were retained: *i*) a minimum of three PE reads or split reads (SR) supporting the given SV event across all samples; *ii*) SV calls with precise breakpoints (flag PRECISE); *iii*) SVs passing the quality filters suggested by DELLY and LUMPY (flag PASS); *iv*) SV length ≥ 50 bp; *v*) complex SVs, consisting of, or overlapping SVs were excluded. SV calls for Cab08 and Char04 were extracted using vcftools v0.1.13(Danecek et al., 2011) 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 v2.25 (Quinlan, 2014) with a minimum reciprocal overlap of 80%. We took the intersect of the DELLY and LUMPY calls to separate SVs into three categories: *i*) shared between methods, which was roughly 74.6% of the SV calls; *ii*) DELLY-specific SVs, and *iii*) LUMPY-specific SVs. We then combined the three sets using SURVIVOR (Jeffares et al., 2017) and intersected it with SMRT-based SV calls to get a shared VCF. 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 (Table S4), each of which with coverage > 10X. The mean mapping depth across accessions was 21.6X. 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. 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 v4.0 pipeline, following (Zhou et al., 2017). SNPs and indels were filtered and annotated using SnpEff v4.0 (Cingolani et al., 2012), following (Zhou et al., 2017). 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 Cab04 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: *i*) a minimum number of supporting four reads in LUMPY calls (flag SU, which equals to SP+PE) *ii*) 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); *iii*) a mapping quality ≥ 20 in DELLY calls (flag MAPQ); *iv*) 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 (MEIs)

We used the filtered BAM files with PCR duplicates masked for each sample as input for detecting polymorphic transposable elements (TEs) with the Mobile Element Locator Tool (MELT) v2.1.4 (Gardner et al., 2017). 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 an Hidden Markov Model (HMM) method to build consensus sequences for the TE families that represented > 4% of the Char04 reference (i.e., LINES: L1; LTR retrotransposons: *Copia* and *Gypsy*; and DNA transposons: MuDR and MULE-MuDR; **Table S2**). 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: *i)* TE variants compared to Char04 genome were detected for each accession individually using IndivAnalysis; *ii*) all polymorphic TE calls from all samples were merged to detect breakpoints of insertions in the reference genome using GroupAnalysis; *iii*) the resulting variants file was then used to call genotypes of all insertions for each sample using the Genotype utility; *iv*) 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 program (Gardner et al., 2017). These four steps were performed for each TE family, separately. In order 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 (i.e., 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 50bp.

Population genetic analyses

Our analyses of the Illumina population data resulted in SV calls for a wide variety of events, including insertions (INS), deletions (DEL), duplications (DUP), inversions (INV), and translocations (TRA). In general, variant calling using short-read alignment allowed to detect only short insertions (INS, Figure S2), and we therefore excluded INS variants from further analyses. Complex variants, which were defined as composite variant 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 (i.e, DEL, DUP, INV, 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 (PCA) for SNP and SV calls using PLINK v1.9 (Purcell et al., 2007)(Figure S6).

SNPs and SVs with a minor allele frequency > 0.1 were used for analyses of linkage disequilibrium (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 within a physical distance of 300 kbp, using PLINK v1.9 (Purcell et al., 2007). The decay of LD against physical distance was estimated using nonlinear regression of pairwise r^2 *vs*. the physical distance between SNPs or SVs mid-positions (Hill and Robertson, 1968).

Since LD decayed within 20 kbp in both the wild and the cultivated samples, we divided the Char04 genome into 24,056 non-overlapping windows of 20 kbp 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 v0.1.13 (Danecek et al., 2011), which calculates indices of the expected genetic variance between and within species allele frequencies. We then averaged F_{ST} values of all sites within each 20 kbp non-overlapping window.

We calculated the unfolded site frequency spectrum (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: 4-fold sites, and non-synonymous sites: 0-fold 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 (CLR) test implemented in SweeD v3.2.1(Pavlidis et al., 2013). Synonymous sites at genomic windows with top 5% CLR 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 (**Figure S7**). 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 v0.912 (Korneliussen et al., 2014) 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 (EM) algorithm was set to 10,000. Based on these population structure results (**Figure S5**), the down-sampled set of 12 cultivars was chosen to represent major genetic clusters and also to represent accessions with the least missing data (**Table S4**).

Distribution of fitness effects (DFE) of SVs

We applied the program DFE-α v2.15 to estimate the distribution of fitness effects (DFE) and the proportion of adaptive variants (α) for non-synonymous SNPs, DELs, DUPs, INVs, TRAs, and MEIs (Keightley and Eyre-Walker, 2007, Eyre-Walker and Keightley, 2009). In these analyses, we used information from synonymous SNPs as the neutral reference, based on the unfolded SFS. First, we fitted a demographic model to the SFS for neutral sites using maximum likelihood (ML). 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 (Keightley and Eyre-Walker, 2007). We performed multiple ML searches, each with a different starting point, and treated the parameter values that produced the highest log-likelihood as the ML 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 gamma distribution and *s* is the mean fitness effect of variants. The ML 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 (Eyre-Walker and Keightley, 2009). We obtained 95% confidence intervals (CIs) for the parameter estimates by analyzing 100 bootstrap replicates of SFS and divergence data sets, which were generated by randomly sampling genes. Following the findings of (Keightley et al., 2016), we used high-quality data from two North American wild *Vitis* species as outgroup (Keightley et al., 2016) to infer the ancestral state of variants. We note, however, that the inference of the ancestral state of SVs are 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 v2.0 (Tataru et al., 2017). The results were similar, so we presented the polyDFE results with 95% CIs obtained from the inferred discretized DFEs from 100 bootstrap data sets.

SVs and sex determination

F_{ST} values for both SNPs and SVs exhibited clear outlier peaks in the sex determination region (**Figure 3**). The SNPs of the sex determination region were phased and imputed based on a genetic map (Hyma et al., 2015) using Shapeit v2.12 (Delaneau et al., 2013), following the study of (Zhou et al., 2017). To examine relationships among different sex haplotypes, we built Maximum Likelihood (ML) trees from SNPs within the region. ML trees were based on 10,000 bootstrap replicates, as implemented in MEGAX (Kumar et al., 2018). We built trees for the two regions, corresponding to the peaks of SNP divergence identified by Zhou et al. (Zhou et al., 2017). BEAST v1.8.0 (Drummond et al., 2012)was applied to calculate genetic divergence, based on 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 million MCMC cycles, The divergence time between haplotypes was bases on a genome-wide divergence time of 46.9 million years ago between M. *rotundifolia* and *Vitis* species. (Ma et al., 2018)

The boundaries of the sex determination region were determined by mapping the coding sequences (CDS) of the chr02:4840000 - 498000 region from PN40024 12X.v2 (Canaguier et al., 2017) to the Char04 and Cab08 references. For both Chardonnay and Cabernet Sauvignon haplotypes, gene models were refined by mapping all the CDS identified in the four haplotypes onto Char04 and Cab08 genome assemblies, separately, using GMAP v.2015-11-20 with default parameters (Wu and Watanabe, 2005).

