Origin of cultivated grapevine inferred from genomes of a global cohort

Yang Dong[1,2,42], Shengchang Duan[1,2,42], Qiuju Xia[3,42], Zhenchang Liang[4,42], Xiao Dong[1,2,42,†], Kristina Margaryan[5,6,42], Mirza Musayev[7,42], Svitlana Goryslavets[8,42], Goran Zdunić[9,42], Pierre-François Bert[10,42], Thierry Lacombe[11,42], Erika Maul[12,42], Peter Nick[13,42], Kakha Bitskinashvili[14,42], György Dénes Bisztray[15,42], Elyashiv Drori[16,17,42], Gabriella De Lorenzis[18,42], Jorge Cunha[19,20,42], Carmen Florentina Popescu[21,42], Rosa A Arroyo-Garcia[22,42], Claire Arnold[23,42], Ali Ergül[24,42], Yifan Zhu[1,42], Chao Ma[25,42], Shufen Wang[1,2], Siqi Liu[1,2], Liu Tang[1,2], Chunping Wang[1,2], Dawei Li[1,2], Yunbing Pan[1,2], Jingxian Li[1,2], Ling Yang[1,2], Xuzhen Li[1,2], Guisheng Xiang[1,2], Zijiang Yang[1,2], Baozheng Chen[1,2], Zhanwu Dai[4], Yi Wang[4], Arsen Arakelyan[5,26,27], Varis Kuliyev[28], Gennady Spotar[8], Nabil Girollet[10], Serge Delrot[10], Nathalie Ollat[10], Patrice This[11], Cécile Marchal[29], Gautier Sarah[11], Valérie Laucou[11], Roberto Bacilieri[11], Pingyin Guan[13], Andreas Jung[30], Michael Riemann[13], Levan Ujmajuridze[14], Tekle Zakalashvili[14], David Maghradze[14], Maria Höhn[15], Gizella Jahnke[15], Erzsébet Kiss[15], Tamás Deák[15], Oshrit Rahimi[16], Sariel Hübner[31], Fabrizio Grassi[32], Francesco Mercati[33], Francesco Sunseri[34], José Eiras-Dias[19,20], Anamaria Mirabela Dumitru[21], David Carrasco[22], Alberto Rodriguez-Izquierdo[22], Gregorio Muñoz[35], Tamer Uysal[36], Cengiz Özer[36], Kemal Kazan[37], Meilong Xu[38], Yunyue Wang[1], Shusheng Zhu[1], Jiang Lu[39], Maoxiang Zhao[25], Lei Wang[25], Songtao Jiu[25], Ying Zhang[40], Lei Sun[40], Huanming Yang[41], Shaohua Li[4], Youyong Zhu[1], Shiping Wang[25,*], Jun Sheng[1,2,*], Wei Chen[1,2,*]

1. State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan Agricultural University, Kunming 650201, China.
2. Yunnan Research Institute for Local Plateau Agriculture and Industry, Kunming 650201, China.
3. State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China.
4. Beijing Key Laboratory of Grape Science and Oenology and Key Laboratory of Plant Resources, Institute of Botany, the Chinese Academy of Sciences, Beijing 100093, China.
5. Institute of Molecular Biology, NAS RA, Yerevan 0014, Armenia.
6. Yerevan State University, Yerevan 0014, Armenia.

7. Genetic Resources Institute, Azerbaijan National Academy of Sciences, Baku, AZ1106, Azerbaijan.
8. National Institute of Viticulture and Winemaking ‘Magarach’, Yalta 298600, Crimea.
9. Institute for Adriatic Crops and Karst Reclamation, Split 21000, Croatia.
10. Bordeaux University, Bordeaux Sciences Agro, INRAE, UMR EGFV, ISVV, Villenave d'ornon 33882, France.
11. AGAP, University of Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier 34398, France.
12. Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Grapevine Breeding Geilweilerhof, Siebeldingen 76833, Germany.
13. Botanical Institute, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany.
14. LEPL Scientific Research Centre of Agriculture, Tbilisi 0159, Georgia.
15. Hungarian University of Agriculture and Life Sciences (MATE), Budapest 1118, Hungary.
16. Department of Chemical Engineering, Ariel University, Ariel 40700, Israel.
17. Eastern Regional R&D Center, Ariel 40700, Israel.
18. Department of Agricultural and Environmental Sciences, University of Milano, Milano 20133, Italy.
19. Instituto Nacional de Investigação Agrária e Veterinária, I.P./INIAV-Dois Portos, Torres Vedras 2565-191, Portugal.
20. Green-it Unit, Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras 2780-157, Portugal.
21. National Research and Development Institute for Biotechnology in Horticulture, Stefanesti, Arges 117715, Romania.
22. Centre for Plant Biotechnology and Genomics, UPM-INIA/CSIC, Pozuelo de Alarcon, Madrid 28223, Spain.
23. University of Lausanne, Lausanne 1015, Switzerland.
24. Biotechnology Institute, Ankara University, Ankara 06135, Türkiye.
25. Department of Plant Science, School of Agriculture and Biology, Shanghai JiaoTong University, Shanghai 200240, China.
26. Armenian Bioinformatics Institute, Yerevan 0014, Armenia.
27. Biomedicine and Pharmacy, RAU, Yerevan 0051, Armenia.

28. Institute of Bioresources, Nakhchivan Branch of the Azerbaijan National Academy of Sciences, Nakhchivan AZ7000, Azerbaijan.
29. Vassal-Montpellier Grapevine Biological Resources Centre, INRAE, Marseillan-plage 34340, France.
30. Historische Rebsorten-Sammlung, Rebschule (K39), Gundheim 67599, Germany.
31. Galilee Research Institute (Migal), Tel-Hai Academic College, Upper Galilee 12210, Israel.
32. Department of Biotechnology and Bioscience, University of Milano-Bicocca, Milano 20126, Italy.
33. Institute of Biosciences and Bioresources, National Research Council, Palermo 90129, Italy.
34. Department AGRARIA, University Mediterranea of Reggio Calabria, Reggio Calabria 89122, Italy.
35. IMIDRA, Alcalá de Henares, Madrid 28805, Spain.
36. Viticulture Research Institute, Ministry of Agriculture and Forestry, Tekirdağ 59200, Türkiye.
37. Commonwealth Scientific and Industrial Research Organization (CSIRO) Agriculture and Food, Queensland Bioscience Precinct, St. Lucia, Queensland 4067, Australia.
38. Institute of Horticulture, Ningxia Academy of Agricultural and Forestry Sciences, Yinchuan 750002, China.
39. Centre for Viticulture and Oenology, School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai 200240, China.
40. Zhengzhou Fruit Research Institutes, CAAS, Zhengzhou 450009, China.
41. BGI-Shenzhen, Shenzhen 518083, China.
42. These authors contributed equally to this work.
† Current Address: Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne 50829, Germany.
* Corresponding authors. Shiping Wang, fruit@sjtu.edu.cn; Jun Sheng, shengjun@dongyang-lab.org; Wei Chen, wchenntr@gmail.com.

Summary (199/200 words)

Grapevine cultivation connects deeply with human agricultural history[1,2], but the origin and dispersal of cultivated grapevine and its relationship with wild progenitor remain contentious[3]. Here we report genome-wide variations of 2,503 cultivated and 1,022 wild accessions of *Vitis vinifera* from all major viticultural regions worldwide. With clearly distinguished wild ecotypes and cultivated grapevine genetic ancestries, we provide evidence for a dual origin of cultivated grapevine in the Near East and Caucasus about 11,000 years ago, thereby endorsing a concurrent origin of table and wine grapes. Subsequent dispersal led to a broad distribution for the Near East domesticates but a limited distribution for the Caucasus domesticates. We reveal that, as the Near East domesticates entered Europe via Anatolia, an ancient wild western ecotype introgression (~10.5 kya) assisted in the creation of muscat grape and various western wine grapevine groups. We find that unique grapevine ancestries were already established by the end of Neolithic (~6.9 kya) and that the process matched early inception of agriculture across Eurasia. Lastly, we show that major grapevine evolutionary events correspond to world climate change. Overall, the defined history of cultivated grapevines is a testament to early human migration and the development of various Eurasian civilizations.

