Heredity (2010) 104, 351–362
© 2010 Macmillan Publishers Limited All rights reserved 0018-067X/10 $32.00
www.nature.com/hdy

ORIGINAL ARTICLE

Evolution of the *VvMybA* gene family, the major determinant of berry colour in cultivated grapevine (*Vitis vinifera* L.)

A Fournier-Level[1], T Lacombe[1], L Le Cunff[1,2], J-M Boursiquot[1,2] and P This[1]

[1]*INRA UMR 1097 Diversité et Adaptation des Plantes Cultivées, Montpellier, France and* [2]*UMT Géno-Vigne, INRA-IFV, Montpellier, France*

Polymorphisms in the grape transcription factor family *VvMybA* are responsible for variation in anthocyanin content in the berries of cultivated grapevine (*Vitis vinifera* L. *subsp. sativa*). Previous study has shown that white grapes arose through the mutation of two adjacent genes: a retroelement insertion in *VvMybA*1 and a single-nucleotide polymorphism mutation in *VvMybA*2. The purpose of this study was to understand how these mutations emerged and affected genetic diversity at neighbouring sites and how they structured the genetic diversity of cultivated grapevines. We sequenced a total of 3225 bp of these genes in a core collection of genetic resources, and carried out empirical selection tests, phylogenetic- and coalescence-based demographic analyses. The insertion in the *VvMybA*1 promoter was shown to have occurred recently, after the mutation of *VvMybA*2, both mutations followed by a selective sweep. The mutational pattern for these colour genes is consistent with progressively relaxed selection from constrained ancestral coloured haplotypes to light coloured and finally white haplotypes. Dynamics of population size in the *VvMybA* genes showed an initial exponential growth, followed by population size stabilization. Most ancestral haplotypes are found in cultivars from western region, whereas recent haplotypes are essentially present in table cultivars from eastern regions where intense breeding practices may have replaced the original diversity. Finally, the emergence of the white allele was followed by a recent strong exponential growth, showing a very fast diffusion of the initial white allele.

Heredity (2010) **104,** 351–362; doi:10.1038/hdy.2009.148; published online 18 November 2009

Keywords: anthocyanin; white grape; grape domestication; selection

Introduction

Major domestication events of today's most widespread crops started between 12 000 and 10 000 years ago, at the end of the Epipalaeolithic, from their wild relatives. This was illustrated by the first domestic rye grains found in Abu Hureyra in Syria (Zohary and Hopf, 2000). These events led to continuous morphological change during the transition from wild plants in natural conditions to domesticated crops in controlled conditions (Brown *et al.*, 2009). Evidence for these domestication events at the molecular level has been widely emphasized in plants, as for example with the *Y1* locus in maize (Palaisa *et al.*, 2004) and the *qSW5* locus in rice (Shomura *et al.*, 2008). This artificial selection process has favoured the fixation of alleles that were considered beneficial in cultural conditions. Consequently, the crop genetic diversity is believed to have been shaped by selection and adaptation, either ecological or anthropic in changing environmental conditions (Ehrenreich and Purugganan, 2006; Brown *et al.*, 2009).

Grape (*Vitis vinifera* L. *subsp. sativa*) was among the first perennial domesticated crops. Its cultivation started in the late Neolithic, with the establishment of viticulture from 5000 BC in Syria (Goor, 1966). Owing to its easy vegetative propagation, grapevine has been widely distributed, exchanged and spread across the Mediterranean basin for millennia (Unwin, 1991). Distinct uses of grapes for wine making in the West and table consumption on the Eastern shores led to strong disruptive selection, with smaller and juicy berried Western cultivars and bigger and fleshy berried Eastern cultivars. The diffusion of grapevines from Transcaucasus naturally led to an East-to-West pattern of colonization, following human migrations (Mc Govern, 2003). Furthermore, propagation by cuttings, generation overlap and possible secondary domestication events, as well as hybridization with sympatric wild relatives (Levadoux, 1956), led to a very complex pattern of admixture (Aradhya *et al.*, 2003; Arroyo-Garcia *et al.*, 2006). Although nuclear and chloroplastic microsatellite markers helped to resolve questions about the grape genetic structure and cultivar evolution (Arroyo-Garcia *et al.*, 2006; This *et al.*, 2006), the molecular bases of trait adaptation are still poorly investigated in non-model organisms. The primary domestication syndromes in grape are hermaphroditism and the size of berries and bunches (Levadoux, 1956). Colour variation is neither considered as a primary target for selection nor correlated with a particular geographic origin (Sefc *et al.*, 2000; Lijavetzky *et al.*, 2006).