We analyzed gene expression data from the three grape flower genders. Raw sequencing data were obtained from the Short Read Archive (SRP041212). Reads were first trimmed using Trimmomatic v.0.36 (Bolger et al., 2014) 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 (Chin et al., 2016) separately, using HISAT2 v.2.0.5 (Kim et al., 2015) with the following options: --end-to-end --sensitive --no-unal. The Bioconductor package GenomicAlignments v.1.12.1 (Lawrence et al., 2013) 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 genders (i.e. Male vs. Female, Male vs. Hermaphrodite, Female vs. Hermaphrodite) was performed using the Bioconductor package DESeq2 v1.16.1 (Love et al., 2014) using samples of the last two flower growth stages as replicates to allow enough statistical power. These same data were analyzed previously using the same methods, based on mapping to the PN40024 reference (Zhou et al., 2017). 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 sex-determination tend to show male-biased expression. The differences between studies reflect mapping biases between the presumed female haplotype in the PN40024 (Picq et al., 2014) 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 color

We compared genomes of two cultivars with dark blue berries (PN and Cab08) with two cultivars with light green berries (Char04 and Sultanina) using pairwise whole-genome

alignments and called SVs using the MUMmer4 pipeline. Dot plots were generated using mumplot from (mumplot -l 100 -c 1000 -d 10 -banded -D 5) for chromosome 2 where the berry color QTL located. For Char04 and Cab08, we verified the SV calls using the Sniffles pipeline (Sedlazeck et al., 2018b) after mapping SMRT reads onto the PN40024 reference genome using both the Minimap2 (Li, 2018) and NGMLR (Sedlazeck et al., 2018b). 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 color in different cultivars.

To identify whether other green berry accessions housed large inversions that include the berry color genes, we determined the orientation of the rearranged chromosome fragments and putative breakpoints from bam files of discordant PE reads and split reads (SP) after mapping short-reads to the PN40024 genome V2.0 (Canaguier et al., 2017). Reads were mapped using the BWA-MEM algorithm in bwa-0.7.12 (Li, 2014). The discordant reads and split reads were extracted using samtools v1.9 (Li et al., 2009) and LUMPY v0.2.13 (Layer et al., 2014). To select breakpoints distinguishing genomes of red- and white-berry cultivars, the discordant, the splitter, and the original bam files were inspected visually using IGV v2.2 (Thorvaldsdóttir et al., 2013).

To detect potentially hemizygous regions on chromosome 2, we calculated runs of homozygosity (ROH) for each sample using the software PLINK v1.9 (Purcell et al., 2007) 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. CNV analyses were conducted in cnv-seq (Xie and Tammi, 2009) using the neighboring grapevines with green and dark blue berry colors with bam file of the former as test while bam file of the later as a reference. The log2 values of the adjusted copy number ratio were plotted in R.

Gene expression analyses of the berry color region utilized the raw RNA-seq data from SRA: SRP049306-SRP049307 (Massonnet et al., 2017). The data were generated from berries sampled during berry development at four stages, including two before and two after veraison, from 10 Italian varieties (5 red and 5 white). RNA-seq data were mapped onto the Char04 reference and analyzed as described in the previous section. Differential gene expression analysis was performed for each berry growth stage,

separately, by comparing samples from red cultivars with berries from with varieties. Genes presented an adjusted *P*-value ≤ 0.05 between red and white cultivars were considered as significantly expressed. Gene expression analyses focused on the 173 genes in the Char04 chromosome 2 inversion.

ACKNOWLEDGEMENTS:

We are grateful for the technical assistance of Rosa Figueroa-Balderas, the services of the Genomics High Throughput Facility at UC Irvine, and the comments of A. Muyle, D. Seymour, D. Koenig, T. Batarseh and G. Martin. This work was supported by seed funding from UC Irvine, NSF grants 1542703 to BSG, NSF grant #1741627 to BSG and DC, and support to DC by J. Lohr Vineyards and Wines, E. & J. Gallo Winery, and the Louis P. Martini Endowment in Viticulture.

FIGURE LEGENDS:

Figure 1: Structural heterozygosity within Chardonnay 04 and comparisons of structural variation between Chardonnay 04 and Cabernet Sauvignon 08. A) The circle plot reports heterozygous SVs within the Char04 genome. The outermost circle denotes the number and size of chromosomes (gray), followed by gene density (red), TE density (black), deletions (orange), duplications (dark red), insertions (green), inversions (blue) and with translocations represented in the middle of the circle in purple. B) A demonstration of hemizygous genes of Char04 supported by both homozygosity and coverage. The vertical colored lines in the grey coverage plot shows heterozygous sites. Both coverage and heterozygous sites support a complete hemizygous gene (Vitvi02g00781), a partial hemizygous gene (Vitvi02g00783). C) A Venn diagram showing the common and specific SVs detected by each method between Cab08 and Char04. The SVs shared between Illumina and Pacbio calls provide the basis for criteria to identify SVs from the diversity panel.

Figure 2: SVs are strongly deleterious and under purifying selection. A) The unfolded site frequency spectrum (SFS) of different types of SVs compared to presumably neutral synonymous SNPs (Syn) and nonsynonymous SNPs (Nsyn) for samples of 12 wild (top) and 12 cultivated (bottom) grapevines. The types of SVs plotted include duplications (DUP), TE polymorphisms (MEI), deletions (DEL), translocations (TRA) and inversions (INV). B) The inferred distribution of fitness effects ($N_e s$) for SVs and nonsynonymous SNPs in wild (left) and cultivated (right) grapevines. C) The proportion of adaptive variation (□) in wild and cultivated grapevines.

Figure 3: Population genetics of SVs associated with grapevine domestication. A) The recessive (number of homozygous SVs per grapevine), heterozygous and additive (the number of heterozygous SVs plus two times the number of homozygous SVs per grapevine) load in wild and cultivated grapevines for SVs compared to presumably neutral sSNPs. 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 between *sylvestris* (n = 12) and *sativa* (n = 50) sample across the genome, based on F_{ST} of SVs within 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 color loci are indicated. The *x*-axis indicates the number and size of chromosomes across the genome. D) The same as panel C, except genetic differentiation is based on SNP data.

Figure 4: Haplotypes of the sex region and the evolution of sex in grapevine. A) Comparison of the sex determination 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. * denotes the gene *VviAPT3* 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 novel functional annotation in Char04 (*INP1*). Protein-coding genes are colored according to their functional annotation. Genes that are not shared among genome assemblies are colored in black. *PAP2*, Phosphatidic acid phosphatase 2; *VviAPT3*, Adenine phosphoribosyltransferase 3; *WRKY21*, WRKY transcription factor 21; *FMO*, Flavin-binding monooxygenase; *PLATZ*, Plant AT-rich sequence and zinc-binding protein; *TPR*, tetratricopeptide repeat; *ETO1*, Ethylene Overproducer-like 1; *KASIII*, 3-Oxoacyl-(acyl-carrier-protein) synthase III; *INP1*, INAPERTURATE POLLEN1; *TPP*, Trehalose-6-phosphate phosphatase. Black arrows highlight genes that are found on a single haplotype in the Cabernet Sauvignon genome and on two haplotypes in Chardonnay. B) Gene expression values of each flower gender 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 $\log_2^{(RPM + 1)}$. C) A phylogeny of the sex determination 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.