Main Text (3839/4,300 words)

Cultivated grapevine (*V. vinifera* ssp. *vinifera*, hereafter *V. vinifera*) with its unmatched cultivar diversity has been heralded as an emblem of cultural identity in major Eurasian civilizations[2,3]. As a food source (table and raisin grapes) and wine-making ingredient (wine grapes), *V. vinifera* has been sharing a close relationship with the human race since the beginning of agriculture[1,4]. This connection prompted intensive research in ampelography, archaeobotany, and historical records to reveal its past history[5]. Preliminary findings contend that *V. vinifera* originated from its wild progenitor *V. vinifera* ssp. *sylvestris* (hereafter *V. sylvestris*) about 8,000 years ago (ya) during the Neolithic agricultural revolution in the Near East[4,6]. In recent years, this proposition received further exploration from various genetic studies[6–13], but key discoveries on the finer details of grapevine domestication were often inconsistent. For instance, some studies argued the existence of domestication centres outside the Near East (e.g.,

western Mediterranean[13], Caucasus[12,14], and Central Asia[12]), which in turn casts doubt on the popular notion of a single past domestication event[10,11]. Additionally, three demographic inferences yielded population split times between *V. vinifera* and *V. sylvestris* at around 15 Kya to 400 Kya, which markedly predate the historical consensus on domestication time[7–9]. As early domesticates spread to other parts of Eurasia via poorly defined migration routes in the ensuing millennia[4], the single-origin theory also brings a debate on the origin order between table and wine grapevines. The popular view proposes a wine grapevine-first model with two types diverging about 2,500 ya[7,10,11]. Moreover, hybridization with local *V. sylvestris* is deemed common in the creation of extant European wine grapes[10,11], but it is not known whether these introgression events occurred early or late in history. Several studies suggest that the earliest cultivation of European wine grapes in France and Iberia postdates 3000 ya[10,15]. Since the abovementioned discrepancies and unknowns result in large part from the inadequate sampling of grapevine accessions and the limited resolution of genetic data in previous analyses, we report here the genomic variation dataset from a global cohort to systematically delineate the structure of *V. sylvestris* and *V. vinifera* genetic diversity, explore the origin of *V. vinifera*, deduce a putative dispersal history, and investigate key domestication traits and introgression signatures.

Genomic Variation Dataset

In order to attain genomic variations, we constructed a chromosomal-level genome assembly for *V. sylvestris* as reference (VS-1 from Tunisia; Extended Data Fig. 1, Supplementary Note 1, Supplementary Tables 1-9). From the 3,304 assembled accessions, good quality Illumina paired-end sequencing data to a 20-fold average coverage were obtained for 3,186 grapevine accessions (2,237 *V. vinifera* and 949 *V. sylvestris*; Supplementary Table 10-13, Extended Data Fig. 3a; see Methods) from a dozen Eurasian germplasm and private collections. We also included genomic data for 339 previously sequenced accessions (266 *V. vinifera* and 73 *V. sylvestris*; Supplementary Table 14) in the analyses[7,8,16], producing the final cohort of 3,525 grapevine accessions (2,503 *V. vinifera* and 1,022 *V. sylvestris*). The alignment of the Illumina reads to the VS-1 reference genome identifies 45,624,306 biallelic SNPs and 7,314,397 biallelic short Indels (≤40 bp; 73.2% shorter than 5 bp; Supplementary Note

2), among which rare alleles (minor allele frequency ≤ 1%) account for the majority (Extended Data Fig. 2, Supplementary Tables 15-22). The intergenic region of the genome encompasses about 64.7% of SNPs and 70.0% of Indels. About 7.0% of SNPs are located in the coding sequence, and the nonsynonymous to synonymous SNP ratio is 1.497. In comparison, only 2.9% of Indels are found in the coding sequence. We also show that 423,625 SNPs are predicted to be deleterious, and 151,721 Indels to cause frameshift mutations in the coding sequence.

Core accessions by viticultural regions

Clones, mutants, synonyms, and homonyms are a common phenomenon in grapevine germplasm and collections[17]. Knowing how our samples are related is a precondition for the successful analyses of population genomic data. By using the identity-by-state sharing pattern estimators, we found 1,534 accessions sharing the genetic profile with at least one other in the cohort, which belong to 498 distinctive genotypes (Extended Data Fig. 3a-c, Supplementary Note 3, Supplementary Table 23). We kept one accession for each distinctive genotype, corrected misidentified accessions, and excluded interspecific hybrids to obtain a core cohort of 2,448 grapevines (1,604 *V. vinifera* and 844 *V. sylvestris*; Extended Data Fig. 3b). These core accessions remain representative of the major viticultural regions[18] in the world (Extended Data Fig. 3d).

Since geographical indication carries economic and cultural significance for grapevines, major viticultural regions have been the preferred grouping method in the evaluation of grapevine genetic diversity[6,12,19]. For the principal component analysis (PCA), the genetic variation among core accessions shows that *V. sylvestris* and *V. vinifera* separately spread out along the first two principal component (PC) axes, with both displaying a crude Near East to Western Europe gradient (Fig. 1a). The PC-based median positions do not precisely mirror the geographical locations on a map (e.g., unlike that in human[20]). Along the PC3 axis, the differentiation of individuals is in large part based on the *V. vinifera* utilization (Extended Data Fig. 4b). In addition, individuals from the same viticultural region are loosely clustered together with no clear boundaries between different viticultural regions. Likewise, the maximum likelihood phylogenetic tree of core accessions displays two clades, which are mainly based on

grapevine utilization (Extended Data Fig. 5). Particularly, the *V. vinifera* accessions from the same viticultural region do not form a monophyletic subclade, but rather scatter in different places of the subtrees. These results collectively demonstrate the disconnection between the fine viticultural geographic pattern and the genetic structures in grapevine[21]. One explanation could be the extensive exchange of superior cultivars across regions throughout history, as subsequent interbreeding for new cultivars would blur the boundaries of established groups and even out the effect of isolation-by-distance (genetic differentiation).

Grapevine grouping by genetic ancestry

In view of the poor resolution of viticultural regions in defining grapevine diversity, we have leveraged genetic ancestry information from an unsupervised ADMIXTURE analysis to categorize core accessions (Fig. 1b, Extended Data Fig. 6a, Supplementary Note 4). At $K=2$, the *V. sylvestris* accessions display various proportions of the east (red) and west (blue) genetic ancestry components. In contrast, all *V. vinifera* accessions at $K=2$ contain a major east (red) ancestry. This observation indicates that *V. vinifera* was derived from wild progenitors of the east (red) ancestry. At $K=8$, hierarchical clustering of ancestry components identifies four *V. sylvestris* groups, each including accessions from distinct geographic regions: the Near East (Syl-E1), the Caucasus (Syl-E2), Central Europe (Syl-W1), and the Iberian Peninsula (Syl-W2; Fig. 1b). *V. sylvestris* accessions collected from other regions show admixed genetic structures (Supplementary Note 4). For cultivated grapevines, six genetic ancestries could designate six distinctive groups (CG1 to CG6), all covering a broad range of viticultural regions (Fig. 1b, Supplementary Note 4). We examined accessions in each group with pure or close to pure ancestries (Extended Data Fig. 6b, 6e, Supplementary Note 4), and ascribed names accordingly to these groups as Near East table grapevines (CG1), Caucasian wine grapevines (CG2), muscat grapevines (CG3), Balkan wine grapevines (CG4), Iberian wine grapevines (CG5), and Western European wine grapevines (CG6). The admixed *V. vinifera* accessions showed different combinations of genetic ancestries (Extended Data Fig. 6c, 6d). In the end, the four *V. sylvestris* and six *V. vinifera* groups could be clearly differentiated in the PCA plots (Fig. 1c, Extended Data Fig. 4c), thus suitable for population genomic investigations.

V. sylvestris diversity and past history

V. sylvestris natural habitats are partitioned by the Mediterranean Sea, the Black Sea, the Alps, and the Zagros Mountain in the western Eurasia continent. According to the genetic ancestries and the occupied corresponding ecological niches, we designate *V. sylvestris* accessions in the Near East and the Caucasus as the eastern ecotype (Syl-E) and accessions in Central Europe and the Iberian Peninsula as the western ecotype (Syl-W; Fig. 2a). This designation is supported by the large between-ecotype fixation index values (e.g., Syl-E1 vs. Syl-W1, F_{ST}=0.340), as opposed to the small within-ecotype fixation index values (Syl-E1 vs. Syl-E2, F_{ST}=0.101; Syl-W1 vs. Syl-W2, F_{ST}=0.072; Extended Data Fig. 7a, Supplementary Table 26). By evaluating nucleotide diversity (π) and individual heterozygosity, we show that the western ecotype (especially Syl-W1) has a significantly lower degree of population polymorphism than its eastern counterpart (Extended Data Fig. 7b, c). Moreover, the linkage disequilibrium decay (LD, r^2) was much slower in Syl-W (1.0-1.6Kb at half of maximum r^2) than in Syl-E (400-600bp at half of maximum r^2; Extended Data Fig. 8). These data demonstrate that the eastern ecotype retains the highest genetic diversity.

Demographic inference shows that the ancient history of *V. sylvestris*, similar to human evolution[22], was influenced by global climate change. Both subgroups of the Syl-E and Syl-W exhibit a remarkable population bottleneck around the time of the Last Glacial Maximum (LGM at ~21 Kya, thousand years ago; ~10-40 Kya for *V. sylvestris* subgroups), with the effective population sizes (N_e) reaching a minimum of 10,000 to 40,000 (Extended Data Fig. 9). In accordance with this result, ecological niche modelling predicts that the areas with suitable environmental conditions for Syl-E and Syl-W (suitability>0.75) became not only limited but also completely separated at the LGM (Fig. 2b). Notably, LGM was associated with a human population bottleneck and later population turnover in Europe[23]. The *V. sylvestris* N_e rebound post LGM was less steep and more prolonged in the Syl-E accessions than in the Syl-W accessions (Extended Data Fig. 9). Nonetheless, the N_e of the Syl-W accessions decreased to lower levels in recent time, which agrees with their reduced genetic diversity.

