










The genetic basis of sex determination in grapes

Mélanie Massonnet ¹, Noé Cochetel ¹, Andrea Minio ¹, Amanda M. Vondras¹, Jerry Lin ¹, Aline Muyle²,
Jadran F. Garcia ¹, Yongfeng Zhou ², Massimo Delledonne ³, Summaira Riaz¹, Rosa Figueroa-Balderas¹,
Brandon S. Gaut ²✉ & Dario Cantu ¹✉

It remains a major challenge to identify the genes and mutations that lead to plant sexual differentiation. Here, we study the structure and evolution of the sex-determining region (SDR) in *Vitis* species. We report an improved, chromosome-scale Cabernet Sauvignon genome sequence and the phased assembly of nine wild and cultivated grape genomes. By resolving twenty *Vitis* SDR haplotypes, we compare male, female, and hermaphrodite haplotype structures and identify sex-linked regions. Coupled with gene expression data, we identify a candidate male-sterility mutation in the *VviINP1* gene and potential female-sterility function associated with the transcription factor *VviYABBY3*. Our data suggest that dioecy has been lost during domestication through a rare recombination event between male and female haplotypes. This work significantly advances the understanding of the genetic basis of sex determination in *Vitis* and provides the information necessary to rapidly identify sex types in grape breeding programs.

¹Department of Viticulture and Enology, University of California Davis, Davis, CA 95616, USA. ²Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA 92697, USA. ³Dipartimento di Biotecnologie, Università degli Studi di Verona, 37134 Verona, Italy. ✉email: bgaut@uci.edu; dacantu@ucdavis.edu

Plant species possess a variety of mating systems. Some are monoecious, with separate male and female flowers on the same plant, and others are hermaphroditic, with bisexual flowers. Occasionally plants have separate male and female individuals, a mating system called dioecy. Dioecy ensures outcrossing, but it occurs in only 5–6% of angiosperms^{1,2}. Despite its rarity, dioecy is widespread phylogenetically, suggesting it has evolved independently on multiple occasions.

Dioecy has been the focus of numerous evolutionary and genetic studies, both because of its multiple evolutionary origins and because several economically important crops are dioecious. A common hypothesis about the origin of dioecy is the two-locus model, which requires that dioecy evolved from an hermaphroditic ancestor in two steps^{1,3,4}. The first step includes a recessive mutation that interrupts male function. Individuals with a homozygous male-sterility mutation retain only female function, and a population with this mutation contains both females and hermaphrodites (gynodioecy). The second step requires a dominant mutation that suppresses female function, leading to males. This two-locus system can maintain separate sexes only if the two loci are completely linked, because recombination between them could restore hermaphrodites^{1,5}. Support for the two-locus model has been found in several species, including papaya⁶ (*Carica papaya*), strawberry⁷ (*Fragaria virginiana*), *Silene latifolia*⁸, *Actinidia* spp.⁹ and grapes¹⁰ (*Vitis* spp.). However, the sterility mutations that cause dioecy have not been identified completely in any species¹¹. Thus far, the best candidates are from asparagus and kiwifruit^{9,12,13}. In asparagus, for example, females lack a gene associated with tapetal development¹³ and mutant males without a putative female-suppressor gene revert to hermaphrodites¹².

Here we study the sex-determining region (SDR) within the genus *Vitis*. All ~70 wild *Vitis* species are dioecious¹⁴, suggesting that dioecy has been conserved since the origin of the genus. One *Vitis* species, the cultivated grapevine (*Vitis vinifera* ssp. *vinifera*; hereafter *Vv vinifera*), has reverted to hermaphroditism, even though its wild ancestor *Vitis vinifera* ssp. *sylvestris* (hereafter *Vv sylvestris*) is dioecious. This shift of mating system occurred during domestication ~8000 years ago¹⁵, perhaps following a rare recombination event between male (M) and female (F) haplotypes^{10,11}. Therefore, *Vitis* spp. have individuals of three types (Fig. 1a, b): (i) males with flowers that have reduced pistils, with neither stigma nor style development, (ii) females with flowers containing reflexed anthers and stamens that release sterile pollen grains¹⁶, and (iii) hermaphrodites within *Vv vinifera*, which have perfect flowers with functional pistils and stamens that bear fertile pollen. The three types are determined by the genotype at the SDR. Males are heterozygous for male and female haplotypes (MF), females are homozygous (FF), and cultivated *Vv vinifera* hermaphrodites are either homozygous for hermaphrodite haplotypes (HH) or heterozygous (HF).

Previous genetic and genomic studies have identified the approximate boundaries of the SDR in grapes^{10,17,18}. In *Vitis* spp., the SDR maps genetically to ~150 kbp of chromosome 2 that contains between 15 and 20 genes^{10,18}. Polymorphisms within the region have high linkage disequilibrium in *Vv sylvestris*, suggesting low or no recombination between M and F haplotypes¹⁰. It has been hypothesized that this region contains the recessive male-sterility and dominant female-sterility alleles predicted by the two-locus model, and their identification has been attempted by comparative gene expression analyses^{10,19}. One candidate gene, the adenine phosphoribosyltransferase gene *VviAPT3*, is expressed in the carpel primordial of male plants, suggesting a role in pistil abortion²⁰.

Until recently, a major limitation in the study of *Vitis* sex determination has been that the *Vv vinifera* reference genome represented only a partially assembled F haplotype²¹. More recent

work has resolved the partial sequence of four SDR haplotypes, including three H and one F haplotypes²². Yet, despite substantial progress, our understanding of the SDR and the potential determinants of sex have been hampered by the absence of information from M haplotypes.

To fill this gap, we report nine phased diploid genomes of cultivated hermaphrodites, and wild male and female grapes. All of these genomes are based on high-coverage, Pacific Biosciences (PacBio) long-read sequencing. For each genome, haplotypes within the genetically defined SDR have been curated manually, and transcripts expressed from the region have been measured during early and late stages of flower development in male, female, and hermaphrodite plants. With these extensive sequence and expression data, we compare the F, H, and M haplotypes to better define the SDR, identify candidate sex-determining genes, and assess whether H haplotypes owe their origin to a recombination event.

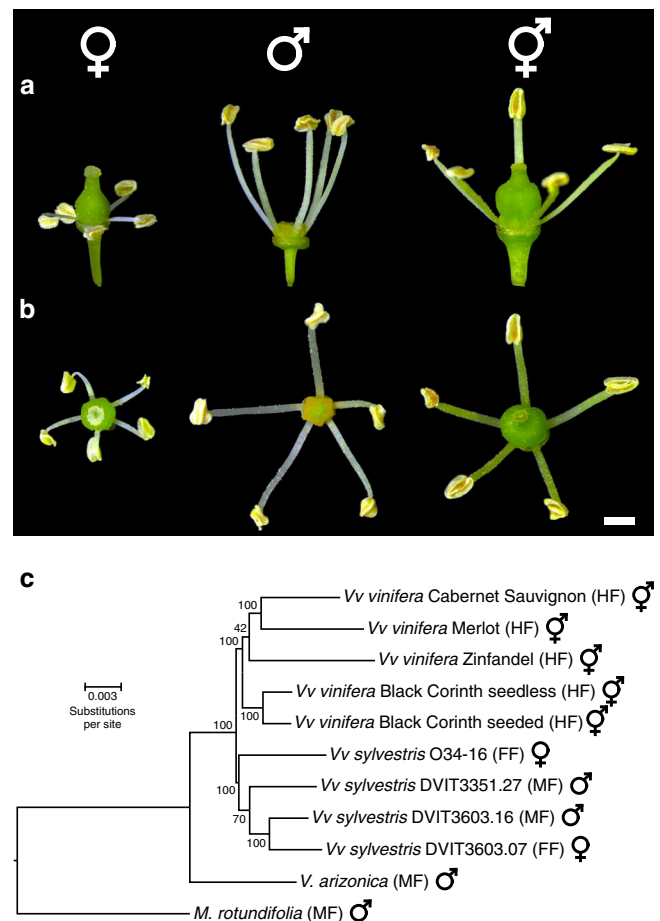


Fig. 1 The morphology of flower sexes in grapes and a phylogenetic analysis of wild and cultivated species. Side view (a) and top view (b) of dioecious *Vv sylvestris* female O34-16 (left), male DVIT3351.27 (middle), and of hermaphrodite *Vv vinifera* Chardonnay (right). Scale bar = 1 mm. c A phylogenetic tree predicted from whole-genome proteome orthology separates species by taxonomy and not by sex genotype. *M. rotundifolia* is an outgroup to the *Vitis* ingroup²³. For each individual, the genotype of the sex-determining region is indicated in parentheses followed by its corresponding sex type. The symbols ♀, ♂ and ♀ represent female, male, and hermaphrodite individuals, respectively. Numbers associated with nodes reflect bootstrap values (see “Methods”). Scale bar is in the unit of the number of substitutions per site.

Results

Sex-specific haplotypes are conserved throughout *Vitis* spp. We sequenced and assembled the complete genomes of eight *Vitis* accessions, including three hermaphrodite *Vv vinifera* cultivars (Merlot, Black Corinth seedless and Black Corinth seeded), four *Vv sylvestris* accessions (two females and two males), and one male *V. arizonica*. In addition, the genome of one male *Muscadinia rotundifolia* was sequenced as a dioecious outgroup to *Vitis* spp.²³ (Fig. 1c). Each genome was based on Single Molecule Real Time (SMRT) DNA sequencing and de novo assembled with FALCON-Unzip²⁴, which produces partially phased diploid genomes. We also included two publicly available genomes from the Cabernet Sauvignon and Zinfandel grape cultivars, both of which were sequenced and assembled with the same approach^{24–26}. All diploid assemblies were highly contiguous and covered approximately twice the expected haploid genome size²¹ (Supplementary Data 1). The Cabernet Sauvignon assembly was improved further with HiC proximity-based ligation, optical mapping, and multiple scaffolding methods, resulting in two phased copies, hap1 and hap2, of all 19 chromosomes.

The SDR was first identified by aligning primer sequences of sex-linked markers^{10,17} to chromosome 2 of the Cabernet Sauvignon hap1 reference. Protein-coding sequences of the Cabernet Sauvignon hap1 SDR were then aligned to the other ten genome assemblies to identify orthologous regions. Next, we manually annotated the SDR of each haplotype in each genome and assigned a sex type (M, F or H) based on known genotypes (Fig. 1c) and known H vs. F haplotype structure²². By these means, we resolved a dataset of 20 *Vitis* SDR haplotypes (5 H, 12 F, and 3 M haplotypes) and the M and F haplotypes of *M. rotundifolia* (Fig. 1c).