Figure 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 Chard04 chromosome 2 that reveals a 4.82 Mb inversion overlapping with the major berry color 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 labeled in green. The y-axis is the log2 of 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. TB and TT are abbreviations for Tempranillo Blanco and Tempranillo Tinto. D) A phylogeny, based on genome-wide SNPs from a selection of grape varieties, with the color of text labels reflecting the color of the berry.

CITATIONS:

Arroyo-García R, Ruiz-García L, Bolling L et al. 2006. Multiple origins of cultivated grapevine (Vitis vinifera L. ssp. sativa) based on chloroplast DNA polymorphisms. Mol Ecol. 15:3707-3714.

Beridze T, Pipia I, Beck J, Hsu SC, Gamkrelidze M, Gogniashvili M, Tabidze V, This P, Bacilieri R, Gotsiridze V. 2011. Plastid DNA sequence diversity in a worldwide set of grapevine cultivars (Vitis vinifera L. subsp. vinifera). Bull Georgian Nat Acad Sci. 5:91-96.

Boetzer M, Pirovano W. 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 15:211.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114-2120.

Boualem A, Troadec C, Camps C, Lemhemdi A, Morin H, Sari M-A, Fraenkel-Zagouri R, Kovalski I, Dogimont C, Perl-Treves R. 2015. A cucurbit androecy gene reveals how unisexual flowers develop and dioecy emerges. Science. 350:688-691.

Bowers J, Boursiquot JM, This P, Chu K, Johansson H, Meredith C. 1999. Historical Genetics: The Parentage of Chardonnay, Gamay, and Other Wine Grapes of Northeastern France. Science. 285:1562-1565.

Canaguier A, Grimplet J, Di Gaspero G et al. 2017. A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genom Data. 14:56-62.

Carbonell-Bejerano P, Royo C, Torres-Pérez R et al. 2017. Catastrophic Unbalanced Genome Rearrangements Cause Somatic Loss of Berry Color in Grapevine. Plant Physiol. 175:786-801.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 13:238.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44:e147.

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. 2018. Hidden genetic variation shapes the structure of functional elements in Drosophila. Nat Genet. 50:20-25.

Charlesworth B, Charlesworth D. 1978. A model for the evolution of dioecy and gynodioecy. American naturalist. 975-997.

Chia JM, Song C, Bradbury PJ et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 44:803-807.

Chin CS, Alexander DH, Marks P et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 10:563-569.

Chin CS, Peluso P, Sedlazeck FJ et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 6:80-92.

Coito JL, Ramos MJ, Cunha J, Silva HG, Amâncio S, Costa MM, Rocheta M. 2017. VviAPRT3 and VviFSEX: Two Genes Involved in Sex Specification Able to Distinguish Different Flower Types in Vitis. Front Plant Sci. 8:98.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674-3676.

Da Silva C, Zamperin G, Ferrarini A et al. 2013. 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.

Daccord N, Celton JM, Linsmith G et al. 2017. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nat Genet. 49:1099-1106.

Danecek P, Auton A, Abecasis G et al. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156-2158.

Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 10:5-6.

Dobritsa AA, Coerper D. 2012. 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.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 29:1969-1973.

English AC, Richards S, Han Y et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 7:e47768.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26:2097-2108.

Fang Z, Pyhajarvi T, Weber AL, Dawe RK, Glaubitz JC, Gonzalez Jde J, Ross-Ibarra C, Doebley J, Morrell PL, Ross-Ibarra J. 2012. Megabase-scale inversion polymorphism in the wild ancestor of maize. Genetics. 191:883-894.

Fechter I, Hausmann L, Daum M, Sörensen TR, Viehöver P, Weisshaar B, Töpfer R. 2012. Candidate genes within a 143 kb region of the flower sex locus in Vitis. Mol Genet Genomics. 287:247-259.

Flagel LE, Willis JH, Vision TJ. 2014. The standing pool of genomic structural variation in a natural population of Mimulus guttatus. Genome Biol Evol. 6:53-64.

Fournier-Level A, Le Cunff L, Gomez C, Doligez A, Ageorges A, Roux C, Bertrand Y, Souquet JM, Cheynier V, This P. 2009. 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.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, 1000 GPC, Devine SE. 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 27:1916-1929.

Gaut BS, Seymour DK, Liu Q, Zhou Y. 2018. Demography and its effects on genomic variation in crop domestication. Nat Plants.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29:1072-1075.

Haas BJ, Delcher AL, Mount SM et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654-5666.

Haas BJ, Papanicolaou A. 2016. TransDecoder (find coding regions within transcripts). Google Scholar.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7.

Harkess A, Zhou J, Xu C et al. 2017. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. Nat Commun. 8:1279.

Henn BM, Botigue LR, Peischl S et al. 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci U S A. 113:E440-9.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. Theor Appl Genet. 38:226-231.

Hyma KE, Barba P, Wang M, Londo JP, Acharya CB, Mitchell SE, Sun Q, Reisch B, Cadle-Davidson L. 2015. Heterozygous Mapping Strategy (HetMappS) for High Resolution Genotyping-By-Sequencing Markers: A Case Study in Grapevine. PLoS One. 10:e0134880.

Jaillon O, Aury JM, Noel B et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 449:463-467.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 8:14061.

Jones P, Binns D, Chang HY et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236-1240.

Kajitani R, Toshimoto K, Noguchi H et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24:1384-1395.

Keightley PD, Campos JL, Booker TR, Charlesworth B. 2016. Inferring the Frequency Spectrum of Derived Variants to Quantify Adaptive Molecular Evolution in Protein-Coding Genes of Drosophila melanogaster. Genetics.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 177:2251-2261.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res. 12:656-664.

Khelik K, Lagesen K, Sandve GK, Rognes T, Nederbragt AJ. 2017. NucDiff: in-depth characterization and annotation of differences between two sets of DNA sequences. BMC Bioinformatics. 18:338.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 12:357-360.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. Science. 304:982.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive. Genome Res. 27:722-736.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics. 5:59.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 15:356.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 35:1547-1549.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357-359.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. PLoS Comput Biol. 9:e1003118.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15:R84.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 30:2843-2851.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094-3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078-2079.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22:1658-1659.

Lieberman-Aiden E, van Berkum NL, Williams L et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 326:289-293.

Lin YC, Wang J, Delhomme N et al. 2018. Functional and evolutionary genomic inferences in in Populus through genome and population sequencing of American and European aspen. Proc Natl Acad Sci U S A. 115:E10970-E10978.

Liu XQ, Ickert-Bond SM, Nie ZL, Zhou Z, Chen LQ, Wen J. 2016. Phylogeny of the Ampelocissus-Vitis clade in Vitaceae supports the New World origin of the grape genus. Mol Phylogenet Evol. 95:217-228.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33:6494-6506.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

Ma Z-Y, Wen J, Ickert-Bond SM, Nie Z-L, Chen L-Q, Liu X-Q. 2018. Phylogenomics, biogeography, and adaptive radiation of grapes. Molecular phylogenetics and evolution.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 14:e1005944.