The stairway plots reveal an additional population bottleneck in all *V. sylvestris* subgroups around 200-600 Kya (Extended Data Fig. 9) during a Pleistocene period characterized by changing climate cycles and hominin expansion[24,25]. This period is congruent with the deduced population split time (median ~200-400 Kya) between Syl-E and Syl-W (Fig. 2a). The slow descent of the split line suggests that the geographic isolation of the two ecotypes was a gradual process (Extended Data Fig. 9). The median population split time between Syl-E1 and Syl-E2 was estimated as ~56 Kya, which corresponds to the modern human migrating out of Africa[26], presumably to escape a dryer climate[27]. In comparison, the median population split time between Syl-W1 and Syl-W2 was abrupt and recent at ~2.5 Kya, when the rise and fall of the Roman Empire was linked to climate fluctuations[28].

Dual origin of *V. vinifera*

The wet climate in Early Holocene (~11.7-8.3 Kya) facilitated the expansion of suitable habitats for both wild ecotypes, with Syl-E enjoying a large geographic span from Central Asia to the Iberian Peninsula (Fig. 2b). This result supports the eastern origin and subsequent continental dispersal of *V. vinifera* (Fig. 3a). Since CG1 shares the main ancestral component with Syl-E1 and CG2 with Syl-E2 (Fig. 1b), we evaluated the possibility of two independent primary domestication events in the past. Indeed, both CG1 and CG2 maintain the highest genetic diversity and manifest the quickest LD decay among all CG groups (Extended Data Fig. 7, 8). They have a lower population differentiation with their corresponding wild ecotypes (Extended Data Fig. 7a). The outgroup f_3 statistics bi-plots also reveal that CG1 and CG2 are genetically closer to Syl-E1 and Syl-E2, respectively (Fig. 3b, Supplementary Table 27). Notably, the population split lines of the CG1/Syl-E2 and CG2/Syl-E1 pairs resemble that of the Syl-E1/Syl-E2 and differ from those of the CG1/Syl-E1 and CG2/Syl-E2 pairs (Fig. 3c, Extended Data Fig. 10). These data collectively support a dual origin of *V. vinifera* and reject the popular theory of a single primary domestication centre[10,11].

The separation of the CG1/Syl-E1 and CG2/Syl-E2 population pairs occurred fairly quickly (Fig. 3c), which is compatible with a clean split scenario. The median population split time is estimated to be ~11 Kya for both pairs, suggesting that the

independent domestication events took place concurrently around the advent of agriculture. As CG1 and CG2 respectively represent table and wine grapevine ancient genetic background (Extended Data Fig. 6e), this analysis also turns down the notion that wine grapevine predates table grapevine[7,10,11].

The geographic distributions of CG1 and CG2 cultivars across Eurasia and North Africa could outline vastly different dissemination routes for the two grapevine groups (Fig. 3a). The CG1 dispersal goes in four directions. The eastward expansion through Central Asia into India and China follows the Inner Asia Mountain Corridor, a path that also witnessed the exchange of other crops (i.e., wheat, barley, and millet) between the West and the East[29]. The northbound expansion showcases the early cultural contact of the Near East over Zagros mountains with the Caucasus[30,31]. The northwest expansion via Anatolia into the Balkan bespeaks the spread of farming into Europe[32,33]. Finally, a westward expansion across North Africa coastlines corroborates with the finding that early Neolithic Moroccans were genetically related to Levantine farmers[34]. In contrast, CG2 individuals were mainly confined to both sides of the Caucasus Mountain, with a limited dispersal route going into the Carpathian Basin by the northern Black Sea. This path implies that CG2 played a negligible role in the formation of wine grapevines in Europe. Altogether, the post-domestication dispersal routes of *V. vinifera* parallel the trails of past human migration.

Selection on sex determination region

In order to reveal domestication signatures in *V. vinifera*, we investigated both Syl-E1/CG1 and Syl-E2/CG2 group pairs by selecting genomic regions that display high levels of nucleotide diversity difference and population differentiation (both top 5%; Extended Data Fig. 11a, Supplementary Table 28). We collated the identified domestication selective sweep regions from the two pairs and found 27 shared ones mainly in the chromosomes 2 and 17 of the VS-1 genome assembly (Supplementary Table 29). In particular, the Chr2:14.28-14.34 Mb selective sweep region overlaps with the grapevine sex determination region (SDR; Extended Data Fig. 11b), which underlies the transition from dioecy in *V. sylvestris* to hermaphroditism in *V. vinifera*[35].

This result corroborates with a previous investigation[8] and confirms that the selection on flower sexual morphs is of great importance during grapevine domestication.

The dioecious grapevine SDR includes the male (M/f) and female (f/f) genotypes, with which independent recombination events at three loci have facilitated the formation of two major hermaphroditic (H1 and H2) haplotypes[35]. The haplotype pairing among f, H1, and H2 yields hermaphroditic H1/f, H2/f, H1/H1, and H1/H2 SDR genotypes, which account for the majority in our *V. vinifera* samples (Extended Data Fig. 11c, Supplementary Table 30). Specifically, the H1/f SDR is universally distributed among all six cultivated grapevine groups, whereas H2/f and H1/H2 SDRs are predominantly found in the Iberian and Western European wine grapevines (CG5 and CG6; Extended Data Fig. 11d). Another interesting finding is the enrichment of homozygous H1/H1 phenotype in CG3 muscat grapevines. The distribution bias of these genotypes implies an independent origin of H2 haplotype and an intensive selection of the muscat grapevines, respectively.

Aside from the known major haplotypes and genotypes, the scale of our grapevine cohort also enables the discovery of accessions containing novel minor haplotypes (male variant Mv, female variant fv, H3, H4, and H5) and genotypes (Mv/f, M/H1, M/H5, H1/fv, H5/f, H4/f, H2/H2, and H2/H3) as a result of recombination events at five different sites in the SDR (Extended Data Fig. 11b, c, and e, Supplementary Table 30). This result not only showcases the SDR diversity in grapevine natural populations, but also assists the construction of a putative recombination history for known SDR haplotypes (Extended Data Fig. 11f). It is clear that a first independent recombination event between the parental M and f haplotypes created Mv (site 4), fv (site 3), H1 (site 2), and H4 (site 1). On this basis, H1 experienced a second independent recombination event with f to produce H3 (site 5) and H5 (site 4), whereas H4 recombined again with f at site 5 to bring about H2. The fact that H4 predates H2 allows us to build a putative evolutionary past of the two haplotypes (Extended Data Fig. 11g). Intriguingly, the origin of H4 can be traced to the Near East in three Syl-E1 *V. sylvestris* accessions (IS164, IS167, and IS180). After human selection, it possibly followed a westward dispersal route to reach the Iberian Peninsula, where it can now be found in an old Iberian

cultivar ‘Malvasia Fina’ (PO153). Given the geographic distribution of H2/f and H1/H2 SDRs, a likely scenario supports that H2 originated from H4 in the Iberian Peninsula and later became dominant during the diversification of Iberian and Western European cultivars.

Syl-W ecotype introgression

The expansion of suitable habitats for Syl-E and Syl-W in early Holocene led to shared areas mainly in the coastal regions of northern Mediterranean and southern Black Sea, the Iberian Peninsula, and an area corresponding to present western France (black area in Fig. 2b). This formed an ecological foundation for the genetic exchange between CG1 and local refugia Syl-W accessions as the early domesticates dispersed into Europe via Anatolia. The pervasive introgression of wild genotypes is well documented in the extant European *V. vinifera* grapevines[10,11], with many old varieties (i.e., ‘Lambrusco’ cultivars) deriving about half of their ancestries from Syl-W (Extended Data Fig. 6d). However, they likely showcase the late diversification effort after the distinct ancestries (CG3-CG6) had been established. To test this, we have chosen cultivars in each group with at least 75% major ancestry (also average Syl-W ancestry in each *V. vinifera* group <3%) to delineate how Syl-W introgression shaped cultivated grapevines. Interestingly, the TreeMix analysis finds one migration edge that points from Syl-W to a population ancestral to CG3-CG6 (estimated weight 0.114; Fig. 4a, Extended Data Fig. 12a), suggesting an ancient introgression event occurred before the diversification of all European grapevines. An additional migration edge also points from Syl-W to CG6 (estimated weight 0.292), which implies that Western European wine grapevines had a unique independent introgression event in the past. This introgression history is supported by various combinations of *D*-statistics testing the gene flow from Syl-W into CG groups (Z-score>3.0, adjusted $P<4.17\times10^{-5}$; Extended Data Fig. 12b, Supplementary Table 31). Additionally, the gene flows from Syl-W into CG3-CG6 inferred from Momi2 all point to their corresponding divergence from CG1, further supporting the introgression history (Extended Data Fig. 12c). Notably, the estimated median divergence times date the creation of Balkan wine grapes (CG4) to 8,070 ya, Iberian wine grapevines (CG5) to 7,740 ya, and Western European wine grapevines to 6,910 ya (Fig. 4b, Extended Data Fig. 12c). These time points accord with the historical