All SDR haplotypes were aligned to the Cabernet Sauvignon hap1 H haplotype to assess structural differences and to identify sex-specific features (Fig. 2; Supplementary Fig. 1). There were obvious differences in length among the sexes. Overall, the haplotypes ranged in size from ~171.6 to ~837.4 kbp (Fig. 2a–c), but F haplotypes (181.4 ± 10.2 kbp) were significantly shorter than M (425.9 ± 274.6 kbp; Mann–Whitney test, P value = 0.0008403) and H (289.2 ± 7.4 kbp; P value = 0.0002334) haplotypes. These length differences reflect the presence of sex-linked structural variants (SVs; >50 bp). For example, the 12 *Vitis* F haplotypes shared eight large deletions relative to the H haplotype of Cabernet Sauvignon, encompassing a total length of 117.4 kbp (Fig. 2b). Most (62.9%) of these F-linked deletions were composed of transposable elements (TE), including LTR, Gypsy, Copia and MuDR elements. Similarly, the three M haplotypes shared two large SVs relative to the H reference, including a 22.6 kbp insertion at position 4,802,134 and a 30 kbp deletion from positions 5,021,983 to 5,052,079. In addition, the *V. arizonica* M haplotype had a unique ~400 kbp insertion, and the M haplotype of *M. rotundifolia* contained an inversion that encompassed 57% of the SDR (Fig. 2a). The SDR haplotypes varied in gene content among sexes. While all 20 *Vitis* SDR haplotypes shared 13 SDR genes, two SVs altered gene content. As previously observed in Zhou et al.²², F haplotypes had an SV relative to H and *Vv sylvestris* M haplotypes that deleted two genes encoding TPR-containing proteins. In addition, a FLAVIN-CONTAINING MONOOXYGENASE (*FMO*) gene was absent from H and M haplotypes relative to F haplotypes. The overall structure and gene content were well-conserved between *Vitis* and *M. rotundifolia* species. Despite a large inversion, gene content and order in the M haplotype of *M. rotundifolia* was similar to *Vitis* M haplotypes, and the F haplotype of *M. rotundifolia* was identical in gene content and order to *Vitis* F haplotypes (Fig. 2d). Finally, H haplotypes were similar in structure to F haplotypes within the

first 60 kbp of the SDR, but they were more similar in structure to M haplotypes downstream of this region (Fig. 2d).

Sex-linked polymorphisms affect protein sequences. We used our sequence alignments to the Cabernet Sauvignon H haplotype to identify SNPs that associate perfectly with sex among *Vitis* spp. All of the F- and M-associated polymorphisms were found from positions 4,801,876 to 5,061,548 on chromosome 2 of the Cabernet Sauvignon hap1 reference (Fig. 3a; Supplementary Data 2; Supplementary Fig. 2), which further confirms and delimits the SDR^{10,18}. In total, 1275 SNPs were shared by all 12 *Vitis* F haplotypes vs. H and M haplotypes, and 270 SNPs were shared by all three M haplotypes vs. F and H haplotypes. Interestingly, M-linked SNPs were densest in the first 8 kbp of the SDR (176 SNPs, from 4,801,876 to 4,809,592), and the first F-linked SNP was ~40 kbp downstream from the dense cluster of M-linked SNPs (4,842,196; Fig. 3a). Sex-specific SNP distributions were largely consistent when including *M. rotundifolia* haplotypes in the comparison, though the number of sex-specific SNPs decreased due to divergence between the two genera (Supplementary Fig. 3).

Many of the sex-linked SNPs altered amino acids (Fig. 3b). Altogether, we found six M-linked nonsynonymous SNPs: two in the YABBY transcription factor (TF)-coding gene *VviYABBY3*²⁷, two in an aldolase-coding gene, one in a gene encoding a trehalose-6-phosphate phosphatase (*TPP*), and one in the third *FMO* gene. These nonsynonymous M-linked SNPs represent potential female-sterility mutations. Similarly, we detected 89 nonsynonymous F-specific SNPs across ten genes (Supplementary Data 2). These included one in *TPP*, one in *VviINP1*, seven in an exostosin-coding gene, three in a 3-ketoacyl-acyl carrier protein synthase III gene (*KASIII*), seven in a PLATZ TF-coding gene (*PLATZ*), 18 in the first *FMO* gene, 26 in the second *FMO*, 11 in the third *FMO*, 11 in the hypothetical protein *VviFSEX*, and four in *VviAPT3*. Three of these SNPs introduce a premature stop codon in the first two of four *FMO* genes (Fig. 3e).

To better understand the history of the SDR and to further identify male-sterility and female-sterility candidate genes, we constructed phylogenies from *Vitis* sequences for each SDR gene (Fig. 3f; Supplementary Fig. 4). Alleles tended to cluster by sex type across most of the SDR with the pattern of clustering varying along the locus (Fig. 3f; Supplementary Fig. 4). The phylogenies of four genes at the beginning of the region (from *VviYABBY3* to the aldolase gene) clustered most M sequences apart from F and H sequence, with the *VviYABBY3* and aldolase alleles forming clades that separated M from F and H orthologs (Fig. 3f). This pattern switched from *TPP* onward; for *TPP*, *VviINP1*, *exostosin*, *KASIII*, *PLATZ*, the three *FMO*, the hypothetical protein gene *VviFSEX*, and *VviAPT3*, F sequences clustered apart from M and H alleles (Fig. 3f). These phylogenies are consistent with the observed clusters of sex-specific polymorphisms (Fig. 3a, b), with F-like H haplotypes at the beginning of the region and M-like H haplotypes towards the end of the region (Fig. 2). Genes at the edges of the region do not cluster haplotypes by sex type, further supporting our inference of SDR boundaries (Fig. 3f; Supplementary Fig. 4). Together, these observations are consistent with the emergence of H haplotypes via a recombination event near the aldolase and *TPP* genes, where the pattern of sex-specific clustering shifts (Fig. 3e).

An INDEL in *VviINP1* is conserved in all female haplotypes. In addition to SNPs, we identified 156 and 25 small INDELS (≤ 50 bp) shared by all F and M haplotypes, respectively, relative to the Cabernet Sauvignon H reference. Only two of these INDELS were within exons (Supplementary Data 3): (i) a 21 bp INDEL in the

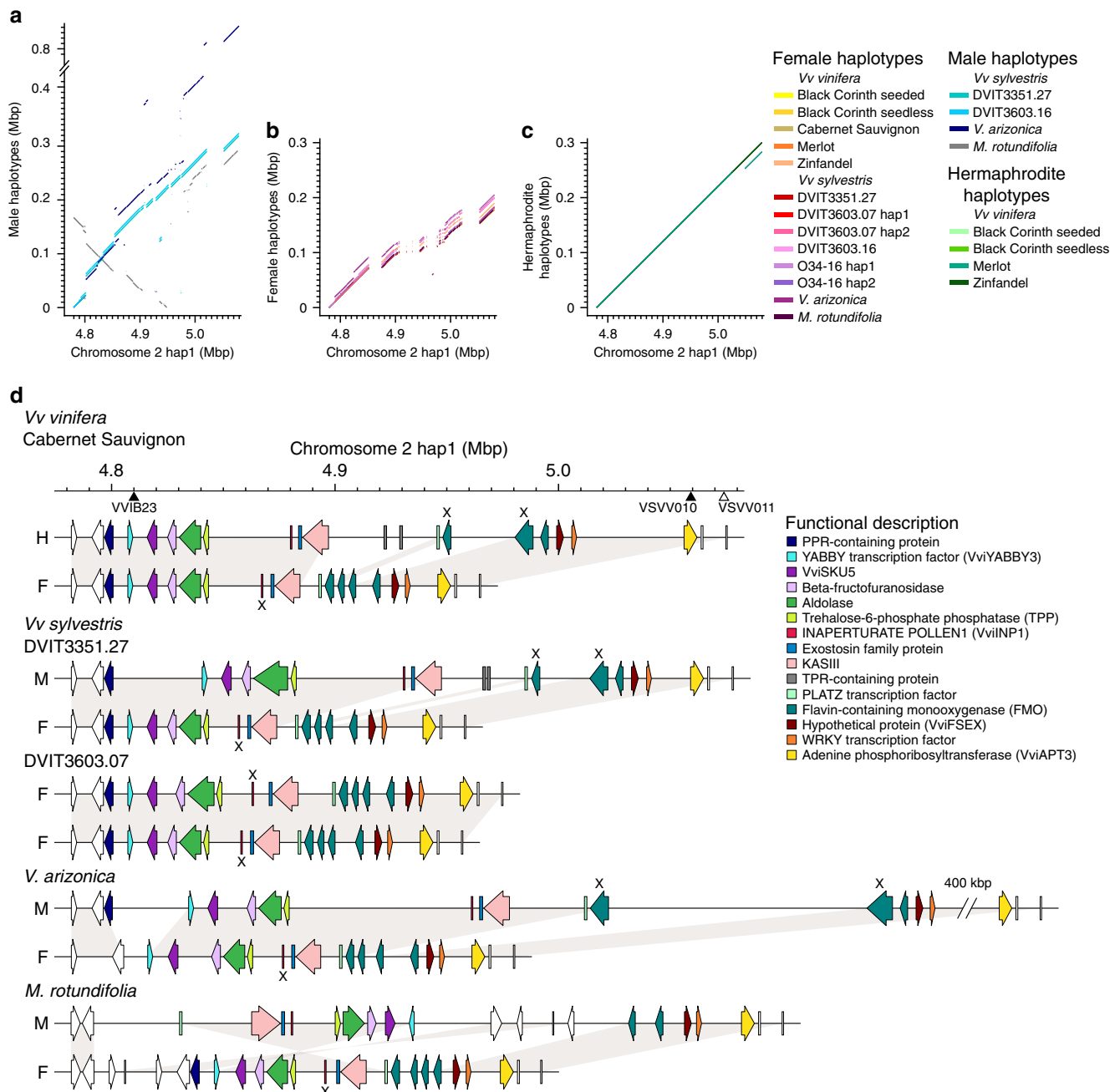


Fig. 2 Sex-linked structural variants and their impact on gene content. Whole-sequence alignments of the sex-determining region (SDR) in M (a), F (b), and H (c) haplotypes against the *Vv vinifera* Cabernet Sauvignon chromosome 2 hap1 (H) reference. The figures illustrate that M haplotypes (a) tend to be longer than either F (b) or H (c) haplotypes and there is a large inversion in the M haplotype of *M. rotundifolia*. **d** Schematic representations of the SDR in four of the 11 genomes analyzed for this study. From top to bottom, the figure illustrates the haplotypes of hermaphrodite *Vv vinifera* Cabernet Sauvignon (HF), male *Vv sylvestris* DVIT3351.27 (MF), female *Vv sylvestris* DVIT3603.07 (FF), male *V. arizonica* (MF), and male *M. rotundifolia* (MF). Each haplotype is annotated with arrows and rectangles to depict annotated genes. White arrows depict genes that are not sex-linked. Genes affected by nonsense mutations are indicated with an X. The scale above the haplotypes denotes the position on the Cabernet Sauvignon reference. The filled, black triangles on this scale mark the position of the sex-linked genetic markers VVIB23¹⁷ and VSVV010¹⁰; the white-filled triangle represents the amplicon VSVV011¹⁰, which is not linked to the SDR. Source data are provided as a Source Data file.