Massonnet M, Fasoli M, Tornielli GB, Altieri M, Sandri M, Zuccolotto P, Paci P, Gardiman M, Zenoni S, Pezzotti M. 2017. Ripening Transcriptomic Program in Red and White Grapevine Varieties Correlates with Berry Skin Anthocyanin Accumulation. Plant Physiol. 174:2376-2396.

Online. MAMAAASMOS.

McGovern P, Jalabadze M, Batiuk S et al. 2017. Early Neolithic wine of Georgia in the South Caucasus. Proc Natl Acad Sci U S A. 114:E10309-E10318.

McKey D, Elias M, Pujol B, Duputié A. 2010. The evolutionary ecology of clonally propagated domesticated plants. New Phytol. 186:318-332.

Migicovsky Z, Sawler J, Gardner KM et al. 2017. Patterns of genomic and phenomic diversity in wine and table grapes. Hortic Res. 4:17035.

Minio A, Lin J, Gaut BS, Cantu D. 2017. How Single Molecule Real-Time Sequencing and Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes. Front Plant Sci. 8:826.

Minio A, Massonnet M, Vondras A, Figueroa-Balderas R, Blanco-Ulate B, Cantu D. 2018. Iso-seq allows genome-independent transcriptome profiling of grape berry development. bioRxiv. 269530

Muyle A, Shearn R, Marais GA. 2017. The Evolution of Sex Chromosomes and Dosage Compensation in Plants. Genome Biol Evol. 9:627-645.

Myles S, Boyko AR, Owens CL et al. 2011. Genetic structure and domestication history of the grape. Proc Natl Acad Sci U S A. 108:3530-3535.

2018. National Grape Research Alliance.

Oberle GD. 1938. A Genetic Study of Floral Morphology and Function in Cultivated Forms of Vitis. Cornell University, June.

OIV. 2015. OIV report on the world vitivinicultural situation.

Parage C, Tavares R, Réty S et al. 2012. Structural, functional, and evolutionary analysis of the unusually large stilbene synthase gene family in grapevine. Plant Physiol. 160:1407-1419.

Pavlidis P, Zivkovic D, Stamatakis A, Alachiotis N. 2013. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol. 30:2224-2234.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33:290-295.

Picq S, Santoni S, Lacombe T et al. 2014. 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.

Purcell S, Neale B, Todd-Brown K et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81:559-575.

Putnam NH, O'Connell BL, Stites JC et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26:342-350.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 47:11.12.1-34.

Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson JV, Bart RS, Verma J, Buckler ES, Lu F. 2017. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat Genet.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 28:i333-i339.

Raymond O, Gouzy J, Just J et al. 2018. The Rosa genome provides new insights into the domestication of modern roses. Nat Genet. 50:772-777.

Roach MJ, Johnson DL, Bohlmann J, van Vuuren HJJ, Jones SJM, Pretorius IS, Schmidt SA, Borneman AR. 2018. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. PLoS Genet. 14:e1007807.

Schmid-Siegert E, Sarkar N, Iseli C et al. 2017. Low number of fixed somatic mutations in a long-lived oak tree. Nat Plants. 3:926-929.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018a. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 19:329-346.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018b. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210-3212.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 6:31.

A.FA. S, HRGP. 2015. RepeatMasker Open-4.0.

Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS. 2018. Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read sequencing. G3 (Bethesda). 8:3143-3154.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34:W435-9.

Sun S, Zhou Y, Chen J et al. 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. Genetics. 207:1103-1119.

This P, Lacombe T, Thomas MR. 2006. Historical origins and genetic diversity of wine grapes. Trends Genet. 22:511-519.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14:178-192.

VanBuren R, Zeng F, Chen C et al. 2015. Origin and domestication of papaya Yh chromosome. Genome Res. 25:524-533.

Velasco R, Zharkikh A, Troggio M et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One. 2:e1326.

Walker AR, Lee E, Bogs J, McDavid DA, Thomas MR, Robinson SP. 2007. White grapes arose through the mutation of two similar and adjacent regulatory genes. Plant J. 49:772-785.

Walker AR, Lee E, Robinson SP. 2006. 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.

Walker BJ, Abeel T, Shea T et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9:e112963.

Wang J, Na JK, Yu Q et al. 2012. Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc Natl Acad Sci U S A. 109:13710-13715.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 21:1859-1875.

Xie C, Tammi MT. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 10:80.

Yakushiji H, Kobayashi S, Goto-Yamamoto N, Tae Jeong S, Sueta T, Mitani N, Azuma A. 2006. 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.

Yao W, Li G, Zhao H, Wang G, Lian X, Xie W. 2015. Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol. 16:187.

Ye C, Hill CM, Wu S, Ruan J, Ma ZS. 2016. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci Rep. 6:31900.

Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS. 2017. Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. Proc Natl Acad Sci U S A.

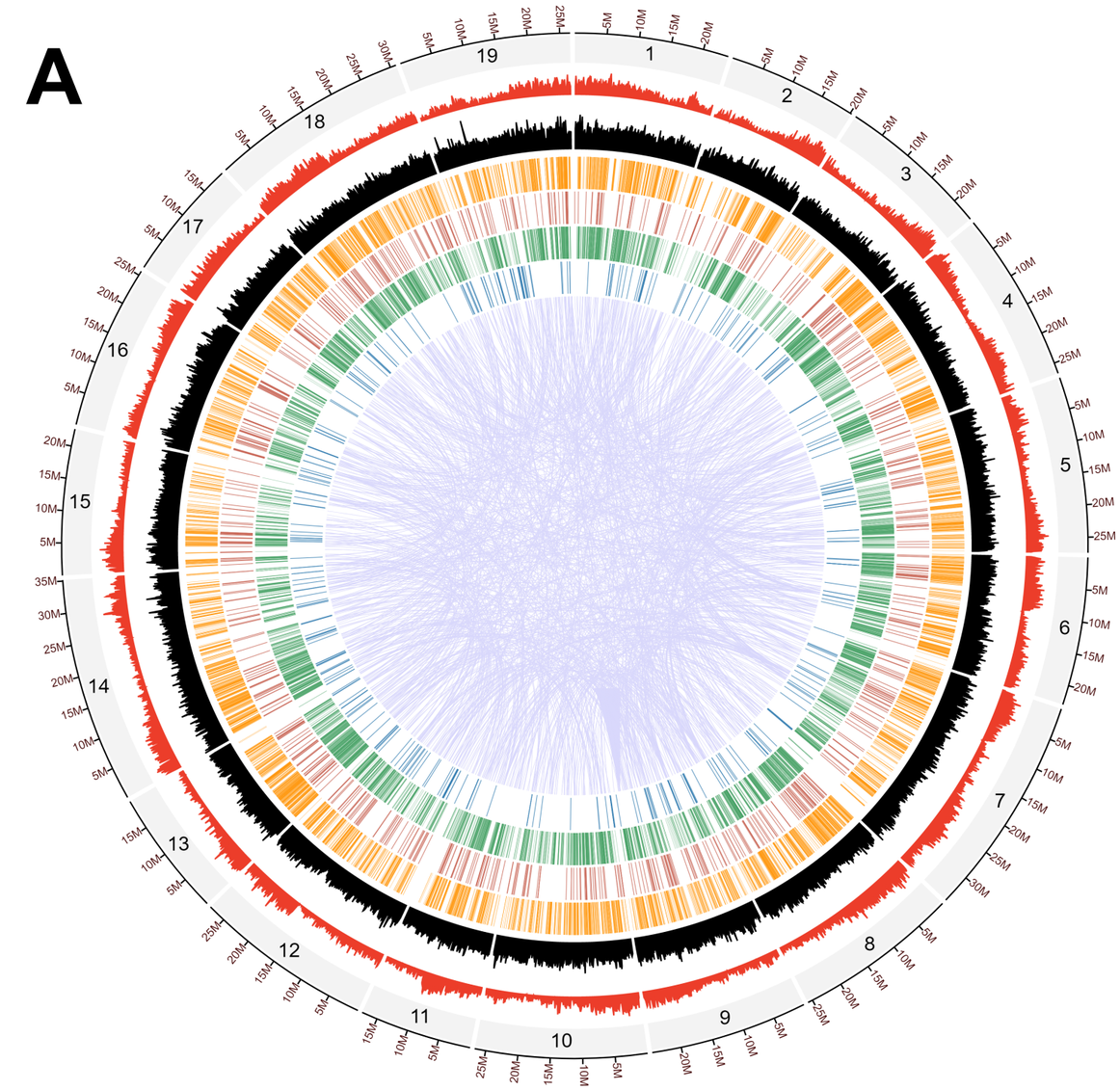
Table 1 Assembly statistics of the Chardonnay genome and two comparatives: the PN40024 reference and the Cabernet Sauvignon (Cab08) assembly.