384 migration of Anatolian farmers into Europe[30,33,36,37], which substantiate the role of
385 viticulture in the formation of Neolithic agricultural societies.
386
387 The migration edge weights, f_4-ratio, and Momi2 estimates collectively show that the
388 ancient introgression from Syl-W accounts for about 11.4-18.0% of the CG3-CG6
389 genomes (Extended Data Fig. 12, Supplementary Table 31). On top of this, the
390 independent introgression contributes about 25.0-30.0% additional Syl-W to the CG6
391 ancestry. We have screened the introgression tracts in CG3-CG6 by choosing the
392 genomic windows having the top 1% d_f and f_{dM} values (Extended Data Fig. 13). A total
393 of ten regions are shared among CG3-CG6 groups, which contain genes that are
394 putatively involved in plant immunity, abiotic stress response, and carbohydrate
395 metabolism (Supplementary Table 32). This result agrees with the proposal that
396 introgression helps grapevines adapt to new environment and become more suitable to
397 wine making[10,11].
398
399 **Muscat grapevine**
400 Muscat grapevine (CG3) is unique for its floral aromas, which are the result of a hard-
401 to-define concoction of monoterpenoids in the fruit[38]. Given its broad geographic
402 distribution (Extended Data Fig. 14a) and very old history, it is difficult to pinpoint the
403 centre of origin. Momi2 estimate predicts a population split from CG1 at around 10,564
404 ya (Extended Data Fig. 12c), which would suggest an origination site close to the Near
405 East. This is supported by the relatively low F_{ST} value and a sizeable gene flow with CG1
406 (Extended Data Figs. 7a, 12c), but very few CG3 cultivars could be located in Anatolia
407 and the surrounding regions. One possible reason is the gradual loss of ancient CG3
408 cultivars throughout history, which could explain the low genetic diversity and high LD
409 extent in the CG3 group compared to others (Extended Data Figs. 7b, 8). Even though
410 the muscat aroma is a complex trait, genome-wide association analysis based on a
411 binary differentiation reveals 18 SNP signatures on chromosomes 5 and 18 (Extended
412 Data Fig. 14b, c, Supplementary Table 33). This set includes a nonsynonymous SNP
413 Chr5:19419686 in the *VvDXS* gene that has been linked to the trait[38]. Examination of
414 the genotype at this locus shows that 108 out of the 135 muscat grapevines (including
415 'Muscat Hamburg', 'Königin der Weingärten', and 'Muscat of Alexandria' commonly

used as parental cultivars) are heterozygous (G/T) and only eight individuals are homozygous (T/T) for the alternative SNP. Additionally, CG3 grapevines without muscat aroma are found to be homozygous for the reference SNP (G/G). This result suggests that selection on this allele might have put some constraint on grapevine fecundity, thereby preventing the alternative SNP from reaching fixation.

Discussion

Our systematic genomic survey of *V. sylvestris* and *V. vinifera* accessions paints a defined picture of the grapevine evolutionary history, which echoes key events in the history of world climate change and human migration (Fig. 5). The Pleistocene era witnessed the continuous fragmentation of habitats, the decline of effective population size, and the separation of ecotypes for *V. sylvestris*. It is highly likely that modern humans extensively utilize grapevines for energy source from late Pleistocene on, but the harsh climate at the time was not fit for agriculture[39]. As the climatic conditions ameliorated at the Pleistocene-Holocene transition, grapevine with its fairly stable perennial yield unsurprisingly became one of the earliest candidates for domestication. The diverse SDR haplotypes suggest that an early goal could be the conscious selection[40] and propagation of rare naturally-occurring hermaphroditic individuals from the *V. sylvestris* population, because they allow mass plantation without male plants. The selection on phenotype, but not on genotype, also implies that the different hermaphroditic haplotypes were subject to a strong genetic drift. This is showcased by the high frequency of H1 and almost extinct H4 in extant cultivars. The Mesolithic and Neolithic period also saw the early dispersal and diversification of grapevines where unique ancestries were established in the Balkan, the Iberia, and the Western Europe with the help of *V. sylvestris* introgression into CG1. This event mirrors early farmer migration in Europe, consolidating the role of viticulture in forming sedentary societies. The last stage since the Bronze Age is characterized by a higher level of cultural exchange, thus the trading of superior grapevine cultivars along trade routes. This is especially evident in the plethora of Italian cultivars with three or more genetic ancestries, and unfortunately poses a challenge to disentangle the genealogical history of each grapevine cultivar[21]. Lastly, genetic reliable wild grapevines from Central Asia, a region battered by climate change and social instability for the past few millennia,

are no longer available to test Vavilov's theory for a diversity centre or a hypothetical turnover of grapevine types due to Islam conversion in the region. These questions may be resolved with the help of paleogenomic data in the future.

Methods (2250/3000 words)

VS-1 genome assembly. The *V. sylvestris* plant VS-1 of Tunisian origin (DVIT2426) was obtained from the grape germplasm and breeding block of the Shanghai Jiaotong University in Shanghai. Fresh young leaves were collected for the extraction of total genomic DNA using the CTAB Plant DNA Extraction Kit (Genenode Biotech Co, Beijing). We obtained 49.5Gb (~100×) PacBio single-molecule real-time (SMRT) reads and 26.7Gb (~54×) circular consensus sequencing (CCS) reads on the PacBio RS II platform from BGI-Wuhan (Wuhan, China) and Berry Genomics (Beijing, China), respectively. We also obtained a total of 170.67Gb (~350×) Illumina paired-end sequencing data and 62.44Gb Hi-C sequencing data from Novogene (Beijing, China).

The details of the genome assembly pipeline can be found in Supplementary Note 1. In brief, we generated a basic contig assembly based on the PacbBio SMRT sequences with NextDenovo (v.2.0.beta.1), from which we removed redundancy with a pipeline provided by Purge Haplotigs[41] and polished residual errors with clean Illumina short reads using Pilon[42]. We next assembled the CCS reads using Canu[43] and aligned it to the SMRT contigs with nucmer[44] to achieve longer contigs. After an additional round of redundancy removal with Purge Haplotigs and contig polish with CCS reads using NextPolish[45], we obtained an assembly of 477.80Mb with a contig N50 size of 13.82Mb. The elongated contigs were then anchored into chromosome scale using a Hi-C proximity-based assembly approach[46,47], where 19 high-confidence clusters representing the haploid chromosomes of *V. sylvestris* were identified, covering 95.04% of the whole assembly. We compared our VS-1 genome assembly with published *V. sylvestris* genomes[48,49] and annotated the protein-coding genes for the ensuing analyses. See Supplementary Note 1 for details.

Sample collection and processing. A total of 23 institutions from 16 nations in the world contributed to the global grapevine cohort[17,50–55], which comprised of 2,269 *V.*

vinifera and 1,035 *V. sylvestris* accessions. The *V. vinifera* accessions were collected from institutional germplasms and private collections. The selection was designed to preferentially include old, autochthonous, and economically important varieties to maximize the spectrum of genetic diversity. The *V. sylvestris* accessions were collected from all major refugia in the world, which spans a large geographical area from Levant and Transcaucasia in the east to the Iberian Peninsula in the west[56]. Total genomic DNA was either obtained from dried grapevine leaf tissues using the CTAB Plant DNA Extraction Kit (Genenode Biotech Co, Beijing) in a wet lab at the Yunnan Agricultural University, or directly sent from collaborators. For the latter, genomic DNA was cleaned once by sodium acetate precipitation and reconstituted in nuclease-free water (Ambion, Texas, USA). Sequencing libraries with an insert size of 350~550 bp were prepared with NEBNext® Ultra™ DNA Library Prep Kit (Illumina, USA) according to the manufacturer's directions. Paired-end sequencing was performed on an Illumina NovaSeq 6000 platform by both Novogene (Beijing, China) and Berry Genomics (Beijing, China). The target sequencing depth was 20× for each accession. After excluding unusable sequencing libraries, we curated raw genome data for 3,270 samples (2,256 *V. vinifera* and 1,014 *V. sylvestris*; success rate 99.4%), totaling 33.96 Tb. On top of these, we also included 271 *V. vinifera* accessions and 73 *V. sylvestris* accessions from previous publications in the following steps [7,8,16]. See Supplementary Note 2 for details.

Variant calling and annotation. The raw sequencing reads were processed to obtain sequencing depth, duplication rate, and percentage of mapping rate for each accession. We denoted any value that was outside mean ± 3S.D. of these parameters to be an outlier, and excluded grapevine samples with outlier parameters from variant calling. With this method, we retained 2,237 *V. vinifera* and 949 *V. sylvestris* samples from our collaboration and 266 *vinifera* and 73 *sylvestris* samples from previous publication, making the final grapevine cohort of 3,525 accessions. A single accession of muscadine grape (ZZ-01) was included as outgroup for the downstream analyses[57]. See Supplementary Note 2 for details.