first FMO-coding gene, which introduced a premature stop codon in H and M alleles, and (ii) an 8 bp INDEL in the grape ortholog of the *A. thaliana* *INP1* (AT4G22600), which caused a frameshift and premature stop codon in all F alleles (Fig. 4a). The 8 bp INDEL in F *VviINP1* alleles leads to the truncation of the protein to 41 amino acids (228 amino acids in H/M haplotypes). Because the same 8 bp deletion was also found in the F haplotype of *M. rotundifolia* (Fig. 4a), this F-specific INDEL likely occurred before

Vitis and *Muscadinia* diverged and is therefore likely to be widespread in *Vitis*. Phylogenetic analysis of the INP1 protein clustered *Vitis* and *Muscadinia* orthologs by sex, confirming the sex-specificity of *INP1* alleles (Fig. 4b; Supplementary Fig. 5).

The sex-specificity of *INP1* alleles provided an opportunity to estimate the divergence date between M and F haplotypes and hence the potential age of dioecy. We calculated the average synonymous distance (dS) between all 52 pairs of F and M alleles

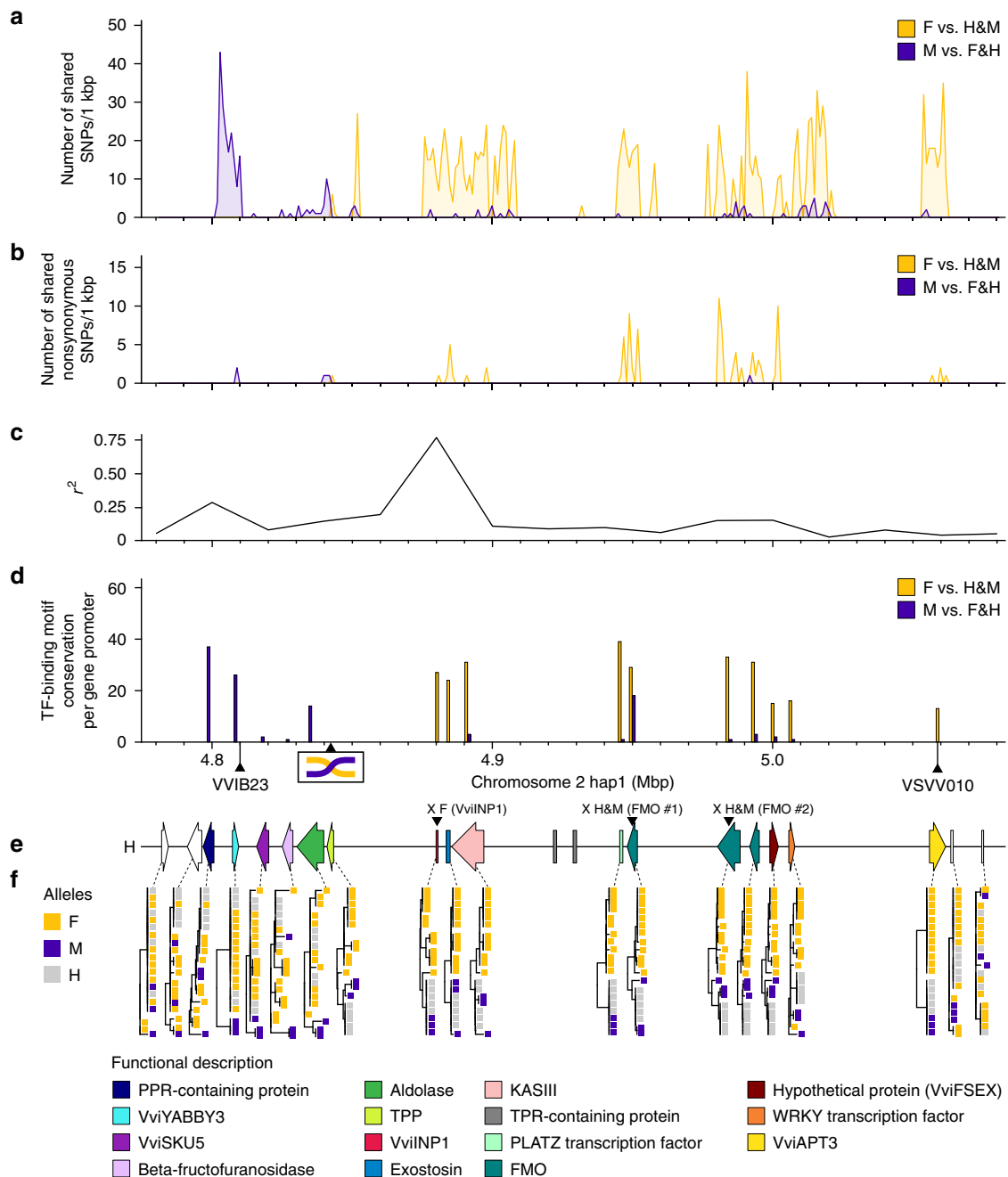


Fig. 3 Sex-linked polymorphisms along the sex-determining region in *Vitis* spp. The number of all (a) and nonsynonymous sex-linked SNPs (b) per kbp across the sex-determining region (SDR). SNPs were identified by aligning all *Vitis* haplotypes to *Vv vinifera* Cabernet Sauvignon hap1 (H). Only SNPs strictly (100%) linked to one sex type were plotted. c Linkage disequilibrium (LD) across the region represented as the median of the squared correlation coefficients (r^2) between all pairs of SNPs calculated within 20 kbp windows. d TF-binding motif conservation per gene promoter detected in the SDR. The x-axis denotes the location on the Cabernet Sauvignon hap1 (H) SDR, and the two black triangles along this axis mark the position of the genetic markers VVIB23¹⁷ and VSVV010¹⁰ that are closely linked to the SDR. The potential position of the recombination event responsible for the reversion to hermaphroditism in domesticated *Vv vinifera* is also indicated. e Gene composition of the H haplotype of *Vv vinifera* Cabernet Sauvignon hap1. Genes are colored as in Fig. 2d and white-colored genes are not sex-linked. Genes affected by nonsense mutations are indicated with an X followed by the affected haplotype and the gene name in parentheses. f Neighbor-joining clustering of the protein sequences encoded by each gene of the SDR. Yellow, purple and gray colors represent F, M and H haplotypes, respectively. Source data underlying Figs. 3a–d, f are provided as a Source Data file.

of *INP1* to be 0.0275 substitutions per base (95% confidence interval: 0.0258–0.0292). Assuming a generation time of 3 years and a nucleotide substitution rate of 2.5×10^{-9} substitutions per base per year²⁸, we infer that M and F alleles diverged ~16.5 million years ago (95% confidence interval: 15.5–17.5), a value

within the range of uncertainty of the estimated split time between *Vitis* and *Muscadinia*²⁹.

To investigate the sex-specificity of the *VviINP1* INDEL further, PCR primers were designed to amplify INDEL-specific alleles of the gene. Seven additional *Vv vinifera* (HF and HH