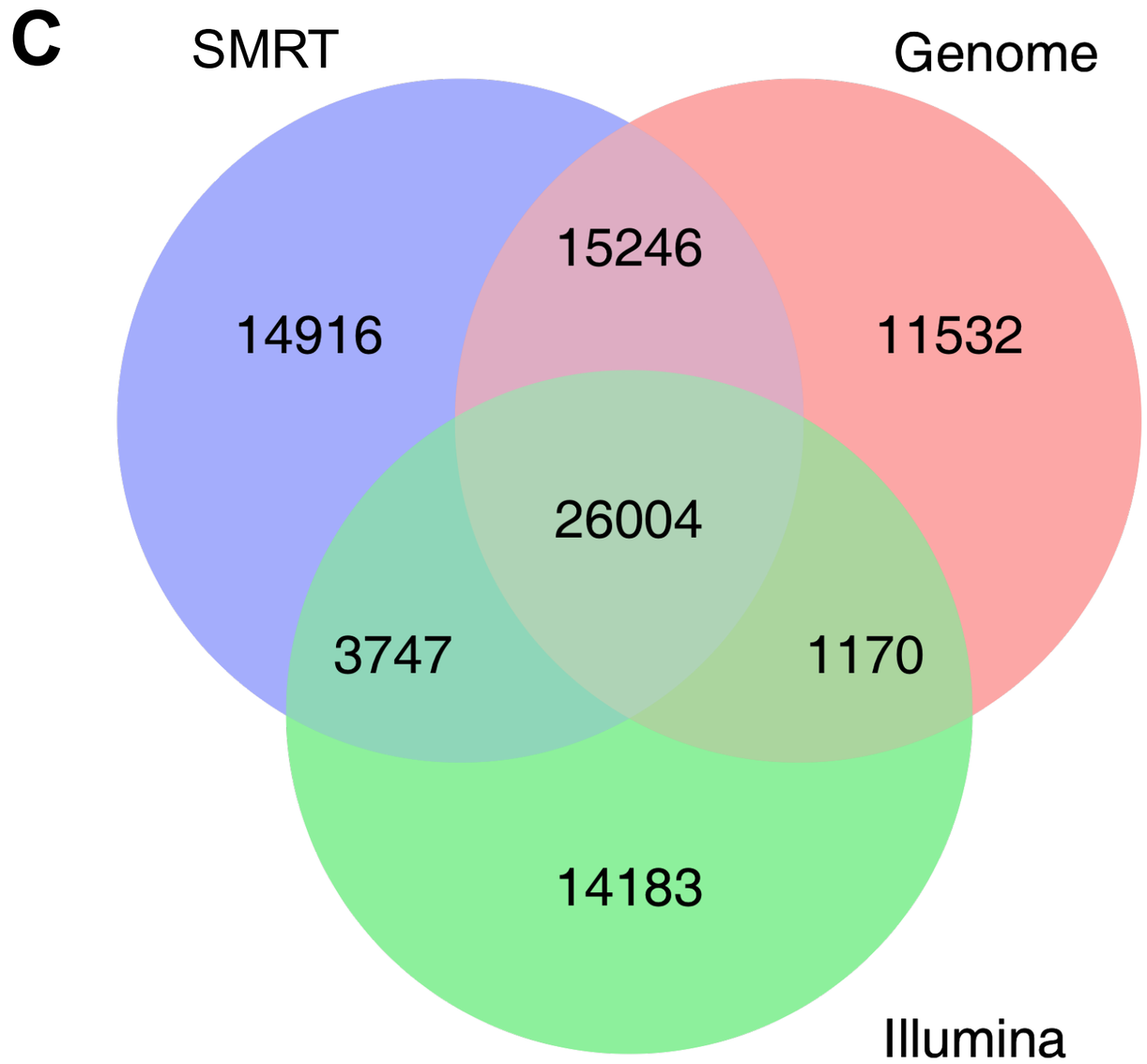
Cultivar	Abbrev.	Assembly statistics			Annotation		
		Assembly size	Contig N50 (Mb)	Scaffold N50 (Mb)	#Genes	%BUSCO	%TE
Chardonnay	Char04[1]	606 Mb	1.24	24.5	38,020	93.4	47.3
Cabernet Sauvignon	Cab08[2]	591 Mb	2.17	-	36,687	92.5	51.1
Pinot Noir	PN40024[3]	486 Mb	0.102	3.4	41,163	96.9	47.0