We used the chromosomes of the VS-1 genome (excluding unanchored sequences) as references in the identification of variants (both SNP and Indel). The variant detection was carried out with GATK3 (v.3.8; https://github.com/broadinstitute/gatk) according to the recommended workflow[58]. In brief, the variants of each accession were called using the GATK HaplotypeCaller, and then a joint-genotyping analysis of the gVCFs was performed on all samples (also separately for *V. vinifera* and *V. sylvestris* samples). In the filtering step, various parameters used in the hard filtering of raw SNPs and Indels were determined according to the recommendation of GATK[58]. As a result, the SNP filter expression was set as "QD<2.0, QUAL<30.0, SOR>3.0, FS>60.0, MQ<40.0, MQRankSum<-10.0, ReadPosRankSum<-8.0". The short Indel filter expression was set as "QD<2.0, QUAL<30.0, SOR>5.0, FS>100.0, InbreedingCoeff<-0.8". SNP density, Indel density and total genetic diversity across each chromosome were calculated with 100 kb sliding window using vcftools (V.0.1.16)[59]. Our called SNP datasets were compared to the 10K grapevine SNP chip[60] and the 472 *Vitis* SNP dataset[8], and further validated with somatic SNPs obtained from a group of Chasselas clones (Supplementary Note 2).

We performed SNP and Indel annotation according to the VS-1 genome using the package ANNOVAR (v.2015-12-14)[61], and predicted the effect of nonsynonymous SNPs on the biological function of proteins with Provean (v.1.1.5)[62].

Genetic clonal accessions. We utilized identity-by-state (IBS) sharing pattern estimators[63–65] to infer relationship among accessions. This approach is superior to the identity-by-descent (IBD) inference in our case in that: (1) it does not require prior knowledge of ancestral pedigree or allele frequencies, and (2) it is robust to SNP ascertainment errors[63–65]. We removed SNPs with low read support (<7 reads) or with high linkage disequilibrium (LD, $r^2 \geq 0.5$) with other SNPs for the analyses. The estimators were calculated with SNPduo (V.2.00a)[63]. By using estimator values from known clonal accession pairs as reference, we set the following three cut-off values: R1≥1.20, IBS2*ratio≥0.99, and KING-robust kinship≥0.3426. We would assume a genetic clonal relationship if two of the above thresholds were met between two

accessions. We kept one accession for each distinctive genotype and marked all other clonal accessions for exclusion from analyses.

Phylogenetic tree. The SNPs were processed using SNPhylo (Version 20180901)[66] with default parameters. The resultant phylip format data were taken to construct a ML phylogenetic tree using RAxML-NG (v.0.9.0)[67] with 32 random search trees and 100 TBE bootstraps. The best tree was chosen according to the maximum Final LogLikelihood value. A muscadine grape was included as outgroup.

Principal Component Analysis and ADMIXTURE. We chose the core set of SNPs (MAF greater than 0.05) for additional pruning. PLINK (v1.90b6.12)[68] was used to remove SNPs having high LD ($r^2 \geq 0.5$) within a continuous window of 50 SNPs (step size 5 SNPs), which yielded 2,669,247 SNPs for both analyses. We performed PCA with GCTA (v.1.26.0)[69] using the default settings. The first three principal components were plotted and colored according to major viticultural region, utilization, and genetic groups, respectively. We also examined the genetic ancestry with ADMIXTURE (v.1.3.0)[70] and determined the choice of K using a 5-fold cross-validation (CV) procedure[71]

Grapevine major group characterization. Linkage disequilibrium (pairwise r^2 values) was calculated across all chromosomes using PopLDdecay (v.3.41)[72] with default parameters. The average nucleotide diversity (π) within continuous 100 kb sliding windows, pairwise population fixation index (F_{ST}), and individual heterozygosity were calculated with VCFtools (v.0.1.16)[59].

Ecological niche modelling. We compiled 41 and 16 different geographical records from all identified Syl-W and Syl-E accessions, respectively for the analysis. The raster files of 19 bioclimatic variables at 2.5 minutes resolution for the Last Glacial Maximum (LGM, ca. 21 ka, v1.2b) and early Holocene (EH, Greenlandian, 11.7-8.326 ka, v1.0) paleoclimate data were obtained from PaleoClim[73]. Since removing highly collinear variables has an insignificant impact on maximum entropy model performance[74], we included all original variables in the analysis. The R package ENMeval (v.0.3.1)[75] was

used to test all combinations of defined settings and perform cross validation for model evaluation. For the Syl-W ecotype, the settings of LQH_1, LQ_2.5 were chosen to measure variable importance for the LGM and EH, respectively, whereas for the Syl-E ecotype, the settings of LQ_1.5 and LQ_4 were selected. Then the projections for habitat suitability were generated in MaxEnt (v.3.4.4)[76] from the ENMeval results with the parameters of 10 subsample replicated runs and 30 random test percentage.

Demographic history inference. First, we employed the MSMC2[77] to infer population size and split time. The input files for MSMC2 were generated with MSMC Tools (https://github.com/stschiff/msmc-tools). In brief, bi-allelic SNP sites with uniquely mapped reads and 0.5 to 2-fold mean coverage depths were used in the analyses, and the remaining genomic regions were masked using the script bamCaller.py. Then all segregating sites within each group were phased using SHAPEIT (v.2.r904)[78]. Single population demographic inference was performed on four individuals (eight haplotypes), whereas population split inference was performed on two individuals (four haplotypes) for each group. Only grapevine accessions with the highest proportion of major ancestries (top 50 or major ancestry > 70%) were randomly chosen for the inference. Single population demographic inference was repeated ten times for each group. Median population split times were deduced from the results of 100 random combinations for each comparison. We used a mutation rate of 5.4×10^{-9} per site per generation and a generation time of 3 years for demographic history inference[8], unless stated otherwise.

The stairway plot 2 (v.2.1)[79] was also used for estimating the population demography history for *V. sylvestris* from SNP frequency spectrum. We filtered out SNP sites in the coding sequence region and masked genomic regions of repetitive elements. For each population, we only included accessions with the highest proportion of major ancestries (50 for Syl-W1, 58 for Syl-W1, 54 for Syl-E1, and 34 for Syl-E2). We estimated folded SFS using easySFS (https://github.com/isaacovercast/easySFS). Population history was predicted by ignoring singletons and 200 bootstraps were run to assess confidence intervals. We plotted the change of estimated median effective population size through time and the associated 95% confidence intervals (2.5% and 97.5% percentiles).

We used Momi2 (v.2.1.19)[80] to explore demographic models for various sets of four populations. Five individuals with the highest proportion of major ancestries were included in each population. We filtered out SNP sites in the coding sequence and genomic regions of repetitive elements. The extracted folded site frequency spectrum (SFS) was split into 100 equal-sized blocks for jackknifing and bootstrapping. One gene flow event and constant population size were assumed for a set of four-population comparison. The split times of Syl-W/Syl-E and Syl-E1/CG1 were based on the MSMC2 results, where the interquartile range (25% to 75%) was fed into Momi2. We fitted 20 independent runs with random starting parameters and selected the demographic model with the biggest log-likelihood value of all runs. Then 100 bootstraps for the best model were implemented by resampling blocks of the SFS to generate confidence intervals.

Selective sweep signals. We investigated the selection signals across the whole genome via a cross comparison of the genetic differentiation (F_{ST}) and nucleotide diversity (π). A 50 kb sliding window with 10 kb step approach was applied to quantify F_{ST} and π by using the VCFtools software (v0.1.16)[59]. The candidates that meet both top 5% of the two values were selected as selective signals.

Treemix. We estimated admixture graphs of grapevine groups using TreeMix (v.1.12), which applies a ML method based on a Gaussian model of allele frequency change[81]. For each group, individuals with at least 75% major ancestries (also average Syl-W ancestry in each *V. vinifera* group <3%) were used. SNPs were filtered for missing calls and monomorphism. The topology of the ML trees changes depending on the number of migration edges (*m*) allowed in the model. The optimal number of migration edges was determined from the range of one to ten using a R packages OptM (v.0.1.6)[82]. The TreeMix program was run with "-bootstrap 1000 -k 500". The Syl-E1 group was set as root. For each migration event, we constructed the tree with migration edges 10 times using random seed. The best outcome was determined by the biggest residual value.

***f*-statistics, Patterson's *D*, and local introgression region.** Individuals with at least 75% major ancestries were used for each group. Outgroup f_3 statistics were calculated using a R package admixr (v.0.9.1)[83] for all possible combinations of grapevine groups with *Vitis rotundifolia* as the outgroup. The Patterson's *D* and f_4 admixture ratio for all possible combinations of trios of the grapevine groups were calculated using Dtrios in Dsuite (v. 0.4 r42)[84] with *V. rotundifolia* as the outgroup. SNPs were filtered for missing calls and monomorphism. To further locate the local introgressed genomic regions, the *df* and f_{dM} statistics were calculated along the whole genome using Dinvestigate in Dsuite with a sliding window of 50 SNPs and a step of 5 SNPs. We defined the putative introgressed regions as those among top 1% of both values and visualized these regions with R.