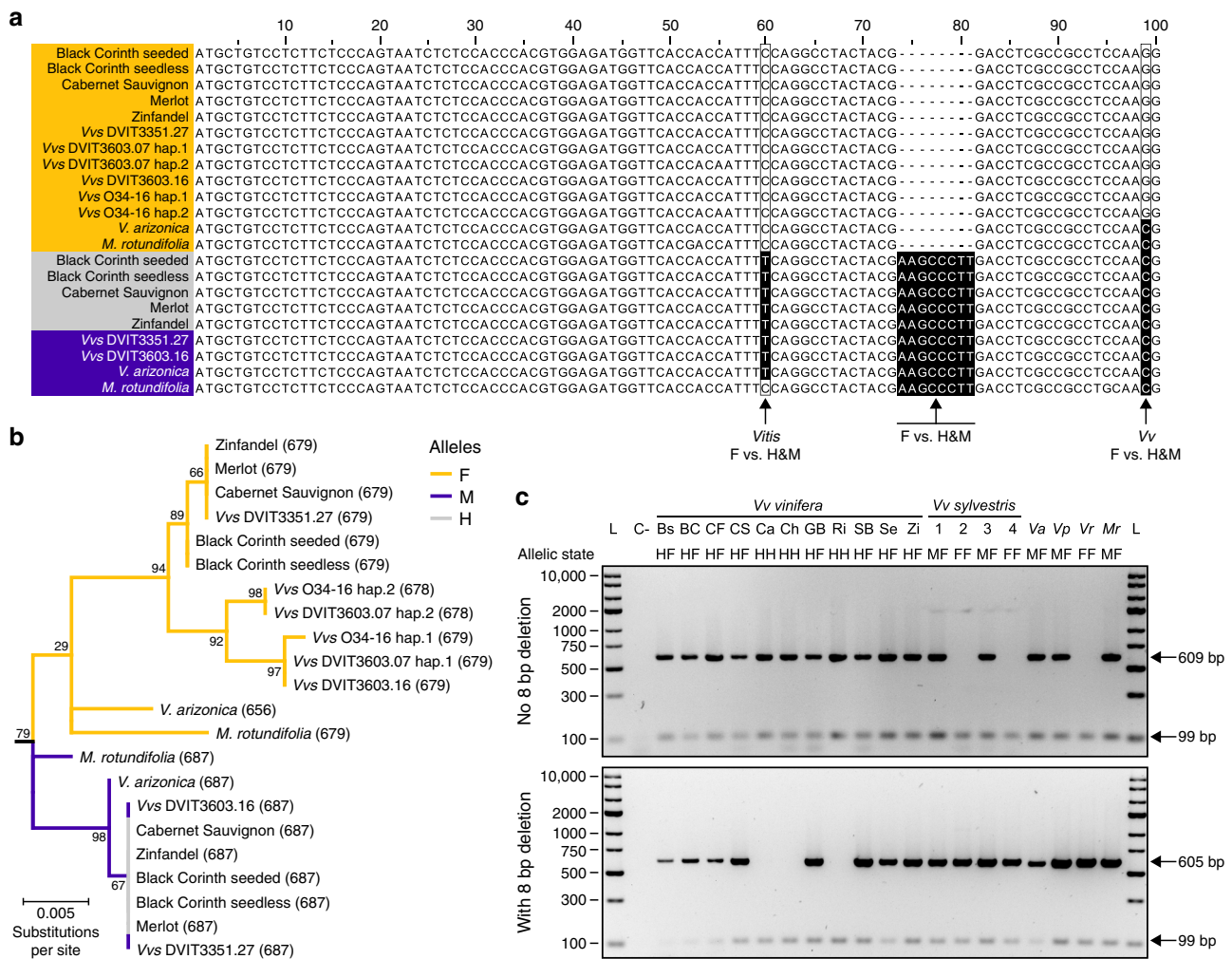


Fig. 4 Mutations and segregation in *VviINP1*. **a** Alignment of the first 100 bp of 20 *VviINP1* coding sequences representing 12 F, 5 H and 3 M haplotypes along with two *M. rotundifolia INP1* coding sequences from F and M haplotypes. Alignment revealed an F-linked 8 bp INDEL in *VviINP1* throughout *Vitis* and shared with *M. rotundifolia*. Yellow, purple and gray colors represent F, M and H haplotypes, respectively. **b** Phylogenetic subtree of the *INP1* coding sequences from *Vitis* spp. and *M. rotundifolia*. The tree was rooted with *INP1* sequences from seven outgroups (see “Methods”). Tree branches are colored according to sex-determining region haplotype. Scale bar is in the unit of the number of substitutions per site. Sequence length in bp is indicated in parentheses. **c** Marker assay amplifying *INP1* fragment without (top panel, 609 bp) or with (bottom panel, 605 bp) the 8 bp deletion. Actin was used as a PCR positive control (99 bp fragment). This assay was performed three times. *Vvs Vv sylvestris*, L Ladder, C – negative control, *Vv vinifera*: Bs seeded Black Corinth, BC seedless Black Corinth, CF Cabernet Franc, CS Cabernet Sauvignon, Ca Carménère, Ch Chardonnay, GB Gouais blanc, Ri Riesling, SB Sauvignon blanc, Se Semillon, Zi Zinfandel, *Vv sylvestris*: 1, DVIT3351.27; 2, DVIT3603.07; 3, DVIT3603.16; 4, O34-16; *Va V. arizonica*, *Vp V. piasezkii*, *Vr V. romanetii*, *Mr M. rotundifolia*. Source data are provided as a Source Data file.

genotypes) accessions and two more distant central Asian *Vitis* species, *V. piasezkii* (male; MF) and *V. romanetii* (female; FF), were analyzed. The presumably functional *VviINP1* allele was present only in genotypes that contained H and M haplotypes, and the INDEL was present only in genotypes containing an F haplotype (Fig. 4c). The same PCR markers were used to screen two F₁ populations. Among the 218 individuals of the *Vv vinifera* × *V. arizonica* F₁ population, 102 individuals were males, 100 females and 16 did not produce any inflorescences, while the *Vv vinifera* × *Vv sylvestris* F₁ population was composed of 92 males, 78 females and 8 plants without inflorescences. In both F₁ populations, sex trait segregation followed a 1:1 ratio, as expected from an FF × MF cross ($\chi^2 = 0.02$, d.f. = 1, *P* value = 0.890; $\chi^2 = 1.15$, d.f. = 1, *P* value = 0.283). In both populations, all F₁ individuals with female flowers were homozygous for the presumably nonfunctional *VviINP1* allele. All F₁ individuals with male flowers carried one functional and one nonfunctional copy of *VviINP1* (Supplementary Figs. 6, 7; Supplementary

Data 4–5). The 8 bp deletion in *VviINP1* as well as all other sex-linked polymorphisms in the SDR were confirmed by replicated bulk analysis of 120 individuals of the *Vv vinifera* × *V. arizonica* F₁ population genotyped by whole-genome sequencing (Supplementary Fig. 8). Finally, it is worth noting that *VviINP1* corresponds to a peak of linkage disequilibrium ($r^2 = 0.77$) across 50 *Vv vinifera* (HF) accessions (Fig. 3c), suggesting a suppression of recombination at this locus.

Because a functional copy of *INP1* is necessary for fertile pollen development in *Zea mays*³⁰, these results support the hypothesis that the 8 bp deletion causes male sterility in homozygous (FF) *Vitis* females and could be responsible for the absence of colpi in *Vv sylvestris* female pollen grains^{16,31}. Together, the sequence, phylogenetic, association, and functional evidences suggest that a recessive allele of *VviINP1* containing an 8 bp deletion interrupts male function, making *VviINP1* a plausible male-sterility candidate.

Sex-linked genes have distinct expression patterns. In order to assess the potential impact of sex-linked polymorphisms on the regulation of SDR genes, we searched for sex-linked TF-binding sites within 3 kbp regions upstream of transcription start sites (Fig. 3d; Supplementary Fig. 9; Supplementary Data 6). M-linked TF-binding motifs were identified upstream of the genes encoding the PPR-containing protein, *VviYABBY3*, the aldolase, *KASIII*, FMOs and *VviFSEX*. Two of the M-linked TF-binding motifs in the promoter region of *VviYABBY3* were associated with flowering and flower development, including SHORT VEGETATIVE PHASE (SVP), which is involved in the control of flowering time by temperature³², and BES1-INTERACTING MYC-LIKE1 (BIM1), a brassinosteroid-signaling component involved in *A. thaliana* male fertility³³. Similarly, we identified TF-binding motifs unique to F haplotypes upstream of *VviINP1*, *exostosin*, *KASIII*, *PLATZ*, FMOs, *VviFSEX*, *WRKY*, and *VviAPT3* (Fig. 3d; Supplementary Data 6). In contrast, all F haplotypes lacked TF-binding sites for bHLH TFs and AGAMOUS-LIKE 3 near the promoter region of *VviAPT3* (Supplementary Fig. 10a).

One gene, *WRKY*, was especially interesting with respect to sex-specific TF-binding sites, because of their potential functional implications. In H and M haplotypes, the *WRKY* promoter region had eleven TF-binding motifs that were absent in all F alleles. These TF-binding sites were associated with TFs that affect flowering time and development in *A. thaliana*³⁴. The remarkable diversity of sex-specific TF-binding motifs suggests the potential for complex regulation of SDR genes by TFs that are located outside the SDR and that could be influenced by environmental factors. Moreover, the distribution of TF-binding sites among haplotypes suggest that many SDR genes may be differentially regulated in a sex-specific manner.

To examine sex-specific regulation, we quantified transcript abundance by RNA-sequencing (RNA-seq) of flower buds from hermaphrodite *Vv vinifera* Chardonnay (HH) and male and female *Vv sylvestris* DVIT3351.27 (MF), and O34-16 (FF). Previous studies have analyzed gene expression in the SDR and have compared flowers of different sexes¹⁹. However, these data were sampled from flowers at early developmental stages and may have missed the late steps of sex determination. Accordingly, we sampled flowers at three developmental stages: (i) during the early development of the reproductive structures, (ii) pre-bloom during pollen maturation, and (iii) at anthesis. Sequencing reads were mapped to both haplotypes of the phased Cabernet Sauvignon genome and expression was compared between individuals at each developmental stage. We focused on genes that showed sex-specific gene expression profiles—e.g., genes that were more highly (or lowly) expressed in the female plant compared to the male plant and to the hermaphrodite.

Thirteen genes fit this criterion in at least one of the developmental stages (Fig. 5a; Supplementary Data 7). For example—and to our surprise—*VviINP1* was significantly more highly expressed in pre- and post-bloom female flowers compared to male and hermaphroditic flowers (adjusted P value ≤ 0.05). Similarly, *WRKY* was more highly expressed in male and hermaphroditic flowers than in female flowers at all developmental stages. Three genes were differentially expressed at only one stage: *TPP*, *aldolase* and *beta-fructofuranosidase* (*BFRUCT*) (Fig. 5a). Two genes exhibited enhanced expression at two or more stages in male flowers: *VviAPT3* was more highly expressed in males at all three stages and *VviYABBY3* was more highly expressed at the two last developmental stages. The *VviAPT3* results were consistent with a previous study showing that high *VviAPT3* transcript abundance was specific to carpel primordia of *Vv sylvestris* male flowers, suggesting a role in carpel abortion²⁰ (Supplementary Fig. 10b). Notably, the high expression of *VviAPT3* in male flowers was specific to the H and M

allele and the expression of the F allele was constantly lower across sex types at each developmental stage (Supplementary Fig. 10b, c). This suggests that the expression level or dosage of the *VviAPT3* M and H alleles might influence sex determination. For *VviYABBY3*, we also confirmed that higher expression was specific to the M allele relative to the F allele in male flowers by aligning RNA-seq reads onto the DVIT3351.27 genome (Supplementary Data 8). Given that sequence polymorphisms in *VviYABBY3* are exclusively M-linked (Fig. 3a), and that the gene resides in the portion of the Cabernet Sauvignon H haplotype that resembles an F haplotype (Fig. 2d), it is reasonable to hypothesize that the M allele for *VviYABBY3* could be associated with female sterility.