Correspondence: Dr P This, UMR Diversité et adaptation des plantes cultivées, INRA, 2 place Viala, Montpellier 34060, France.
E-mail: this@supagro.inra.fr

Received 3 June 2009; revised 24 August 2009; accepted 17 September 2009; published online 18 November 2009

Grape berry colour is due to the presence of a single pigment family, the anthocyanins, which vary highly in concentration among grape cultivars (Mazza, 1995). While in the ancestral wild vines population, no colour variation was reported, neither at the interspecific level (Cadle-Davidson and Owens, 2008) nor at the intraspecific level (Olmo, 1976). The diversification of the grape colour locus led to the emergence of white genotypes, a derived state (Zohary and Hopf, 2000) that arose from ancestral coloured vine populations. Between these two pools, we observe a continuous range of pale-coloured genotypes, from pink to red. Thus, in contrast to the primary domestication traits, growers within each grapevine growing region have selected distinct colour variants, leading to a diversification process.

Recent investigations on the genetic bases of the trait provided evidence that a single gene cluster, located on chromosome 2, is responsible for most of this variation in colour, and that colour phenotype is due to the combined additive effect of the *VvMybA* gene alleles (Fournier-Level *et al*., 2009). This locus is constituted of a cluster of three MYB-type transcription factor genes, among which *VvMybA*1 and *VvMybA*2 were shown to be functionally involved in berry pigmentation (Kobayashi *et al*., 2002; Walker *et al*., 2007). A third expressed gene, *VvMybA*3, is statistically associated with berry colour determinism but not functionally validated (Fournier-Level *et al*., 2009). The rise of the white berry genotype was recently demonstrated to be the result of the silencing of the first two functional genes (Kobayashi *et al*., 2005; Walker *et al*., 2007). *VvMybA*1 gene silencing is due to the insertion of *Gret*1, a gipsy-type retrotransposon in its promoter (Lijavetzky *et al*., 2006; This *et al*., 2007). For *VvMybA*2, the single-nucleotide polymorphism (SNP) K980 in the coding sequence, which modifies a putative α-helix of the R2R3 recognition domain, also led to a non-functional gene (Walker *et al*., 2007).

Since the publication of the genome sequence of *Vitis vinifera* (Jaillon *et al*., 2007), the resequencing of genes of interest in large genetic resource collections has become feasible, helping to answer hypotheses about the evolution of grape traits. We analysed the sequence polymorphism of these three genes in a core collection of cultivated grape genetic resources (Barnaud *et al*., 2006). The aim was to investigate how *Gret1* and K980 affected the diversity of the colour locus in the cultivated compartment, and whether we could detect different patterns of evolutionary dynamics between the coloured and the white genetic variants at the sequence level. Finally, this study provides a framework for the investigation of perennial crop evolution, where population genetics has very distinct features compared with naturally evolving annual plant populations.

To gain insights into perennial crop evolution, affected by multiple hybridization phenomena and generation overlap, we considered the information at the haplotype level (a genotype possibly being a combination of two haplotypes sharing very distant ancestry). Linked haplotypic combinations of phased polymorphisms at three of the *VvMybA* genes formed the foundation of this study, allowing us to reconstruct the mutational coalescent lineage. After detecting departure from neutrality in the haplogroups, defined by the presence or absence of these mutations, we analysed the emergence of these triggering mutations in the evolutionary process through phylogenetic analyses. Finally, we investigated how the demographic history of the colour locus has been affected by these successive mutational events, using Bayesian reconstruction of effective population size dynamics.

Materials and methods

Plant material

The reference plant material consisted of 137 individuals from a core collection of *Vitis vinifera* L. *subsp. sativa*, maximizing agro-morphological diversity for 50 qualitative and quantitative traits (Barnaud *et al*., 2006). Names, origins and colour profile of the cultivars are shown in Supplementary Table 1. The outgroup consisted of four Asian *Vitis sp.* genotypes: *Vitis balanseana* (Vba), *Vitis amurensis* (Vam), *Vitis pentagona* (Vpe) and *Vitis coignetiae* (Vco). The material is conserved at INRA Domaine de Vassal (http://bioweb.ensam.inra.fr/collections_vigne/).