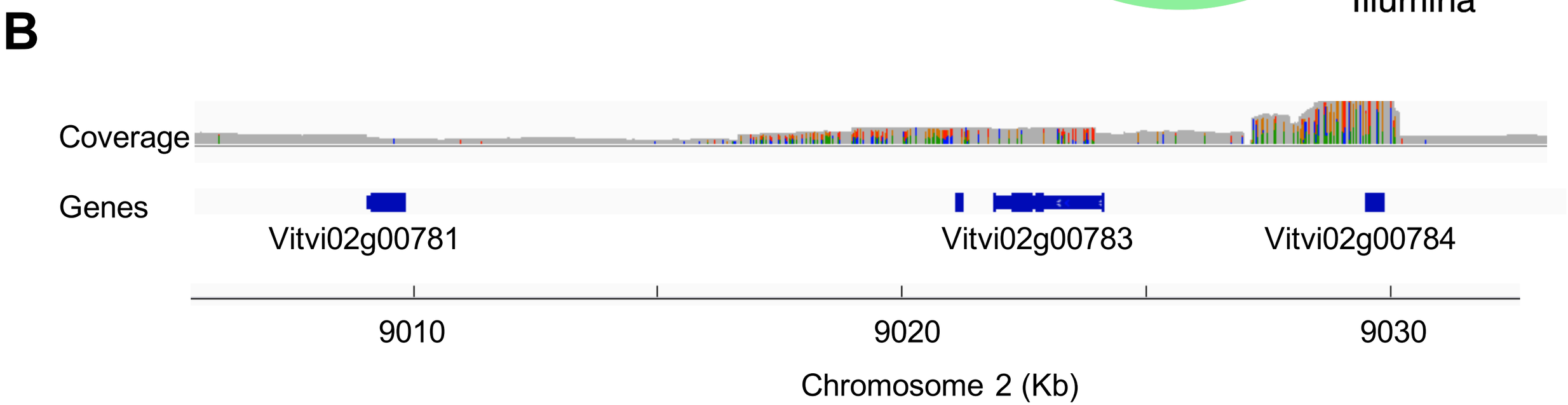
[1] This paper.

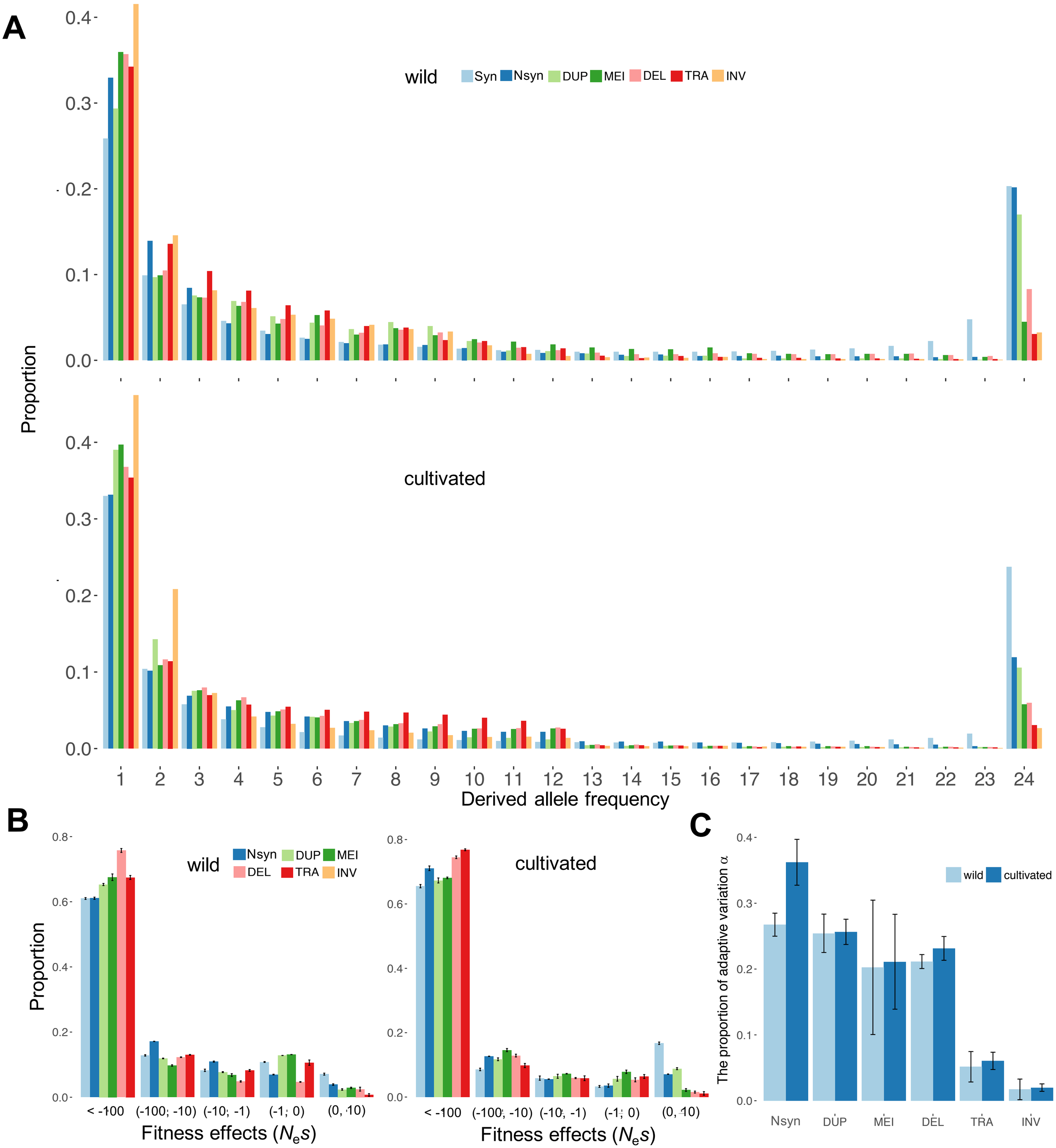
[2] Reference (Chin et al., 2016)