The differentially expressed SDR genes included TFs and genes associated with hormone signaling that may play significant regulatory roles in coexpression networks³⁵. To assess the relationships between sex-linked genes and other genes throughout the genome that may affect development, we performed a Weighted Gene Co-expression Network Analysis across developmental stages (WGCNA³⁶) (Supplementary Data 9). Six groups of coexpressed genes (6830 genes in total) were positively or negatively correlated with one of the three sex phenotypes ($|\text{Pearson correlation}| > 0.9$, P value $< 8 \times 10^{-11}$; Supplementary Fig. 11). For example, the magenta module of 1176 coexpressed genes (Fig. 5b) was correlated with male sex (Pearson correlation = 0.97; P value = 2×10^{-16}). This module included two genes in the SDR encoding a PPR-containing protein and *VviAPT3*. The module also included genes involved in hormonal signaling, like two uridine diphosphate glycosyltransferases (UGTs) that are orthologous to *AtUGT85A1* and *AtUGT85A3* and could be involved in active cytokinin homeostasis³⁷. *WRKY* expression was central (i.e. highly connected) in the red module (1512 genes; Fig. 5c) that was negatively correlated with female sex (Pearson correlation = -0.92 ; P value = 6×10^{-12}). This module included an ortholog of *A. thaliana* *SEPALLATA1* (*SEPI*; AT5G15800), an MADS-box gene necessary for floral organ development³⁸. Altogether, these data suggest that sex-linked polymorphisms affect the regulation of some SDR genes and also that some of these genes are highly connected within coexpression modules that participate in sex determination and other developmental processes.

Discussion

Dioecy in *Vitis* is interesting because grapevine is one of a few ancestrally dioecious crops that reverted to hermaphroditism during domestication¹¹. Structurally, wild *Vitis* spp. have perfect flowers (Fig. 1a, b), but male and female flowers lack functional pistils and pollen, respectively. As is predicted for most angiosperms, these wild *Vitis* flowers are likely derived from hermaphroditic ancestors, with dioecy resulting from independent, sequential male-sterility and female-sterility mutations³. Under a two-locus model, the first step to dioecy is the evolution of a recessive male-sterility mutation, and the second step is the formation of a dominant female-sterility mutation^{3,39}.

By mapping the *Vitis* SD haplotypes onto an H reference from Cabernet Sauvignon, we discovered distinct patterns of M-linked and F-linked polymorphisms. M-linked polymorphisms occur in the 5' region spanning from the promoter of the PPR-containing protein through to the *TPP* gene. In contrast, F-linked polymorphisms span from *TPP* to *VviAPT3* (Fig. 6a). Under the two-locus model of the origin of dioecy, the dominant female-sterility allele is expected to be unique to M haplotypes and therefore located in the region where M haplotypes differ from F and H haplotypes. This narrows a search for the female-sterility locus around the region where M-linked polymorphisms are elevated

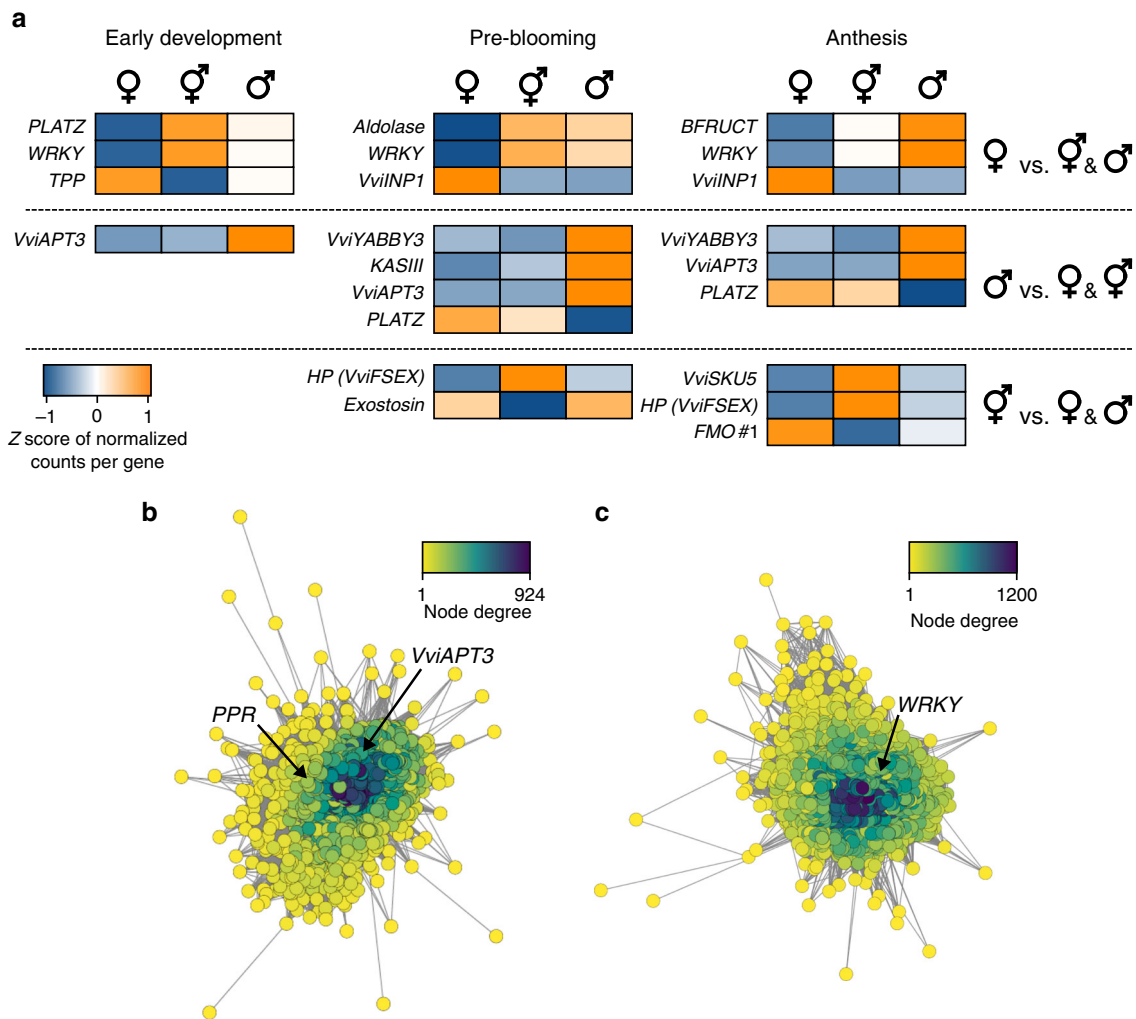


Fig. 5 Transcriptome and gene network analyses. **a** Sex-determining region genes have sex-linked expression at each floral stage. Genes are classified in three groups based on their expression pattern. Only genes differentially expressed in one flower sex compared to the two other sex types are shown. The colors of the heat map depict the Z score of the normalized counts per gene. Gene coexpression networks of the module magenta (**b**), positively correlated with male sex, and red (**c**), negatively correlated with female sex. Node color indicates the degree of connectivity in the network. Positions of the gene encoding a PPR-containing protein, *VviAPT3* and *WRKY* in their corresponding networks are highlighted. HP hypothetical protein. Source data underlying Fig. 5a are provided as a Source Data file.

(Fig. 3a–e). Notably, M-linked SNPs cluster near *VviYABBY3*, and the *VviYABBY3* protein sequence clearly differentiates M from F and H haplotypes (Fig. 3f). In addition, *VviYABBY3* contains two M-specific TF-binding sites and exhibits an M-linked gene expression pattern during flower development (Fig. 5a). We therefore hypothesize that one of the key steps in the transition to dioecy was either caused by the amino acid change in the *VviYABBY3* protein and/or the upregulation of the *VviYABBY3* gene that caused female sterility in males. Functional information about the YABBY gene family is consistent with this hypothesis. In *A. thaliana*, YABBY genes are involved in floral and lateral organ development⁴⁰, specifically the development of carpels and the ovule outer integument⁴¹. Though *VviYABBY3* is not yet characterized, expression of *V. pseudoreticulata* *VpYABBY1* and *VpYABBY2* in *A. thaliana* (and of *VvYABBY4* in tomato) implicates these genes in leaf and carpel development^{27,42}.

Similarly, F-linked polymorphisms define a region of the SDR that is likely to house the hypothesized recessive male-sterility mutation. F-linked polymorphisms were observed in the latter portion of the SDR from *TPP* to *VviAPT3* (Figs. 3 and 6a). Of the

genes in the region, *WRKY* and *VviINP1* are the most noteworthy. The gene *WRKY* is poorly expressed in females relative to males (Fig. 5a) and is part of a coexpression module that is negatively correlated with female sex (Fig. 5c); thus, in theory, it is possible that *WRKY* expression could contribute to male sterility. However, in our view, the 8 bp deletion in the *VviINP1* gene is the most likely cause of male sterility (Fig. 4a). This homozygous deletion causes a frameshift and premature stop codon in the F allele throughout the genus (Fig. 4c), suggesting it has been maintained throughout *Vitis* history. In *A. thaliana* and maize, *INP1* participates in pollen aperture formation^{30,43}; a loss-of-function mutation in grapevine may explain females with sterile inaperturate pollen³¹. All F₁ males contained at least one functional *VviINP1* (Fig. 4c; Supplementary Figs. 6 and 7), suggesting that if the 8 bp deletion in *VviINP1* causes male sterility it is a recessive allele. Why *VviINP1* was more highly expressed in female flowers is not clear (Fig. 5a), especially given our hypothesis that the deletion event makes the F allele of *VviINP1* protein nonfunctional. One possibility is that high expression constitutes a kind of compensatory effect similar to those reported for CRISPR knockouts⁴⁴. Given our hypothesis that the