Molecular analysis

A square inch (80–100 mg) of fresh young leaf was harvested from each genotype. DNA was extracted using Qiagen DNA Plant Mini Kit (Qiagen SA, Courtaboeuf, France) with minor modifications, as described by Adam-Blondon *et al*. (2004).

We aimed to sequence all three *VvMybA* genes that have been shown to have a quantitative effect on colour phenotype. Amplification primers were designed using Primer 3 software (Lincoln S., MIT, Boston, MA, USA). Half of these were anchored in the coding sequence of the genes and the other half in the flanking non-coding and non-conserved regions; the primers are listed in Fournier-Level *et al*. (2009). PCR fragments were amplified, sequenced and analysed as described by Le Cunff *et al*. (2008).

Presence of the *Gret1* retroelement in the *MybA1* locus promoter was genotyped using two forward primers, one specific for the wild allele placed before the insert and the other specific for the white allele inside the insert, and a common reverse primer placed in the first gene exon (Kobayashi *et al*., 2005; This *et al*., 2007). PCR conditions were identical to those described by Kobayashi *et al*. (2005). Amplified fragments described above were then bulked and run together on a 1% agarose gel, stained with ethidium bromide and photographed under ultraviolet light.

Haplotype reconstruction

Owing to the unknown phase of the polymorphisms, we used an algorithmic technique to consistently identify the succession of linked polymorphism along a gene sequence. Haplotypes of *VvMyb*A were reconstructed using a partition-ligation-expectation-maximization algorithm as described in Qin *et al*. (2002) and implemented in PHASE v2.1 (Stephens, University of Washington, Seattle, WA, USA) (Stephens and Scheet, 2005), using a 200 burn-in period with 200 iterations in total, a thinning interval of 1 and 10 repeats. The algorithm was run several times, validating convergence. For molecular evolution estimates, the reconstruction was carried out separately on each gene (*VvMybA*1, *VvMybA*2 or *VvMybA*3). Given their physical linkage within a 43-kb genomic region, the algorithm was run again on the phased genes to reconstruct an entire haplotype combining the information of all three genes.

Identification of recombination events

The reconstructed haplotypes were submitted to three recombination detection tests implemented in the Recombination Detection Program v3beta.28 (Martin, University of Manchester, UK) (Martin *et al*., 2005b) using a $\alpha = 0.05$ threshold with 50 permutations for each test. We used the Bootscan/Recscan method with a window size of 200 bp, with a point at every 20 bp and 100 bootstraps (Martin *et al*., 2005a), and the MaxChi method with a window size of 10 variable sites (Maynard Smith, 1992) and the 3SEQ method (RDP v3) (Boni *et al*., 2007); the use of three alternative methods aimed to ensure consistency. Haplotypes showing a significant recombination probability were excluded from the analyses.

Genetic distance-based and phylogenetic analyses

The genetic distance-based tree was calculated using an unweighted neighbour-joining method implemented in the DARwin software package v5.0.148 (http://darwin.cirad.fr/darwin/Home.php). Dissimilarity matrix was calculated using the Nei distance index with γ parameter set at 0.3. Network analysis was carried out using the median-joining method as described in Bandelt *et al*. (1999) and implemented in Network v4.5.02 (Fluxus Technology, Sudbury, UK). Supplementary analysis was carried out using only the polymorphisms shown to be associated with berry colour and with frequency superior to 0.1 in Fournier-Level *et al*. (2009). We reconstructed another haplotype set with 18 SNPs. A phylogenetic tree was built using a maximum likelihood approach under the Jukes–Cantor model. We used the DNAml program from the Phylip package v3.66 (Felsenstein, University of Washington, Seattle, WA, USA) (Felsenstein, 1981), with the default settings and integrating a global rearrangement of the branches. The likelihood ratio test was maintained to ascertain significance of the branches. The tree was then edited using DARwin v5.0.148.