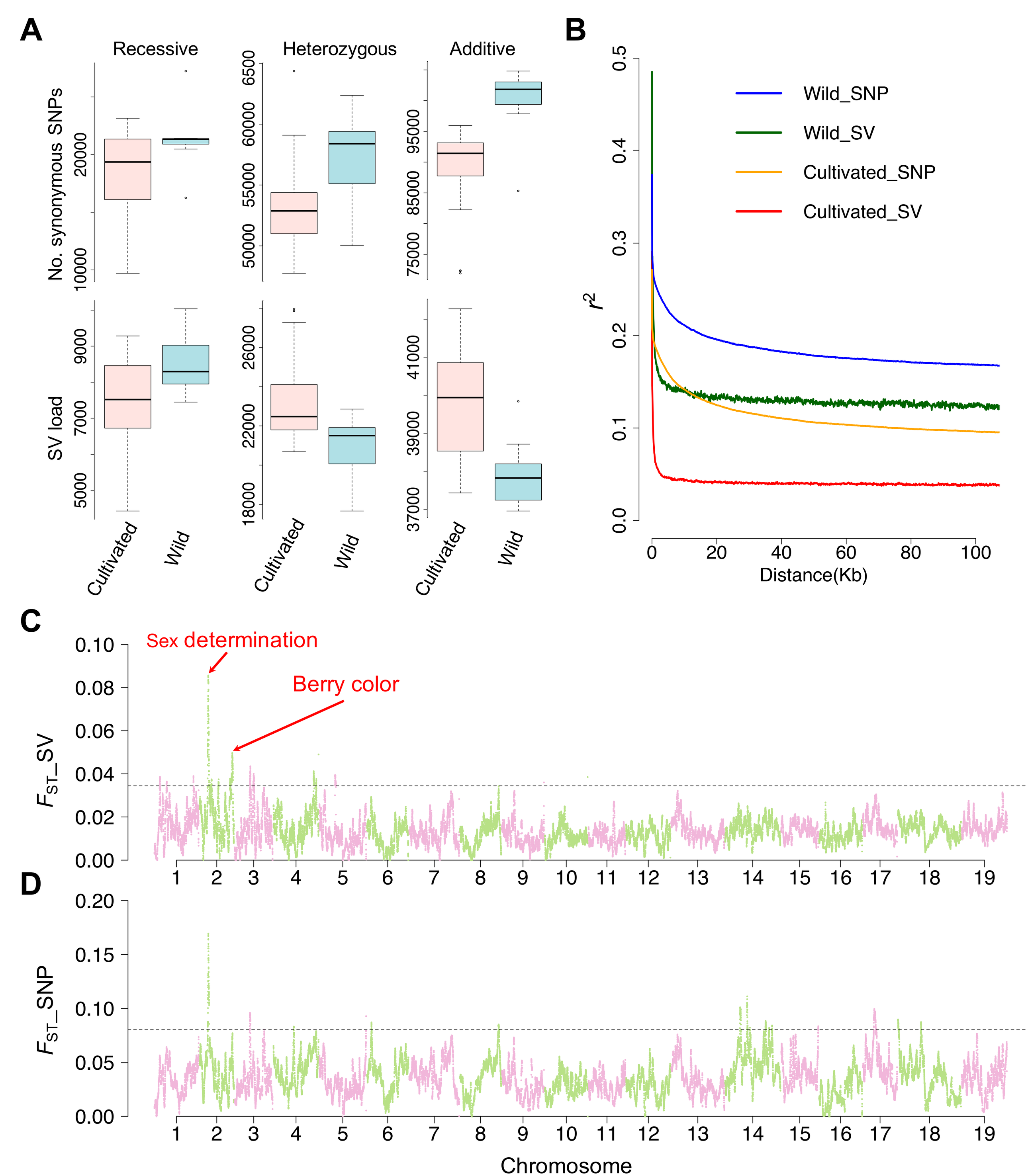
[3] Reference (Canaguier et al., 2017)