Genome-wide association study. We performed a genome-wide association study on muscat and non-muscat grapevines using fastGWA-GLMM method[85] in GCTA (v.1.93.3beta)[69]. For the binary categorization, the muscat phenotype (n=135, Supplementary Table 1 and 14) was defined as 1 and non-muscat phenotype (n=158) as 0. The non-muscat grapevine were selected from CG1, the earliest domesticates. SNPs with missing calls greater than 0.2 and minor allele frequency less than 0.01 were filtered. We defined the whole-genome significance cut-off with $-\log_{10}(P) = 6$.

Data availability

The VS-1 genome assembly is available at the China National Centre for Bioinformation under the project number PRJCA009324. The raw resequencing data are available at the China National Centre for Bioinformation under the project number PRJCA009314.

Code availability

Details regarding the software packages and versions used in the analyses are included in the Methods and Supplementary Note.

References

1. McGovern, P. E., Hartung, U., Badler, V. R., Glusker, D. L. & Exner, L. J. The beginnings of winemaking and viniculture in the ancient Near East and Egypt. *Expedition* **39**, 3–21 (1997).

2. This, P., Lacombe, T. & Thomas, M. R. Historical origins and genetic diversity of wine grapes. *Trends Genet* **22**, 511–519 (2006).

3. Grassi, F. & Lorenzis, G. D. Back to the Origins: Background and Perspectives of Grapevine Domestication. *Int J Mol Sci* **22**, 4518 (2021).

4. Zohary, D., Hopf, M. & Weiss, E. *Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*. (Oxford University Press, 2012).

5. Cantu, D. & Walker, M. A. *The Grape Genome*. (Springer Nature Switzerland AG, 2019).

6. Myles, S. *et al.* Genetic structure and domestication history of the grape. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3530–3535 (2011).

7. 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. U.S.A.* **114**, 11715–11720 (2017).

8. Liang, Z. *et al.* Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. *Nat Commun* **10**, 1190 (2019).

9. Sivan, A. *et al.* Genomic evidence supports an independent history of Levantine and Eurasian grapevines. *Plants People Planet* **3**, 414–427 (2021).

10. Freitas, S. *et al.* Pervasive hybridization with local wild relatives in Western European grapevine varieties. *Sci Adv* **7**, eabi8584 (2021).

11. Magris, G. *et al.* The genomes of 204 Vitis vinifera accessions reveal the origin of European wine grapes. *Nat Commun* **12**, 7240 (2021).

12. Riaz, S. *et al.* Genetic diversity analysis of cultivated and wild grapevine (Vitis vinifera L.) accessions around the Mediterranean basin and Central Asia. *Bmc Plant Biol* **18**, 137 (2018).

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

14. McGovern, P. *et al.* Early Neolithic wine of Georgia in the South Caucasus. *Proc. Natl. Acad. Sci U.S.A.* **114**, E10309–E10318 (2017).

15. Ramos-Madrigal, J. *et al.* Palaeogenomic insights into the origins of French grapevine diversity. *Nature Plants* **5**, 595–603 (2019).

16. Roach, M. J. *et al.* Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLoS Genet* **14**, e1007807 (2018).

17. Lacombe, T. *et al.* Large-scale parentage analysis in an extended set of grapevine cultivars (Vitis vinifera L.). *Theor Appl Genet* **126**, 401–414 (2013).

18. Bacilieri, R. *et al.* Genetic structure in cultivated grapevines is linked to geography and human selection. *Bmc Plant Biol* **13**, 25–25 (2013).

19. Emanuelli, F. *et al.* Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *Bmc Plant Biol* **13**, 39 (2013).

20. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

21. Mercati, F. *et al.* Integrated Bayesian Approaches Shed Light on the Dissemination Routes of the Eurasian Grapevine Germplasm. *Front Plant Sci* **12**, 692661 (2021).

22. deMenocal, P. B. Climate and Human Evolution. *Science* **331**, 540–542 (2011).

23. Posth, C. *et al.* Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe. *Curr Biol* **26**, 827–833 (2016).

24. Hosfield, R. & Cole, J. Early hominins in north-west Europe: A punctuated long chronology? *Quaternary Sci Rev* **190**, 148–160 (2018).

25. Timmermann, A. *et al.* Climate effects on archaic human habitats and species successions. *Nature* 1–7 (2022) doi:10.1038/s41586-022-04600-9.

26. Eriksson, A. *et al.* Late Pleistocene climate change and the global expansion of anatomically modern humans. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16089–16094 (2012).

27. Schaebitz, F. *et al.* Hydroclimate changes in eastern Africa over the past 200,000 years may have influenced early human dispersal. *Commun Earth Environ* **2**, 123 (2021).

28. Büntgen, U. *et al.* 2500 Years of European Climate Variability and Human Susceptibility. *Science* **331**, 578–582 (2011).

29. Stevens, C. J. *et al.* Between China and South Asia: A Middle Asian corridor of crop dispersal and agricultural innovation in the Bronze Age. *Holocene* **26**, 1541–1555 (2016).

30. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).

31. Wang, C.-C. *et al.* Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nat Commun* **10**, 590 (2019).

32. Pinhasi, R., Fort, J. & Ammerman, A. J. Tracing the Origin and Spread of Agriculture in Europe. *Plos Biol* **3**, e410 (2005).

736 33. Mathieson, I. *et al.* The Genomic History of Southeastern Europe. *Nature* **555**, 197–203
737 (2018).

738 34. Fregel, R. *et al.* Ancient genomes from North Africa evidence prehistoric migrations to
739 the Maghreb from both the Levant and Europe. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6774–6779
740 (2018).

741 35. Zou, C. *et al.* Multiple independent recombinations led to hermaphroditism in grapevine.
742 *Proc. Natl. Acad. Sci. U.S.A.* **118**, (2021).

743 36. Olalde, I. *et al.* The genomic history of the Iberian Peninsula over the past 8000 years.
744 *Science* **363**, 1230–1234 (2019).

745 37. Brunel, S. *et al.* Ancient genomes from present-day France unveil 7,000 years of its
746 demographic history. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 12791–12798 (2020).

747 38. Emanuelli, F. *et al.* A candidate gene association study on muscat flavor in grapevine
748 (Vitis vinifera L.). *Bmc Plant Biol* **10**, 241–241 (2010).

749 39. Richerson, P. J., Boyd, R. & Bettinger, R. L. Was Agriculture Impossible during the
750 Pleistocene but Mandatory during the Holocene? A Climate Change Hypothesis. *Am*
751 *Antiquity* **66**, 387–411 (2001).

752 40. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication
753 and diversification. *Nat Rev Genet* **14**, 840–852 (2013).

754 41. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig
755 reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).

756 42. Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant
757 Detection and Genome Assembly Improvement. *Plos One* **9**, e112963 (2014).

758 43. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer
759 weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

760 44. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *Plos*
761 *Comput Biol* **14**, e1005944 (2018).

762 45. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool
763 for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).

764 46. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution
765 Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).

766 47. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields
767 chromosome-length scaffolds. *Science* **356**, 92–95 (2017).

768 48. Massonnet, M. *et al.* The genetic basis of sex determination in grapes. *Nat Commun* **11**,
769 2902 (2020).

770 49. Badouin, H. *et al.* The wild grape genome sequence provides insights into the transition
771 from dioecy to hermaphroditism during grape domestication. *Genome Biol* **21**, 223 (2020).

772 50. Margaryan, K., Melyan, G., Röckel, F., Töpfer, R. & Maul, E. Genetic Diversity of
773 Armenian Grapevine (Vitis vinifera L.) Germplasm: Molecular Characterization and
774 Parentage Analysis. *Biology* **10**, 1279 (2021).

775 51. Ergül, A., Perez-Rivera, G., Söylemezoğlu, G., Kazan, K. & Arroyo-Garcia, R. Genetic
776 diversity in Anatolian wild grapes (Vitis vinifera subsp. sylvestris) estimated by SSR
777 markers. *Plant Genetic Resour* **9**, 375–383 (2011).

778 52. Zdunić, G. *et al.* Genetic Structure and Relationships among Wild and Cultivated
779 Grapevines from Central Europe and Part of the Western Balkan Peninsula. *Genes-basel* **11**,
780 962 (2020).

781 53. Laucou, V. *et al.* High throughput analysis of grape genetic diversity as a tool for
782 germplasm collection management. *Theor Appl Genet* **122**, 1233–1245 (2011).

783 54. Lózsa, R., Xia, N., Deák, T. & Bisztray, G. D. Chloroplast diversity indicates two
784 independent maternal lineages in cultivated grapevine (Vitis vinifera L. subsp. vinifera).
785 *Genet Resour Crop Ev* **62**, 419–429 (2015).