Molecular evolution analysis

The presence of two polymorphisms with a major influence on grape colour (Fournier-Level *et al*., 2009) helped us to define four haplogroups, among which one was discarded from the analyses because of its recombined origin. These haplogroups were treated separately in the analyses. Molecular diversity parameter estimates were calculated using DnaSP v4.50 (Rozas, Universitat de Barcelona, Spain) (Rozas *et al*., 2003). Per site nucleotide diversity (Π) was calculated according to equation 10.5 of Nei and Tajima (1987). Theta estimates (θ) were obtained according to Watterson's calculation on segregating sites (Watterson, 1975). Tajima's *D*-test was performed according to Tajima (1989).

Coalescence and population size estimation

The optimal substitution model for DNA sequences was determined using MODELTEST v3.7 (Posada, University of Vigo, Spain) (Posada and Crandall, 1998). The optimal substitution model for our data appeared to be HKY + I + G (Hasegawa *et al*., 1985), which is a time-reversible model with some invariant sites and reversible gamma-distributed mutated sites. Differential coalescent-based estimates of the effective population size were calculated using Beast v1.4.8 (Rambaut, Auckland University, Auckland, New Zealand) (Drummond *et al*., 2002; Drummond and Rambaut, 2007) under the relaxed molecular clock hypothesis (Drummond *et al*., 2006), with time following normal distribution of parameters (0,1), and closely related sequence evolution rate following a gamma distribution with parameters (2,1). Each MCMC simulation was performed using a 10-000 burn-in period with a total of 10 000 000 runs and a thinning interval of 10 000 runs. We used the Bayesian Skyline Plot option of Beast v1.4.8 to determine the population size dynamic as a derivative of time in number of mutations per site of the effective population size. For the population growth priors, the Bayesian skyline coalescent model performed systematically better than both exponential and expansion coalescent models. Growth dynamics were then tested taking growth distribution as prior for each haplogroup set (N + Rs + B/Rs/B in the paraphyletic model and N + Rs + B/Rs + B/B in the monophyletic model, respectively, see Figure 1 for haplogroup definition): a uniform, a normal or an exponential growth distribution, thus generating nine paraphyletic and nine monophyletic alternative models. The prior mean time to the most recent common ancestor for each haplogroup set was centred on 1 for the N + Rs + B haplogroup, on 0.8 for Rs haplogroup in paraphyletic model/Rs + B haplogroup in monophyletic model and on 0.7 for B haplogroup, in order to mimic progressive emergence in time. Comparison of alternative models was performed using the Bayesian Factor method, which compares the log harmonic mean likelihood of two alternative models as described in Kass and Raftery (1995), modified in Suchard *et al*. (2001) and implemented in Beast v1.4.8.

Results

Haplotype reconstruction and recombination detection

We analysed the sequences of three genes of the *VvMybA* family involved in grape colour determinism (Kobayashi *et al*., 2005; This *et al*., 2007; Walker *et al*., 2007; Fournier-Level *et al*., 2009), which led to a data set of 3225 bp for each individual. These genes are located in a 43-kb segment on the long arm of the chromosome 2, between the position 14 146 500 and 14 191 500 and corresponded to the isogenes GSVIVT00022659001, GSVIVT00022658001 and GSVIVT00022656001. The description of the polymorphisms is presented in Table 1, the names and genotypes of the sample and complete sequence data set are available in Supplementary Table 1 and Supplementary Data.

The phase reconstruction of each gene haplotype led to the identification of 43 haplotypes of *VvMybA*1, 37 haplotypes of *VvMybA*2 and 71 haplotypes of *VvMybA*3, including haplotypes of recombined origin (Supplementary Data). Although *VvMybA*2 showed the highest number of SNPs, it was also the gene that showed the least number of haplotypes. In contrast, *VvMybA*3, with a restricted number of polymorphisms, presented a high number of haplotypes (Table 1). What we refer to as macrohaplotypes in the following sections are the combination of all the polymorphisms of the three adjacent genes. The reconstruction, with all 62 segregating polymorphisms plus *Gret1*, led to a total of 132 macrohaplotypes.