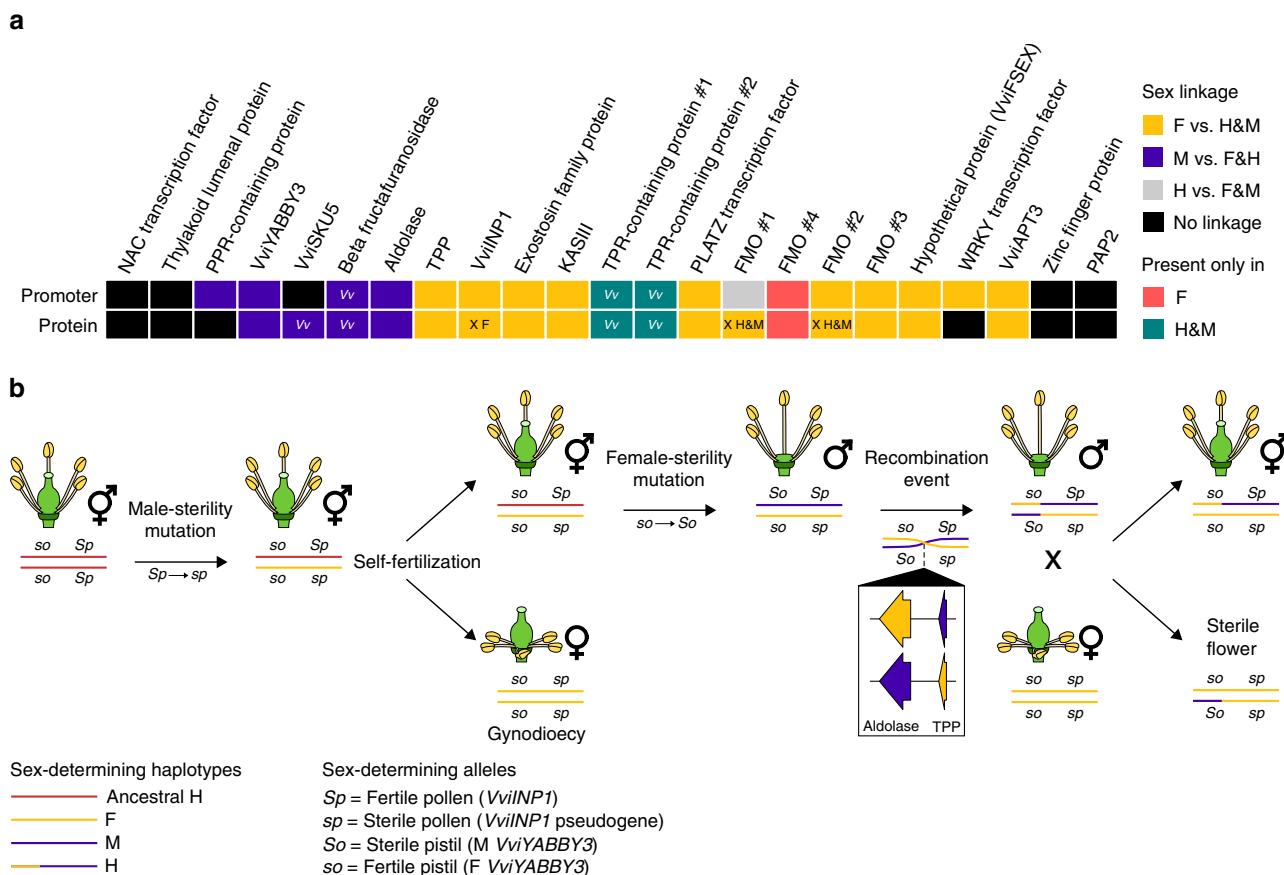


Fig. 6 Model of the evolution of sex determination in grapevine. a A graphical representation of the association between sex and the observed polymorphisms in promoter regions (top row) and encoded proteins (bottom row) of each gene present in the sex-determining region. Genes affected by nonsense mutations are indicated with an X with the affected haplotype. Sex linkage observed only in *Vitis vinifera* species are indicated with a Vv. **b** A potential model for the evolution of dioecy in *Vitis* and its relatives and the reversion to hermaphroditism in cultivated *Vv vinifera*. From left to right, a hermaphroditic ancestor gave rise to male-sterility mutation to produce gynodioecious individuals. Here we denote the recessive male-sterility mutation as *sp*, following the convention of Oberle³⁹. Next, male flowers originated as a consequence of a female-sterility mutation(s), labeled *So*, again following the convention of Oberle³⁹. According to our data, we hypothesize that a rare recombination event occurred in a *Vv sylvestris* male, leading to H haplotype and hermaphrodite individuals in domesticated *Vv vinifera* cultivars. The symbols ♀, ♂ and ♀ represent female, male and hermaphrodite individuals, respectively.

VviINP1 deletion leads to male sterility, an important future step will be functional confirmation that a homozygous deletion engineered into a hermaphrodite produces female flowers.

Like *VviINP1*, *VviAPT3* is in the latter portion of the SDR and exhibits F-linked polymorphism, and so its location suggests a possible role in male sterility. However, *VviAPT3* expression was higher in male flowers and its coexpression module correlated with the male sex. These results are consistent with the gene expression pattern described by Coito et al.²⁰. These data implicate *VviAPT3* in flower development and sex determination but suggest its mechanism of action is complex and requires further study to understand fully.

It has been hypothesized that recombination between M and F haplotypes caused reversion to H^{10,11,45}. Several pieces of evidence support the hypothesis that H haplotypes arose from a recombination event, including their intermediate length and their similarity in structure and sequence to F haplotypes in the first portion of the SDR and to M haplotypes in the latter portion of the SDR (Fig. 6a). Based on our data, we could also localize the hypothetical recombination event. Phylogenetic evidence supports that the recombination event occurred between the *aldolase* and *TPP* genes, and polymorphism information supports that it could have occurred within *TPP* gene, because the gene sequence contained both M- and F-linked nonsynonymous SNPs

(Figs. 3a–f, 6b). We note, however, that there is evidence that H haplotypes originated more than once in domesticated grapevine^{10,45}, suggesting that there could be different recombination breakpoints across H haplotypes.

Finally, we address one more question about recombination: if recombination can occur between F and M alleles, then what has kept the two haplotypes distinct for so long, given that dioecy has been maintained in the wild since the origin of the genus? This is an especially important question given the hypothesis that the rarity of dioecy among angiosperms is due to easy reversion to hermaphroditism⁴⁶. We speculate that recombination between M and F haplotypes is deterred by at least three features of the *Vitis* SDR. The first is that the close linkage of sex-determining genes may simply reflect physical closeness³⁹. If we are correct in our hypotheses that *VviYABBY3* and *VviINP1* are the sterility genes, then recombination events must occur in <100 kbp that separates the two genes to produce an H haplotype. The second is that not all recombination events will be successful in nature: only 50% of correct recombinants will become hermaphrodites (Fig. 6b), and there can be fitness costs associated with hermaphroditism³. Finally, we suspect that differences in the structure and length of M and F haplotypes, which are largely attributable to TE accumulation in intergenic space, limit recombination, because recombination can be slowed by differences in TE content

between alleles⁴⁷. In this context, the inversion in *M. rotundifolia*, which affects 57% of the M haplotype relative to the F haplotype, may be an especially effective deterrent, because inversions can be barriers to recombination^{6,48}. We suspect that these three features contribute to the conspicuous absence of hermaphroditic grapes in the wild.

Methods

Plant material. Different plant tissues were collected from several genotypes for genome sequencing, RNA-seq and marker assay. All plant material was immediately frozen and ground to powder in liquid nitrogen after collection. For genome sequencing, young leaves were collected from three hermaphrodite *Vv vinifera* (Merlot clone FPS 15; Black Corinth with parthenocarpic fruit FPS 02.1; Black Corinth with seeded fruit; Supplementary Fig. 12), four dioecious *Vv sylvestris* (male DVIT3351.27 collected from Armenia; female O34-16 collected from Iran; female DVIT3603.07 and male DVIT3603.16 collected from Azerbaijan), one dioecious male *V. arizonica* (b40-14), and one dioecious male *M. rotundifolia* (Trayshed). For RNA-seq, inflorescences were collected in April and May 2019 from vines at the University of California Davis (Davis, CA, USA). These times correspond to two developmental stages⁴⁹: flowers pressed together (E-L 15) and full flowering with 50% caps off (E-L 23; Supplementary Fig. 13). Three genotypes were sampled, including *Vv vinifera* cv. Chardonnay clone FPS 04, male *Vv sylvestris* DVIT3351.27, and female O34-16. Floral buds were sampled from three inflorescences at E-L 15 (time 1). At E-L 23, pre- (time 2) and post-bloom (time 3) flowers were collected separately, each represented by three biological replicates. Young leaves from additional genotypes were collected for a marker assay. This included seven *Vv vinifera*, *V. piasezkii* and *V. romanetii* and two F₁ populations (Supplementary Table 1). One F₁ population was the result of a cross between the pistillate *Vv vinifera* F2-35 (Carignane × Cabernet Sauvignon) and *V. arizonica* b42-26. The other F₁ population was produced by crossing female *Vv vinifera* 08326-61 (selfing of Cabernet Franc) and *Vv sylvestris* DVIT3351.27. The sexes of F₁ individuals were evaluated by flower morphology and the presence of fruit.

Library preparation and sequencing. High molecular weight genomic DNA (gDNA) was isolated as in Chin et al.²⁴ from each of the nine previously described samples. DNA purity, quantity, and integrity were evaluated with a Nanodrop 2000 spectrophotometer (Thermo Scientific, IL, USA), Qubit 2.0 Fluorometer together with the DNA High Sensitivity kit (Life Technologies, CA, USA), and by pulsed-field gel electrophoresis, respectively. SMRTbell libraries were prepared as described in Minio et al.⁵⁰ and sequenced on a PacBio Sequel system using V3 chemistry, and on a PacBio RS II (Pacific Biosciences, CA, USA) using P6-C4 chemistry for *Vv vinifera* cv. Merlot (DNA Technology Core Facility, University of California, Davis). Summary statistics of SMRT sequencing are provided in Supplementary Data 1. DNA extractions from the two F₁ populations were performed using a modified CTAB protocol¹⁷. DNA concentration was measured using Qubit 2.0 Fluorometer. For the bulk segregant analysis, 120 DNA samples from the F₁ population *Vv vinifera* F2-35 × *V. arizonica* b42-26, 60 females and 60 males, were selected to constitute four pools of 30 individuals each, two made of DNA from male individuals and two made of DNA from female individuals. A total of 1 µg DNA per sample was used as the input material. DNA-seq libraries were prepared using the Kapa LTP library prep kit (Kapa Biosystems, MA, USA). Libraries were evaluated for quantity and quality with the High Sensitivity chip on a Bioanalyzer 2100 (Agilent Technologies, CA, USA) and sequenced in paired-end 150 bp reads on an Illumina HiSeq4000. Total RNA were extracted from floral buds as described in Rapiçavoli et al.⁵¹ from each of the three genotypes described above. cDNA libraries were prepared using the Illumina TruSeq RNA sample preparation kit v.2 (Illumina, CA, USA) and sequenced in single-end 100 bp reads on an Illumina HiSeq4000 (Supplementary Data 10). Four samples from the male genotype (three replicates at time 2 and one at time 3) were re-sequenced in paired-end 150 bp reads to improve library quality and depth.