786 55. Cunha, J. *et al.* Genetic Relationships Among Portuguese Cultivated and Wild Vitis
787 vinifera L. Germplasm. *Front Plant Sci* **11**, 127 (2020).

788 56. Grassi, F., Mattia, F. D., Zecca, G., Sala, F. & Labra, M. Historical isolation and
789 Quaternary range expansion of divergent lineages in wild grapevine. *Biol J Linn Soc* **95**, 611–
790 619 (2008).

791 57. Dong, X. *et al.* VitisGDB: The Multifunctional Database for Grapevine Breeding and
792 Genetics. *Mol Plant* **13**, 1098–1100 (2020).

793 58. Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome
794 Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinform* **43**, 11.10.1-11.10.33 (2013).

795 59. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158
796 (2011).

797 60. Laucou, V. *et al.* Extended diversity analysis of cultivated grapevine Vitis vinifera with
798 10K genome-wide SNPs. *Plos One* **13**, e0192540 (2018).

799 61. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
800 variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

801 62. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of
802 amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).

803 63. Roberson, E. D. O. & Pevsner, J. Visualization of Shared Genomic Regions and Meiotic
804 Recombination in High-Density SNP Data. *Plos One* **4**, e6711 (2009).

64. Waples, R. K., Albrechtsen, A. & Moltke, I. Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Mol Ecol* **28**, 35–48 (2019).

65. Stevens, E. L. *et al.* Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. *Plos Genet* **7**, e1002287 (2011).

66. Lee, T.-H., Guo, H., Wang, X., Kim, C. & Paterson, A. H. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**, 162 (2014).

67. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

68. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).

69. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am J Hum Genetics* **88**, 76–82 (2011).

70. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).

71. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *Bmc Bioinformatics* **12**, 246 (2011).

72. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2018).

73. Brown, J. L., Hill, D. J., Dolan, A. M., Carnaval, A. C. & Haywood, A. M. PaleoClim, high spatial resolution paleoclimate surfaces for global land areas. *Sci Data* **5**, 180254 (2018).

74. Feng, X., Park, D. S., Liang, Y., Pandey, R. & Papeş, M. Collinearity in ecological niche modeling: Confusions and challenges. *Ecol Evol* **9**, 10365–10376 (2019).

75. Muscarella, R. *et al.* ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods Ecol Evol* **5**, 1198–1205 (2014).

76. Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. & Blair, M. E. Opening the black box: an open-source release of Maxent. *Ecography* **40**, 887–893 (2017).

77. Schiffels, S. & Wang, K. MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. in *Statistical Population Genomics* (ed. Dutheil, J. Y.) vol. 2090 147–166 (Humana, 2020).

78. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179–181 (2012).

79. Liu, X. & Fu, Y.-X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol* **21**, 280 (2020).

80. Kamm, J., Terhorst, J., Durbin, R. & Song, Y. S. Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *J Am Stat Assoc* **115**, 1–16 (2019).

81. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *Plos Genet* **8**, e1002967 (2012).

82. Fitak, R. R. OptM: estimating the optimal number of migration edges on population trees using Treemix. *Biol Methods Protoc* **6**, bpab017 (2021).

83. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).

84. Malinsky, M., Matschiner, M. & Svardal, H. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol Ecol Resour* **21**, 584–595 (2021).

85. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data. *Nat Genet* **53**, 1616–1621 (2021).

Acknowledgements

We thank Frédérique Pelsy, Laurence Garmendia Auckenthaler, Anne-Françoise Adam-Blondon, Christèle Cornier, Pál Kozma, Olivier Bachmann, François Gillet, Jean-Michel Gobat, Sandrine Dedet, Joachim Daumann, Kerstin Huber, Valentina Risovannaya, Alla Polulyah, Bordenave Louis, Maria Lafargue, Goutouly Jean-Pascal, Gagik Melyan, Dorin Ioan Sumedrea, and technical staff from EGFV and UEVB for their assistance in the sample collection and laboratory work. We thank Peter Kupfer, Elisha D. O. Roberson, and Desislava Petkova for their comments. Support for this project was provided by the Natural Science Foundation of China (32070599) and Yunnan Agricultural University Research Fund (A2032002519) to W.C.; by China Agriculture Research System of MOF and MARA (CARS-29) to S.W.; by 20APP-4E007, DAAD (57552334), and Alliance of International Science Organization (ANSO-CR-PP-2020-04-A) to K.M.; by Ministerio de Ciencia, Innovación y Universidades and Agencia Estatal de Investigación of Spain (RTI2018-094470-R-C21) to R.A.A.G.; by Predoctoral Fellowship (PRE2019-088446) to A.R.-I.; by Israel Ministry of Science and Technology (90-23-020-12) to E.D.; by Fondation Giacomi and Swiss National Science Foundation (SNSF 43307) to C.A.; by European Regional Fund (KK.05.1.1.02.0010) to G.Z.; by

Georgian state budget to L.U., K.B., and T.Z.; by TUBITAK and Ministry of Agriculture and Forestry of Republic of Türkiye (Grant No. 105 G 078) to A.E.

Author contributions

Y.D., Z.L., S.W., J.S., and W.C. conceived this project. Z.L., K.M., M.M., S.G., G.Z., P.-F.B., T.L., E.M., P.N., K.B., G.D.B., E.D., G.D.L., J.C., C.F.P., R.A.A.G., C.A., A.E., Z.D., V.K., G.S., N.G., S.D., N.O., P.T., C.M., V.L., A.J., L.U., T.Z., D.M., M.H., G.J., E.K., T.D., F.G., F.M., F.S., J.E.D., A.M.D., D.C., G.M., T.U., C.Ö., K.K., M.X., Jiang L, M.Z., L.W., S.J., Y.Z., L.S., and S.L. collected and validated grapevine material. Y.Z., C.M., S.W., S.L., L.T., C.W., D.L., Y.P., Jingxian L., L.Y., X.L., G.X., Z.Y., B.C., Y.W., P.G., M.R., O.R., A.R.-I., Y.W., and S.Z. performed laboratory work. Y.D., H.Y., Y.Z., S.W., J.S., and W.C. supervised the work. S.D., Q.X., and X.D. analysed the data. All authors participated in the interpretation of the data. Y.D., S.D., Q.X., X.D., and W.C. wrote the paper with input from all co-authors.

Competing interests

A.J. is the founder and owner of Historische Rebsorten vineyard. All other authors declare no competing interests.

Correspondence

Correspondence should be addressed to S.W., J.S., or W.C.

Figure Legend

Figure 1. Genetic diversity of global core *V. sylvestris* and *V. vinifera* accessions. a, Principal component analysis of the 2,448 core grapevine accessions. PC1 vs. PC2 projection according to major viticultural regions (Extended Data Fig. 3d). A large square or circle highlights the median position. Grey star shows the position of VS-1. **b**, ADMIXTURE clustering (at K=2 and 8) of the 2,448 core grapevine accessions. Four groups (Syl-W1/2 and Syl-E1/2) with distinct ancestries are identified in *V. sylvestris* and six groups (CG1 to CG6) with distinct ancestries in *V. vinifera*. Pie charts show the geographic locations of the accessions in each group. Gray colour represents minor locations. **c**, PC2 vs. PC3 projection according to grapevine groups. N. East, Near East; F. East, Far East; N. World, New World; C. Asia, Central Asia; Rus/Ukr, Russia/Ukraine; E. Euro, East Europe; C. Euro, Central Europe; W. Euro, West Europe; Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Figure 2. The population history of *V. sylvestris* ecotypes. **a**, Geographic isolation and population separation of *V. sylvestris* ecotypes. Left, distribution of *V. sylvestris* ecotype on present day map. Pie charts show the mean ancestry proportion at K=8 at each location with the same colour scheme in Fig. 1b. Right, estimated split times (100 runs for each comparison) among *V. sylvestris* ecotypes using relative cross-coalescence rate (0.5) analyses with MSMC2. Four random haplotypes in each population. Red bars, median value with 95% confidence interval. **b**, Ecological niche modelling of the suitable habitats for *V. sylvestris* ecotypes at the Last Glacial Maximum (~21 Kya) and early Holocene (~11.7-8.3 Kya). Colour scale shows suitability score. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype.