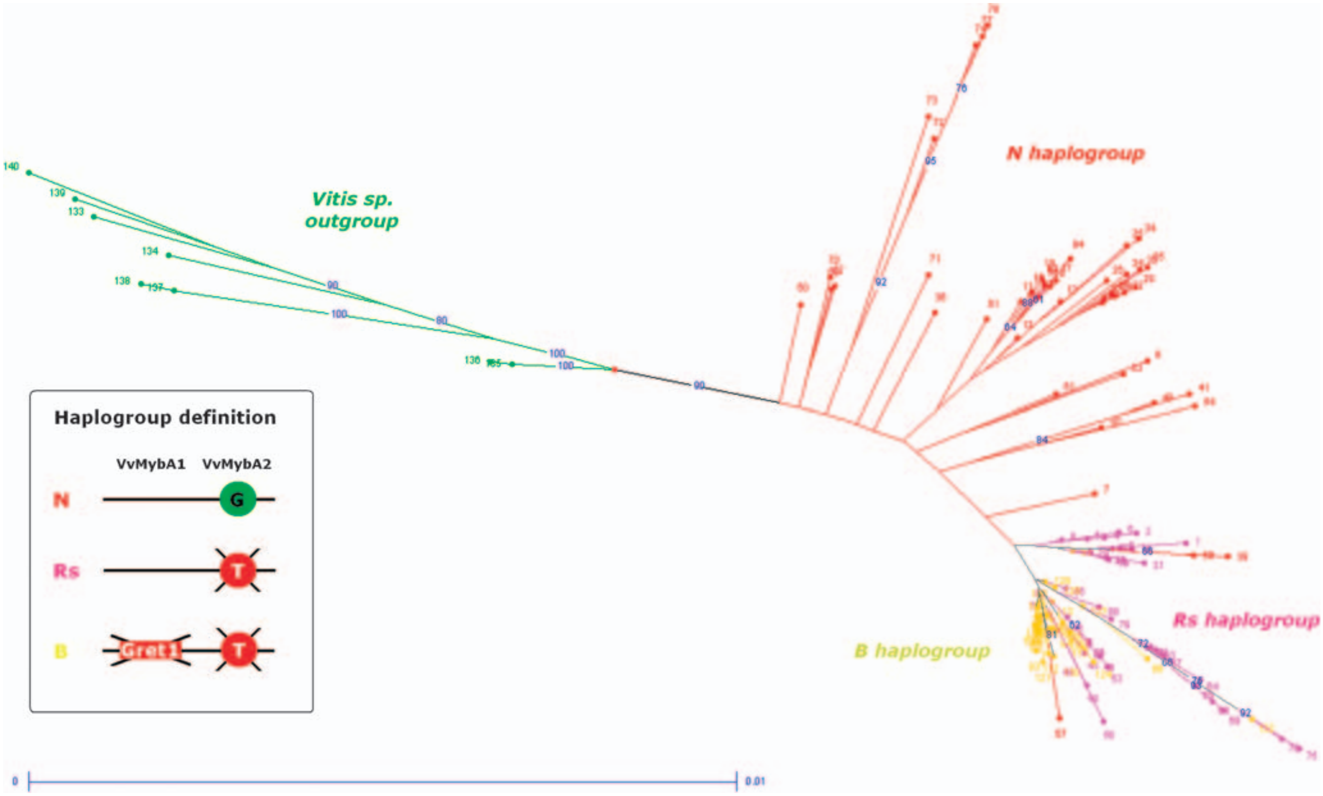


Figure 1 Neighbour-joining tree for the *Vitis vinifera ssp sativa* sample and four *Vitis sp* as an outgroup performed on the three *VvMybA* genes using a neighbour-joining method as implemented in DARwin. Only the bootstraps superior to 50 are presented. The N haplogroup carried no *Gret1* but a functional G allele at K980 corresponding to the 'coloured' haplogroup, the Rs haplogroup carried no *Gret1* but a mutated T allele at K980 corresponding to an 'altered colour' haplogroup, the B haplogroup carried *Gret1* and a mutated T allele at K980, corresponding to a 'white' haplogroup. The Rec haplogroup that carried *Gret1* and the functional G at K980, corresponded to a low-frequency recombined 'white' haplogroup and was discarded.