Genome assembly and annotation. Genome assembly of hermaphrodite *Vv vinifera* cv. Merlot was performed at DNAnexus (Mountain View, CA, USA) (Supplementary Method 1). For each of the other eight genotypes, SMRT reads were assembled with a custom FALCON-Unzip pipeline⁵² reported in <https://github.com/andreaminio/FalconUnzip-DClab>. Repetitive regions were marked using the DAMasker TANmask and REPmask modules⁵³ before and after SMRT read error correction. Then, error-corrected reads were assembled with FALCON v.2017.06.28-18.01²⁴. This included setting different minimum seed-read lengths (length_cutoff parameter) to improve the contiguity of the primary assembly (Supplementary Data 1). Haplotype phasing was carried out using FALCON-Unzip. FALCON-Unzip was designed to combine single-nucleotide polymorphisms and structural variants to separate long sequencing reads based on their haplotype, which are then assembled into separate contigs²⁴. FALCON-Unzip was shown to successfully phase heterozygous regions in plants, including grapes^{24,52}. FALCON-Unzip was performed with default settings. Primary contigs and haplotypes were both polished with Arrow from ConsensusCore2 v.3.0.0. To further improve sequence contiguity, primary contigs of Cabernet Sauvignon²⁴ and the

other nine assemblies were scaffolded using SSPACE-Longread v.1.1⁵⁴ and gaps were closed with PBJelly from PBsuite v.15.8.4⁵⁵. Summary statistics of the ten genome assemblies is provided in Supplementary Data 1. Additional scaffolding and gap-closing steps were performed on the Cabernet Sauvignon genome assembly to construct pseudomolecules (see details in Supplementary Method 2). For all new nine assemblies, genome annotation was performed as described previously for *Vv vinifera* cv. Zinfandel²⁶. Details are provided in Supplementary Method 3.

SDR localization and haplotype reconstruction. The grape SDR was identified by aligning the SSR marker VVIB23¹⁷ and genes previously associated with the SDR¹⁰ (Supplementary Table 2) to the chromosome-scaled Cabernet Sauvignon genome assembly. Protein-coding sequences (CDS) of Cabernet Sauvignon hap1 in this genomic region were then aligned to the ten other genome assemblies with GMAP v.2015-09-29⁵⁶ to identify homologous regions. When the alignments of SDR-associated sequences were fragmented (i.e. with genes aligned to multiple contigs), BLAT v.36x2⁵⁷ was used to determine the overlap between sequences and contigs were manually joined. Junction gaps of ten bases were added between overlaps. A schematic representation of the haplotypes was made using the Gviz Bioconductor package v.1.20.0⁵⁸.

Transcription factor-binding site analysis. For each haplotype, promoter sequences were extracted for all the genes of the SDR using the R package GenomicFeatures v.1.36.4⁵⁹ with a maximum of 3 kbp upstream regions from the gene transcriptional start sites. TF-binding sites were identified using the R package TFBSTools v.1.22.0⁶⁰ and the JASPAR database (R package JASPAR2018 v.1.1.1⁶¹).

Linkage disequilibrium analysis. Illumina whole-genome resequencing data from 50 *vinifera* accessions from previous studies^{22,28} (mean depth = 21.6×) were trimmed using Trimmomatic v.0.36⁶² to remove adaptor sequences and bases for which average quality per base dropped below 20 in 4 bp window. Trimmed paired-end reads were aligned onto Cabernet Sauvignon primary pseudomolecules (hap1) using Minimap2 v.2.17⁶³. Alignments with a mapping quality <10 were removed using Samtools v.1.9⁶⁴. PCR duplicates introduced during library preparation were filtered in MarkDuplicates in the picard-tools v.1.119 (<https://github.com/broadinstitute/picard>). INDEL realignment was performed using Realigner-TargetCreator and IndelRealigner (GATK v.3.5⁶⁵). Sequence variants were called using HaplotypeCaller (GATK v.4.1.2.0⁶⁵). SNPs were filtered using VariationFilter (GATK v.4.1.2.0⁶⁵), according to the following criteria: variant quality (QD) > 2.0, quality score (QUAL) > 40.0, mapping quality (MQ) > 30.0, and <80% missing genotypes across all samples. Linkage disequilibrium measured as the correlation coefficient of the frequencies (r^2) were calculated using Plink v.2.0⁶⁶ for pairwise SNPs with a minor allele frequency (MAF) > 0.05 across the whole SDR (chr2:4750000..5100000).

Whole-sequence alignments and structural variation analysis. Pairwise alignments of all the haplotypes were performed using NUCmer from MUMmer v.4.0.0⁶⁷ and the --mum option using Cabernet Sauvignon hap1 as a reference. Structural variants (SVs; >50 bp) including deletions, insertions, duplications, inversions, translocations, and complex insertion-deletions (CIDs) and short INDELS (<50 bp) were called using show-diff and show-snps, respectively. SNPs were called using show-snps from MUMmer v.4.0.0 and the -f 1 filter. Comparison between haplotypes for SVs and SNPs was performed using multiinter from BEDTools v.2.19.1⁶⁸. Polymorphisms were considered as fully sex-linked only if they were strictly fixed in one sex haplotype compared to the other two sex haplotypes. SNPs and INDELS were confirmed by manually inspecting the alignments of each genotype whole-genome short- and long-read sequences onto their corresponding genomes. Alignments were visualized using Integrative Genomics Viewer (IGV) v.2.4.14 and phasing the two haplotypes⁶⁹.

Phylogenetic analysis. Phylogenetic analysis of the 11 genomes was based on orthology inference (see Supplementary Method 4). Phylogenetic analysis of the proteins and promoter regions of genes in the SDR were conducted with MEGA7⁷⁰ using the Neighbor-Joining method⁷¹ and 1000 replicates. Evolutionary distances were computed using the Poisson correction method⁷² and are expressed as the number of amino acids or base substitutions per site. All positions with less than 5% site coverage were eliminated. Phylogenetic analysis of the *INP1* coding sequences from *Vitis* spp. and *M. rotundifolia* was performed with seven outgroups: three Brassicaceae: *Matthiola incana*, *A. thaliana*, *Capsella rubella*, *Solanum lycopersicum* (Solanaceae), *Eschscholzia californica* (Papaveraceae), and two Poaceae, *Zea mays* and *Brachypodium distachyon*³⁰ (Supplementary Fig. 5), using the Maximum Likelihood method based on the Tamura-Nei model⁷³ and 1000 iterations in MEGA7⁷⁰.

All 52 possible pair of sequences were constructed between the F and M alleles of *INP1* coding sequences in *Vitis* and *M. rotundifolia*, based on an alignment of 687 nucleotides. The average synonymous distance (dS) was computed for each sequence pair using the yn00 program in the PAML package v.4.9⁷⁴. The resulting distribution of dS values was not significantly different from a normal distribution

using a Shapiro test ($W = 0.95765$, P value = 0.06189) and hence normality was assumed to construct 95% confidence intervals.

Transcript abundance quantification. Illumina reads were trimmed using Trimmomatic v.0.36⁶² and the following settings: LEADING:3 TRAILING:3 SLIDINGWINDOW:10:20 CROP:100 MINLEN:36. For the four samples with paired-end reads, only the first mate was used and samples were randomly down-sampled to 26 million reads with seqtk v.1.0-r57-dirty (<https://github.com/lh3/seqtk>) before trimming. These reads were then mapped onto Cabernet Sauvignon phased genome and *Vv sylvestris* DVIT3351.27 genome using HISAT2 v.2.0.5⁷⁵ with the following settings: --end-to-end --sensitive --no-unal -q -t --non-deterministic -k 1 (Supplementary Data 10). The Bioconductor package GenomicAlignments v.1.12.2⁵⁹ was used to count reads per gene locus. Read-count normalization and statistical testing of differential expression were performed using DESeq2 v.1.16.1⁷⁶. Coexpression analysis was conducted with WGCNA v.1.66³⁶ using log₂-transformed normalized counts (log₂(normalized counts + 1)). A soft-thresholding power of 12 with a scale-free model fitting index $R^2 > 0.92$ was applied with a minimum module size of 30. The weighted network was converted into an unweighted network preserving all connections with topological overlap mapping metric (TOM) > 0.02, and imported into Cytoscape v.3.7.2⁷⁷.

Marker assay. Three primers (two forward, one reverse) were designed to determine the presence or absence of an 8 bp deletion in *VviINP1*: INP1_HM-F, 5'-CCTACTACGAAGCCCTTGACC-3'; INP1_F-F, 5'-CAGGCCTACTACGGAC CTC-3'; INP1-R, 5'-CCGTGCAGCGTTC AATTCAGT-3'. *Actin*-specific primers⁷⁸ were used to provide a positive amplification control. The PCR conditions included a 3 min denaturation at 95 °C, followed by 35 cycles of 35 s at 95 °C, 30 s at 60 °C and 40 s at 72 °C, and a final extension for 5 min at 72 °C.

Bulk segregant analysis. Illumina reads from the four pools (mean depth = 172×) were trimmed using Trimmomatic v.0.36⁶² and the following settings: LEADING:7 TRAILING:7 SLIDINGWINDOW:10:20 MINLEN:36. Trimmed paired-end reads were aligned onto Cabernet Sauvignon hap1 genome using BWA v.0.7.17-r1188⁷⁹ with default parameters. PCR and optical duplicates were removed with picard-tools v.2.0.1 (<http://broadinstitute.github.io/picard/>). Sequence variants were called using UnifiedGenotyper from GATK v.3.5-0-g36282e4⁶⁵. Variants were filtered according to the following criteria: read depth (DP) > 5 and quality score (QUAL) > 40.

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

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this Article is available as a Supplementary Information file. The datasets generated and analyzed during the current study are available from the corresponding authors upon request. Sequencing data are accessible through NCBI under the BioProject ID PRJNA593045. Genome sequences and gene annotation files are available at <https://doi.org/10.5281/zenodo.3827985>. The source data underlying Figs. 2, 3a–d, f, 4a, b, and 5a, as well as Supplementary Figs. 1–5 and 8–10 are provided as a Source Data file. Source data are provided with this paper.

Code availability

The script HaploSync is available at Github (<https://github.com/andreaminio/HaploSync>). Source data are provided with this paper.