Figure 3. Independent domestications of *V. vinifera* in the Near East and Caucasus. **a**, Geographic distribution of CG1 and CG2 in relation to the domestication centres. Major dispersal route shown by solid lines with arrows. Putative dispersal route shown by dashed line with arrow. **b**, Outgroup f_3 statistics biplots measuring genetic similarity between CGs, Syl-W, and Syl-E. Rotund, *Muscadinia rotundifolia*. Stars mark the f_3 statistics for Syl-W1/Syl-W2, Syl-E1/Syl-E2, and CG1/CG2 pairs, respectively. **c**, Estimated split times among Syl-E1/2 and CG1/2 populations using relative-cross-

coalescence rate (0.5) analyses with MSMC2 (left). Four haplotypes in each population with 100 runs for each comparison (right). Red bars, median value with 95% confidence interval. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Figure 4. Early diversification of *V. vinifera* in Europe. **a**, Introgression from Syl-W into European *V. vinifera* groups revealed by TreeMix with four migration edges. **b**, Origination of European *V. vinifera* groups (CG4-CG6) by the end of Neolithic. Geographic distribution of CG groups shown by colour circles. Dispersal route of CG1 into Europe shown by a solid line with arrow. Population split times from Momi2 estimates in Extended Data Fig. 12. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Figure 5. Schematic graph of grapevine evolutionary history. Key events in the evolutionary history of grapevines are shown side by side with major events in global climate change and human migration. LGM, Last Glacial Maximum; Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Extended Data Figure Legend

Extended Data Fig. 1. The genome assembly of a *V. sylvestris* accession 'VS-1'. a, Pseudo-chromosomes of the VS-1 genome assembly. Numbers corresponds to the chromosome number used in the *V. vinifera* genome assembly PN40024 (12X.v2). b, Syntenic relationship between the VS-1 genome assembly and PN40024 (12X.v2). c, Comparison of the anchored chromosome lengths in the VS-1 and PN40024 (12X.v2) genome assemblies.

Extended Data Fig. 2. Characterization of SNPs and small Indels from 3,648 *V. sylvestris* and *V. vinifera* accessions. **a**, Density plot of SNPs, small Indels (<40 bp), and nucleotide diversity (π) across 19 chromosomes of the VS-1 genome. **b**, Tabulation of SNPs and small Indels according to the different locations in the genome. **c**, Frequency spectrum of SNPs according to the minor allele frequency brackets and functional annotation. **d**, Size frequency of small Indels in the genome. **e**, Frequency spectrum of small Indels according to the minor allele frequency brackets.

Extended Data Fig. 3. Identification of core *V. sylvestris* and *V. vinifera* accessions in the total sample cohort. **a**, Schematic flowchart for the acquirement of 2,448 core *V. sylvestris* and *V. vinifera* accessions from the total cohort. **b**, Geographical locations of the 2,448 core grapevine accessions around the world. **c**, Identification of clonal, close-cross (e.g., backcross), parent-offspring, and full sibling relationships among 3,525 accessions according to identity-by-state (IBS) sharing patterns. The majority of clonal relationships are among *V. vinifera* individuals and shared by less than five accessions. PO, parent offspring; FS, full sibling; IBS, identity-by-state. **d**, Categorization of core accessions according to the major viticultural regions. N. East, Near East; F. East, Far East; N. World, New World; C. Asia, Central Asia; Rus/Ukr, Russia/Ukraine; E. Euro, East Europe; C. Euro, Central Europe; W. Euro, West Europe.

Extended Data Fig. 4. Principal component analyses of 2,448 core grapevine accessions. The projections are coloured according to major viticultural regions (**a**), grapevine utilization (**b**), and major grapevine groups (**c**). The large square and circle

in (**a**) represent the median positions. Uncategorized and admixed accessions are greyed out.

Extended Data Fig. 5. Maximum likelihood phylogenetic tree of 2,448 core grapevine accessions. **a**, Circular presentation of the maximum likelihood phylogenetic tree with 100 TBE bootstraps. Two major clades are zoomed-in. Each clade contains two smaller clusters. *V. sylvestris* from Near East is located in the clade with a majority of table grapes. *V. sylvestris* from Caucasus and the rest of Europe is located in the clade with a majority of wine grapes. Stars show TBE values greater than 0.70. Small dark circles and blue circles in the zoomed-in clades represent clasped accessions for clarity. **b**, The proportion of table, wine, table/wine, and other types of grapevines in each cluster. C. Asia, Central Asia; E. Euro, East Europe; C. Euro, Central Europe; W. Euro, West Europe.

Extended Data Fig. 6. Categorization of core accessions according to ancestry. **a**, ADMIXTURE clustering of core accessions from $K=2$ to 8. **b**, Representative cultivars from the six *V. vinifera* groups (CG1-CG6) with pure or close to pure ancestries. **c**, Representative admixed *V. vinifera* cultivars with two major ancestry sources. **d**, Representative admixed accessions with a sizeable wild western ecotype component (sky blue Syl-W1 and pink Syl-W2). e, Tri-plot of *V. vinifera* cultivars according to the proportions of *K*2, *K*5, and the other *K*s, showing *K*2 and *K*5 ancestries are associated with table grapevines and all other ancestries with wine grapevines. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Extended Data Fig. 7. Genetic diversity of major grapevine groups with distinct ancestry. a, Pairwise fixation index F_{ST} of major grapevine groups. Yellow colour represents larger population differentiation. Two red boxes show that CG1 is closer to Syl-E1 and CG2 is closer to Syl-E2. **b**, Nucleotide diversity (π, 100 kb window size) distribution of major grapevine groups. **c**, Individual heterozygosity distribution of major grapevine groups. Solid and dashed lines represent median and interquartile range. White diamonds represent mean values. For mean comparisons, $P<0.05$ for a<b<e<c<d from Brown-Forsythe and Welch ANOVA test with Games-Howell post

hoc multiple comparisons. Graph drawn according to the ancestry colour palette. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Extended Data Fig. 8. Linkage disequilibrium in the major grapevine groups. Linkage disequilibrium (LD, r^2) decay of *V. sylvestris* (**a**) and *V. vinifera* (**b**) major groups both show that grapes of the Near East (red lines) and Caucasian (teal lines) descents have the smallest LD extents at around 400 – 500 bp. **c**, LD decay of *V. sylvestris* is only slightly slower than that of *V. vinifera*. **d**, Inverse correlation of LD at 1 Kb and nucleotide diversity (π) from major grapevine groups. Graph drawn according to the ancestry colour palette. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Extended Data Fig. 9. Demographic history of *V. sylvestris* grapevines. **a**, Representative demographic histories of *V. sylvestris* populations from 10^7 to 10^3 years ago deduced from MSMC2. Each line shows estimation from eight haplotypes of four accessions. **b**, Representative split lines among *V. sylvestris* populations based on relative cross- coalescence rate (RCCR) analyses from MSMC2. **c**, Demographic histories of *V. sylvestris* populations deduced from Stairway Plot 2. Red line: median of 200 inferences. Black line: 75% confidence interval. Grey line: 95% confidence interval. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Extended Data Fig. 10. Population split between *V. sylvestris* and *V. vinifera*. Representative split lines between each *V. sylvestris* population and all *V. vinifera* groups based on relative cross-coalescence rate (RCCR) analyses from MSMC2.

Extended Data Fig. 11. Selection and evolution of the sex determination region in the core grapevine accessions. **a**, Identification of domestication selective sweep regions in Syl-E1/CG1 (left) and Syl-E2/CG2 (right) comparison pairs. Red dots have top 5% of F_{ST} and nucleotide diversity. **b**, The sex determination region (SDR) in VS-1 and PN40024 (12X.v2). Selective region marked in light blue. Syntenic genes linked

by grey boxes. Gene shown as blue and yellow boxes. Red triangles indicate identified recombination sites. **c**, SDR genotypes from associated SNPs reveal five recombination sites (dashed lines) and genotype diversity. **d**, Distribution of SDR genotypes in the six major grapevine groups. **e**, Major and minor haplotypes deduced from SDR genotypes. Purple shows female haplotype. Yellow shows male haplotype. Dashed lines show recombination sites. **f**, Recombination history of all SDR haplotypes. **g**, Putative dispersal route of the H4 haplotype and the origination of H2 haplotype. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Extended Data Fig. 12. Introgression of Syl-W and the origination of European grapevines. **a**, Tree structures inferred by TreeMix with zero and four migration edges (m=4). Outgroup is set as Syl-E1. Residual matrices for the two trees are shown. Optimal number of migration edges indicated by the red circle. Migration edges more than four do not substantially increase the composite likelihood $L(m)$. Four migration edges increase the proportion of variance explained from 90.2% ($m=0$) to 99.5%. b, Verification of introgression events with *D*-statistics. Positive numbers indicate gene flow from P3 to P2. Z-score>3. **c**, Four population simulation of split times and genetic introgression using Momi2. Median numbers are obtained from 100 bootstrap runs and marked in the graphs. Syl-W, *V. sylvestris* western ecotype; Syl-E, *V. sylvestris* eastern ecotype; CG, cultivated grapevine.

Extended Data Fig. 13. Local introgression tracts of Syl-W in four *V. vinifera* grapevines. Colour scheme show the relative density of identified introgression tracts. Each tract contains 50 SNPs.

Extended Data Fig. 14. Grapevine group CG3 and muscat flavour. a, Geographic distribution of CG3 grapevines. **b**, Identification of SNPs associated with muscat flavour using FastGWA-GLMM. The significance threshold is set at $-\log_{10}(p)=6.0$. **c**, Zoomed-in genomic regions with significant SNP signatures. Genes closest to the SNPs are coloured in red. The non-synonymous SNP Chr5:19419698 and the corresponding *VvDXS* gene are shown in blue.