Table 1 Pattern of diversity and neutrality tests for the three *VvMybA* genes and the three haplogroups excluding the recombined haplotypes

	VvMybA1	VvMybA 2	VvMybA3	*All genes*	*All genes with recombination*
	860 bp	*1236 bp*	*1159 bp*	*3225 bp*	
All groups					
S	16	27	19	62 (63^a)	62 (63^a)
H	36	35	66	116 (117^a)	131 (132^a)
π	0.0039	0.0033	0.0027	0.0032	0.0035
θ	0.003	0.0036	0.0027	0.0031	0.0031
$D_{\mathbf{Tajima}}$	0.75	−0.21	−0.004	0.12	0.35
N haplogroup					
S	16	26	19	61	
H	22	27	37	48	
π	0.0054	0.0056	0.0052	0.0054	
θ	0.0041	0.0046	0.0036	0.0041	
$D_{\mathbf{Tajima}}$	0.96	0.69	1.34	1.04	
Rs haplogroup					
S	16	4	19	39	
H	18	5	26	38	
π	0.0057	0.00024	0.00285	0.0027	
θ	0.0041	0.00072	0.00364	0.0027	
$D_{\mathbf{Tajima}}$	1.14	−1.49*	−0.69	−0.01	
B haplogroup					
S	14	5	15	34	
H	9	7	16	33	
π	0.00071	0.00038	0.00041	0.00048	
θ	0.0029	0.00072	0.0023	0.0019	
$D_{\mathbf{Tajima}}$	−1.95*	−0.94	−2.15**	−2.17**	

The transitions from the N to the Rs haplogroup in *VvMybA*1 and from the Rs to the B haplogroup in *VvMybA*2 show significant selection. *S* indicates number of segregating sites and *H* haplotype number, *$P<0.01$, **$P>0.001$.
[a]Data including the *Gret1* retroelement.

To avoid bias in the neighbour-joining calculation and coalescent inference, recombination detection was performed with high stringency on the global sets of macrohaplotypes; 31 recombination events were detected and 15 were considered significant (data not shown). Recombination is known to introduce an excess of long external branches that generated a long-branch attraction effect in the phylogenetic inference (Sanderson *et al*., 2000), which led to a loss of power in empirical selection tests (Tenaillon and Tiffin, 2008) and to the loss of the molecular clock (Schierup and Hein, 2000). The 15 recombined macrohaplotypes were removed from the analyses as advocated in Avise (1994), Posada and Crandall (2002) and Gantenbein *et al*. (2005) because of misleading tree building algorithms.

Selection and diversification driven by the mutational events

Considering the two polymorphisms that have been shown to be functionally responsible for berry colour variation (Kobayashi *et al*., 2005; Walker *et al*., 2007), namely *Gret1* and K980, four macrohaplotype clusters or 'haplogroups' were defined by the presence or absence of the *Gret1* retroelement in the *VvMybA*1 promoter, and by the presence of a functional G allele or a mutated T allele at the SNP K980 in the *VvMybA*2 coding sequence (see Figure 1 for haplogroup description). Almost all the macrohaplotypes carrying the *Gret1* insertion and a functional G allele at K980 (12 of 15) were identified as recombined and had been already removed. The three remaining macrohaplotypes, derived from recombined macrohaplotype by a single mutation, were thus removed from the analyses. This haplogroup had a frequency of 0.05 in the entire sample and is referred as Rec in Supplementary Data.

We observed a progressive reduction of diversity in the number of segregating SNP (*S*) and the number of macrohaplotypes (*H*), diversity (Π) and polymorphism (*θ*) during the transition from the N to the B haplogroup (Table 1). Both N and Rs haplogroups showed quite similar Π and *θ* values, whereas the B haplogroup showed a Π value four times lower than *θ* (Table 1). Considering the genes independently, *VvMybA*1 appeared more polymorphic in the N and Rs haplogroups than in the B haplogroup, and *VvMybA*2 appeared more polymorphic in the N haplogroup than in the Rs and B haplogroups. *VvMybA*3 appeared very variable, given that 19 segregating SNPs allowed us to reconstruct 71 haplotypes, which showed independent neutral mutations. Tajima's *D*-tests were significant in the B haplogroup for *VvMybA*1, *VvMybA*3 and the combined genes, and in the Rs haplogroup for *VvMybA*2 genes (Table 1). This showed a deviation from neutrality within the Rs and the B haplogroups. Considering now the genotype combination, only nine varieties appeared homozygous for N haplotypes and one for the Rs haplotypes, while in the same time 34 coloured varieties are combination of either N or Rs haplotype with a B haplotype. Among the white varieties, 19 appeared homozygous for haplotype 51.