Received: 6 January 2020; Accepted: 19 May 2020;

Published online: 09 June 2020

References

- Westergaard, M. The mechanism of sex determination in dioecious flowering plants. *Adv. Genet.* **9**, 217–281 (1958).
- Charlesworth, D. in *Evolution—essays in honour of John Maynard Smith* 237–268 (Cambridge University Press, Cambridge, 1985).
- Charlesworth, B. & Charlesworth, D. A. Model for the evolution of dioecy and gynodioecy. *Am. Nat.* **112**, 975–997 (1978).
- Charlesworth, D. Plant sex chromosomes. *Annu. Rev. Plant Biol.* **67**, 397–420 (2016).
- Bull, J. J. *Evolution of Sex Determining Mechanism* (Benjamin/Cummings, Menlo Park, 1983).
- Wang, J. et al. Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc. Natl. Acad. Sci. USA* **109**, 13710–13715 (2012).
- Spigler, R. B., Lewers, K. S., Main, D. S. & Ashman, T. L. Genetic mapping of sex determination in a wild strawberry, *Fragaria virginiana*, reveals earliest form of sex chromosome. *Heredity* **101**, 507–517 (2008).
- Fujita, N. et al. Narrowing down the mapping of plant sex-determination regions using new Y chromosome-specific markers and heavy-ion beam irradiation induced Y deletion mutants in *Silene latifolia*. *G3* **2**, 271–278 (2011).
- Akagi, T. et al. Two Y-chromosome-encoded genes determine sex in kiwifruit. *Nat. Plants* **5**, 801–809 (2019).
- Picq, S. et al. A small XY chromosomal region explains sex determination in wild dioecious *V. vinifera* and the reversal to hermaphroditism in domesticated grapevines. *BMC Plant Biol.* **14**, 229 (2014).
- Henry, I. M., Akagi, T., Tao, R. & Comai, L. One hundred ways to invent the sexes: theoretical and observed paths to dioecy in plants. *Annu. Rev. Plant Biol.* **69**, 553–575 (2018).
- Harkess, A. et al. Sex determination by two Y-linked genes in garden asparagus. *Plant Cell* <https://doi.org/10.1105/tpc.19.0085> (2020).
- Tsugama, D. et al. A putative MYB35 ortholog is a candidate for the sex-determining genes in *Asparagus officinalis*. *Sci. Rep.* **7**, 41497 (2017).
- Moore, M. O. Classification and systematics of eastern North American *Vitis* L. (Vitaceae) north of Mexico. *Sida* **14**, 345 (1991).
- This, P., Lacombe, T. & Thomas, M. R. Historical origins and genetic diversity of wine grapes. *Trends Genet.* **22**, 511–519 (2006).
- Gallardo, A., Ocete, R., Angeles López, M., Lara, M. & Rivera, D. Assessment of pollen dimorphism in populations of *Vitis vinifera* L. subsp. *sylvestris* (Gmelin) Hegi in Spain. *Vitis* **48**, 59–62 (2009).
- Riaz, S., Krivanek, A. F., Xu, K. & Walker, M. A. Refined mapping of the Pierce's disease resistance locus, Pdr1, and sex on an extended genetic map of *Vitis rupestris* × *V. arizonica*. *Theor. Appl. Genet.* **113**, 1317–1329 (2006).
- Fechter, I. et al. Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. *Mol. Genet. Genomics* **287**, 247–259 (2012).
- Ramos, M. J. et al. Flower development and sex specification in wild grapevine. *BMC Genomics* **15**, 1095 (2014).
- Coito, J. L. et al. VviAPRT3 and VviFSEX: two genes involved in sex specification able to distinguish different flower types in *Vitis*. *Front. Plant Sci.* **8**, 98 (2017).
- Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
- Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
- Liu, X. Q., Ickert-Bond, S. M., Nie, J. L., Chen, Q. L. & Wen, J. Phylogeny of the Ampelocissus-Vitis Glade in Vitaceae supports the New World origin of the grape genus. *Mol. Phylogenet. Evol.* **95**, 217–228 (2016).
- Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
- Minio, A., Lin, J., Gaut, B. S. & Cantu, D. How single molecule real-time sequencing and haplotype phasing have enabled reference grade diploid genome assembly of wine grapes. *Front. Plant Sci.* **8**, 826 (2017).
- Vondras, A. M. et al. The genomic diversification of grapevine clones. *BMC Genomics* **20**, 972 (2019).
- Zhang, S. et al. Genome-wide analysis of the YABBY gene family in grapevine and functional characterization of VvYABBY4. *Front. Plant Sci.* **10**, 1207 (2019).
- Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D. & Gaut, B. S. Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. *Proc. Natl. Acad. Sci. USA* **114**, 11715–11720 (2017).
- Wan, Y. et al. A phylogenetic analysis of the grape genus (*Vitis* L.) reveals broad reticulation and concurrent diversification during neogene and quaternary climate change. *BMC Evol. Biol.* **13**, 141 (2013).
- Li, P. et al. INP1 involvement in pollen aperture formation is evolutionarily conserved and may require species-specific partners. *J. Exp. Bot.* **69**, 983–996 (2018).
- Caporali, E., Spada, A., Marziani, G., Failla, O. & Scienza, A. The arrest of development of abortive reproductive organs in the unisexual flower of *Vitis vinifera* ssp. *silvestris*. *Sex. Plant Reprod.* **15**, 291–300 (2003).
- Lee, J. H. et al. Role of SVP in the control of flowering time by ambient temperature in Arabidopsis. *Genes Dev.* **21**, 397–402 (2007).
- Xing, S. et al. SPL8 acts together with the brassinosteroid-signaling component BIM1 in controlling *Arabidopsis thaliana* male fertility. *Plants (Basel)* **2**, 416–428 (2013).
- Shim, J. S., Kubota, A. & Imaizumi, T. Circadian clock and photoperiodic flowering in Arabidopsis: CONSTANS is a hub for signal integration. *Plant Physiol.* **173**, 5–15 (2017).
- Serin, E. A., Nijveen, H., Hilhorst, H. W. & Ligterink, W. Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.* **7**, 444 (2016).
- Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).

37. Cucinotta, M. et al. CUP-SHAPED COTYLEDON1 (CUC1) and CUC2 regulate cytokinin homeostasis to determine ovule number in *Arabidopsis*. *J. Exp. Bot.* **69**, 5169–5176 (2018).
38. Pelaz, S., Ditta, G. S., Baumann, E., Wisman, E. & Yanofsky, M. F. B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature* **405**, 200–203 (2000).
39. Oberle, G. O. A genetic study of variations in floral morphology and function in cultivated forms of *Vitis*. *N. Y. Agric. Exp. Sta. Tech. Bull.* **250**, 1–63 (1938).
40. Chen, Q. et al. The *Arabidopsis* FILAMENTOUS FLOWER gene is required for flower formation. *Development* **126**, 2715–2726 (1999).
41. Villanueva, J. M. et al. INNER NO OUTER regulates abaxial–adaxial patterning in *Arabidopsis* ovules. *Genes Dev.* **13**, 3160 (1999).
42. Xiang, J. et al. Isolation and characterization of two VpYABBY genes from wild Chinese *Vitis pseudoreticulata*. *Protoplasma* **250**, 1315–1325 (2013).
43. Dobritsa, A. A. & Coerper, D. The novel plant protein INAPERTURATE POLLEN1 marks distinct cellular domains and controls formation of apertures in the *Arabidopsis* pollen exine. *Plant Cell* **24**, 4452–4464 (2012).
44. Smits, A. H. et al. Biological plasticity rescues target activity in CRISPR knock outs. *Nat. Methods* **16**, 1087–1093 (2019).
45. Zhou, Y., Muyle, A. & Gaut, B. S. in *The Grape Genome* 39–55 (Springer, Cham, 2019).
46. Käfer, J., Marais, G. A. & Pannell, J. R. On the rarity of dioecy in flowering plants. *Mol. Ecol.* **26**, 1225–1241 (2017).
47. He, L. & Dooner, H. K. Haplotype structure strongly affects recombination in a maize genetic interval polymorphic for Helitron and retrotransposon insertions. *Proc. Natl. Acad. Sci. USA* **106**, 8410–8416 (2009).
48. Lemaitre, C. et al. Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome Biol. Evol.* **1**, 56–66 (2009).
49. Coombe, B. G. Growth stages of the grapevine: adoption of a system for identifying grapevine growth stages. *Aust. J. Grape Wine Res.* **1**, 104–110 (1995).
50. Minio, A. et al. Iso-Seq allows genome-independent transcriptome profiling of grape berry development. *G3* **9**, 755–767 (2019).
51. Rapicavoli, J. N. et al. Lipopolysaccharide O-antigen delays plant innate immune recognition of *Xylella fastidiosa*. *Nat. Commun.* **9**, 390 (2018).
52. Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A. & Cantu, D. Diploid genome assembly of the wine grape Carménère. *G3* **9**, 1331–1337 (2019).
53. Myers, G. Efficient local alignment discovery amongst noisy long reads. In (eds Brown, D. and Morgenstern, B.) *International Workshop on Algorithms in Bioinformatics* 52–67 (Springer, Berlin, 2014).
54. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinform.* **15**, 211 (2014).
55. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
56. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
57. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
58. Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and Bioconductor. *Methods Mol. Biol.* **1418**, 335–351 (2016).
59. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
60. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555–1556 (2016).
61. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).
62. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
63. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
64. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
66. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
67. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
68. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinform.* **47**, 11.12.1–11.12.34 (2014).
69. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
70. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
71. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
72. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* 97–166 (Academic Press, New York, 1965).
73. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).
74. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
75. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
76. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
77. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
78. Licausi, F. et al. Genomic and transcriptomic analysis of the AP2/ERF superfamily in *Vitis vinifera*. *BMC Genomics* **11**, 719 (2010).
79. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

Acknowledgements

This work was funded by the NSF grant #1741627 to B.S.G. and D.C., and partially supported by funds to D.C. from the E.&J. Gallo Winery, J. Lohr Vineyards and Wines, the Chilean Economic Development Agency (CORFO; Project 13CEI2-21852), Viña San Pedro, Viña Concha y Toro, and the Louis P. Martini Endowment in Viticulture. A. Muyle is supported by HFSP Fellowship LT000496/2018-L.

Author contributions

M.M. and D.C. designed the project. M.M., J.L., S.R. and R.F.-B. collected samples. A. Minio assembled all phased genomes. M.M., N.C., J.L., J.F.G., S.R. and R.F.-B. performed the phenotyping. R.F.-B. extracted DNA and RNA and prepared sequencing libraries. M.M. and J.L. performed the marker assay. M.D. performed optical mapping. M.M., N.C., A.Minio, A.M.V., J.F.G. and Y.Z. performed data analyses. M.M., N.C., A.Muyle, A.M.V., B.S.G., and D.C. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-16700-z>.

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

Peer review information *Nature Communications* thanks Deborah Charlesworth, Fabrizio Grassi, Margarida Rocheta and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

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



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

© The Author(s) 2020