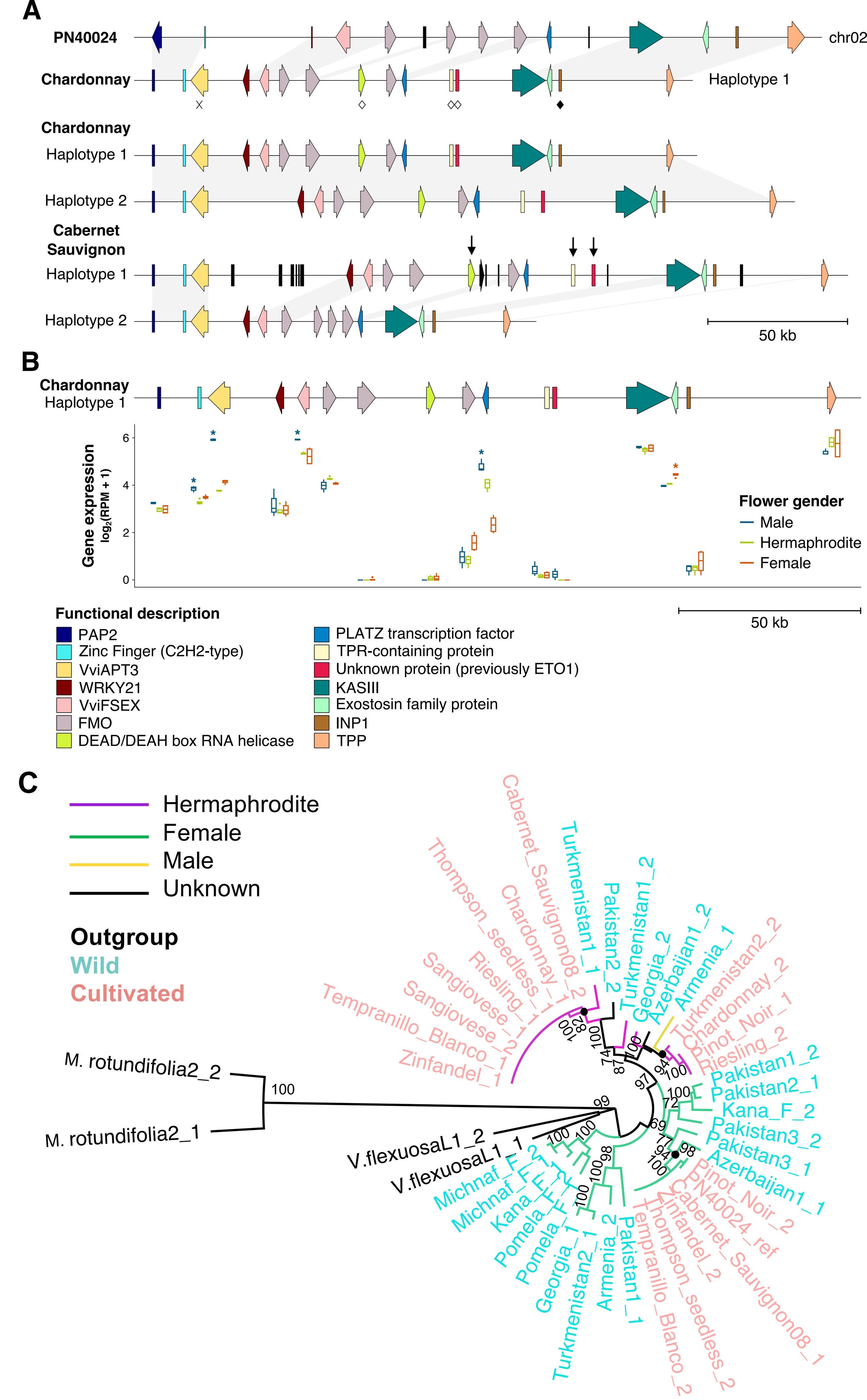

A

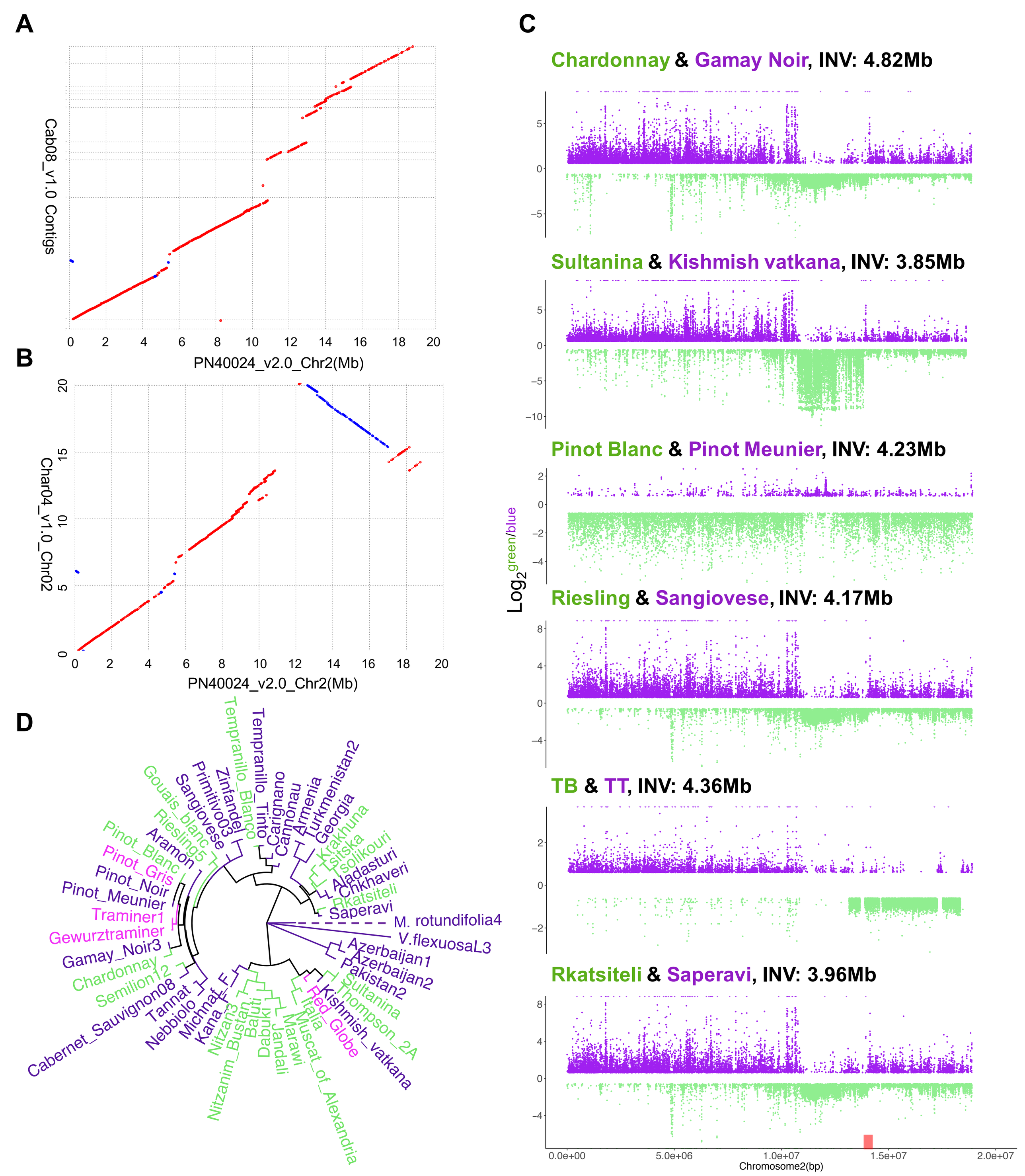

C
SMRT
Genome
Illumina
14916
15246
11532
26004
3747
1170
14183


B
Coverage
Genes
Vitvi02g00781
Vitvi02g00783
Vitvi02g00784
9010
9020
9030
Chromosome 2 (Kb)


A
wild
Syn
Nsyn
DUP
MEI
DEL
TRA
INV
cultivated
Proportion
Derived allele frequency
B
wild
cultivated
Nsyn
DUP
MEI
DEL
TRA
INV
Proportion
< -100
(-100, -10)
(-10, -1)
(-1, 0)
(0, 10)
Fitness effects ($N_e s$)
C
wild
cultivated
The proportion of adaptive variation α
Nsyn
DUP
MEI
DEL
TRA
INV


A
Recessive
Heterozygous
Additive
No. synonymous SNPs
SV load
Cultivated
Wild
B
Wild_SNP
Wild_SV
Cultivated_SNP
Cultivated_SV
r2
Distance(Kb)
C
Sex determination
Berry color
FST_SV
D
FST_SNP
Chromosome


A
PN40024
chr02
Chardonnay
Haplotype 1
Chardonnay
Haplotype 1
Haplotype 2
Cabernet Sauvignon
Haplotype 1
Haplotype 2
50 kb
B
Chardonnay
Haplotype 1
Gene expression
log2(RPM + 1)
Flower gender
Male
Hermaphrodite
Female
50 kb
Functional description
PAP2
Zinc Finger (C2H2-type)
VviAPT3
WRKY21
VviFSEX
FMO
DEAD/DEAH box RNA helicase
PLATZ transcription factor
TPR-containing protein
Unknown protein (previously ETO1)
KASIII
Exostosin family protein
INP1
TPP
C
Hermaphrodite
Female
Male
Unknown
Outgroup
Wild
Cultivated
M. rotundifolia2_2
M. rotundifolia2_1
V.flexuosaL1_2
V.flexuosaL1_1
Cabernet_Sauvignon08_2
Turkmenistan1_1
Pakistan2_2
Turkmenistan1_2
Georgia_2
Azerbaijan1_2
Armenia_1
Turkmenistan2_2
Chardonnay_2
Pinot_Noir_1
Riesling_2
Pakistan1_2
Pakistan2_1
Kana_F_2
Pakistan3_2
Pakistan3_1
Azerbaijan1_1
Pinot_Noir_2
PN40024_ref
Cabernet_Sauvignon08_1
Zinfandel_2
Thompson_seedless_2
Tempranillo_Blanco_2
Pakistan1_1
Armenia_2
Turkmenistan2_1
Georgia_1
Pomela_F_1
Pomela_F_2
Kana_F_1
Michnaf_F_1
Michnaf_F_2
Thompson_seedless_1
Chardonnay_1
Riesling_1
Sangiovese_1
Sangiovese_2
Tempranillo_Blanco_1
Zinfandel_1


A
Cab08_v1.0 Contigs
PN40024_v2.0_Chr2(Mb)
B
Char04_v1.0_Chr02
PN40024_v2.0_Chr2(Mb)
C
Chardonnay & Gamay Noir, INV: 4.82Mb
Sultanina & Kishmish vatkana, INV: 3.85Mb
Pinot Blanc & Pinot Meunier, INV: 4.23Mb
Riesling & Sangiovese, INV: 4.17Mb
TB & TT, INV: 4.36Mb
Rkatsiteli & Saperavi, INV: 3.96Mb
Log2green/blue
Chromosome2(bp)
D
Cabernet_Sauvignon08
Gamay_Noir3
Gewurztraminer
Traminer1
Pinot_Noir
Pinot_Meunier
Pinot_Gris
Pinot_Blanc
Chardonnay
Semilion12
Aramon
Rieslingo5
Gouais_blanc
Sangiovese
Primitivo03
Zinfandel
Tannat
Nebbiolo
Michnat_F
Kana_F
Nitzanim_Bustan
Nitzanim_F
Baluti
Dabuki
Jandali
Marawi
Muscat_of_Alexandria
Red_Globe
Italia
Kishmish_vatkana
Thompson_2A
Sultanina
Pakistan2
Azerbaijan2
Azerbaijan1
V.flexuosaL3
M. rotundifolia4
Saperavi
Rkatsiteli
Chkhaveri
Aladasturi
Solikouri
Tsitska
Krakhuna
Georgia
Turkmenistan2
Armenia
Cannonau
Carignano
Tinto
Temprano
Tempranillo_Blanco