The white allele arose progressively through mutation of K980 followed by insertion of Gret1

We refereed to four Asian genotypes of the genus *Vitis* as an outgroup (Figure 1). We clearly observed, with good bootstrap support, that the N haplogroup is the ancestral

Table 2 Name, origin and use of the cultivars carrying the probable transitional haplotypes of the N and Rs haplogroups

Haplotype	*Vassal accession number*	*Name*	*Origin*	*Ancestry*	*Age estimate*	*Use*	*Genetic pool*
Ancestral haplotype (60)	1269Mtp1	Corvina vincentina	Italy	?	Old	Wine	West
Reference N haplotype (11, 16, 17, 23 and 130)	129Mtp12	Chatus	France/Italy	Pinot x ?	Medium/old	Wine	West
	159Mtp12	Durif	France	Syrah x Peloursin	1880	Wine	West
	154Mtp1	Joubertin	France	Persan x Peloursin	1835	Wine	West
	1991Mtp1	Meunier court maillé	France	Pinot x ?	Clonal event	Wine	West
	487Mtp1	Petit Bouschet x Aramon no°4	France (Languedoc)	Petit Bouschet x Morrastel	Recent	Wine	West
	1216Mtp1	Mancin	France (aquitaine)	?	Medium/old	Wine	West
Last N haplotype (39, 57, 59)	2131Mtp1	Malvarisco	Portugal	Tinto cao 1488 x Alfrocheiro preto 50	Medium	Wine	West
	1529Mtp1	Tinta Madeira	Portugal	?	Medium/old	Wine	West
	18Mtp17	Carignan	Spain	?	Old	Wine	East
First Rs haplotype (4, 5, 8)	1595Mtp3	Hans Rot	Germany	Velteliner rouge 284 x ?	Medium	Wine	West
	1575Mtp1	Heunisch rot	Austria	Gouais x ?	Medium	Wine	West
	2812Mtp1	Jardovany Fekete	Hungary	Furmint x ?	Medium	Wine	West
	1944Mtp1	Kincsem	Hungary	Mathiasz Janosne x Kövidinka	1917	Mixed	East
Last Rs haplotype (44)	2608Mtp1	Chenin rose	France	Mutation of Chenin	Clonal event	Wine	West
	1575Mtp1	Heunisch rot	Austria	Gouais x ?	Medium	Wine	West
	2567Mtp1	Örökké piros	Hungary	Mathiasz Janosne x ?	Very recent	Table	East
	1235Mtp1	Vernaccina	Italy ?	?	Old	Mixed	East

Haplotype numbers in parentheses refer to the network analysis.
?, unknown parentage.

haplogroup and that neither the *Gret1* retrotransposon nor K980 were found in the four accessions of *Vitis* sp. investigated. Macrohaplotype 60, found in cultivar ‘Corvina vicentina’ (Table 2), had the smallest genetic distance to the outgroup haplotypes. This macrohaplotype may represent the most ancestral form of *VvMybA* combination present in our sample. Owing to weak bootstrap support, the neighbour-joining tree provides only limited information about the actual topology within the *Vitis vinifera* haplogroups, but nonetheless justifies the coherence of haplogroup distinction.

To show all the connections likely to have occurred between the macrohaplotypes, a network was constructed with 63 informative sites (Figure 2). The network showed 395 potential mutation steps and 86 mutations for the shortest tree, enlightening homoplasy or time reversibility. In this network, the N, Rs and B haplogroups clustered easily. The network analysis showed no likely direct connection between the N and B haplogroups. The most frequent macrohaplotype within haplogroup N (41) only had a frequency of 0.06, the most frequent macrohaplotype within haplogroup Rs (44) had a frequency of 0.1, and the most frequent macrohaplotype within the B haplogroup (51) had a higher frequency of 0.51. Interestingly, the most frequent macrohaplotypes of both the Rs and B haplogroups were placed within the torso section of the network, whereas most of the frequent macrohaplotypes in the N haplogroup were displayed on external branches. Mean pairwise distances were greater in the N haplogroup compared with the Rs and B haplogroups (Figure 2). The star-like diversification pattern around macrohaplotypes 11, 16, 17, 23 and 130 of the N haplogroup, close to ancestral macrohaplotype 60 and found in old cultivars (Table 2), tend to favour these sequences as being the closest to the initially domesticated gene pool.

All coloured table grape genotypes except two appeared to carry at least one allele from the Rs haplogroup ($\chi^2 = 1.0$, $P < 0.001$). The two most frequent macrohaplotypes of each of the haplogroups Rs and B differed only by the presence or absence of the *Gret1* retroelement in the *VvMybA*1 promoter, macrohaplotype 44 thus being the most likely background for the insertion.

We calculated a supplementary phylogenetic tree (Supplementary Figure 1) with only polymorphisms with frequency >0.1 but including the recombined haplotypes and the Rec haplogroup. It demonstrated that the Rec haplogroup has no consistency and is branched at three different nodes of the tree. Furthermore, none of the macrohaplotypes of the Rec haplogroup is connected at the interface between either the N and Rs haplogroups or the Rs and B haplogroups. Finally, integrating the Rec haplogroup in the reduced analysis did not change the topology presented in this study. Moreover, the sequencing of the 3′-LTR/insertion part of *Gret1* showed no variation (data not shown), supporting the hypothesis of a single insertion event that occurred in the B haplogroup. Two other likely paths of the network could alternatively result from *Gret1* excision, as observed by Lijavetzky *et al*. (2006).

Population size dynamics following mutational events

To infer the effective size changes at the *VvMybA* locus, a coalescent-based approach was used to test the likelihood of different population growth dynamics under various hypotheses. We compared different scenarios to find which of them showed the greatest likelihood of emergence of the B haplogroup from the Rs haplogroup (monophyletic model). This was compared with the scenarios of the independent emergence of both two haplogroups from an N + Rs + B ancestral background (paraphyletic model) (Figure 3 and see the Materials and methods section for scenarios description).

For coalescent population size dynamics, the best results were obtained by considering initial growth as exponential, successively followed by uniform growth and another exponential growth phase, both for the paraphyletic (H_0) and for the monophyletic model (H_1). The optimal two alternative models showed a Bayesian factor of 7.1 in favour of the monophyletic model (Figure 3), representing ‘positive’ support for the monophyletic scenario according to the interpretation of Kass and Raftery (1995). Alternatively, a monophyletic model with a later emergence of the Rs haplogroup (instead of the B) was tested and performed worst (data not shown). We then applied the Bayesian Skyline analysis to follow the demographic evolution of the whole sample that would reflect how the two successive selective sweeps undergone by the *VvMybA*s gene cluster could affect local mutational dynamics (Figure 4). The output of the Bayesian Skyline clearly showed two exponential growth sequences, an old one of medium intensity and a recent very strong one. Replacing the population size change in the mutational history of the *VvMybAs* gene cluster, the ancestral population was shown to have undergone an initial exponential growth phase. This was followed by a stabilization of the population size, implying that K980 led to no change in population size and finally a strong exponential growth following the insertion of *Gret1*, suggesting this late event has had a key role in grape diffusion.

Discussion

Sampling haplotypes from a core collection

Instead of taking each cultivar, a diploid genotype, as the unit of the work, a haplotype-based approach was conducted as performed successfully in human (Tishkoff *et al*., 2007) and plant evolutionary studies (de Alencar Figueiredo *et al*., 2008). This was done to overcome the difficulty of interpreting migration and hybridization phenomena linked to an ancient and widespread diffusion of grapevines. Several methods of haplotype reconstruction were developed, among which the partition-ligation-expectation-maximization method (Qin *et al*., 2002) appears very robust and provides stable results, although it has heavy computing requirements (see for review Niu, 2004). Owing to the initial pattern of genotypic diversity, with a great number of genotypes presenting the most common B macrohaplotype, the algorithm could easily converge within a short burn-in period (500 samplings), making the haplotype reconstruction consistent.

As shown by the network analysis (Figure 2), we observed few extremely frequent macrohaplotypes but many infrequent to singleton macrohaplotypes. This feature can be explained because samples were taken from a core collection, designed on morphological traits

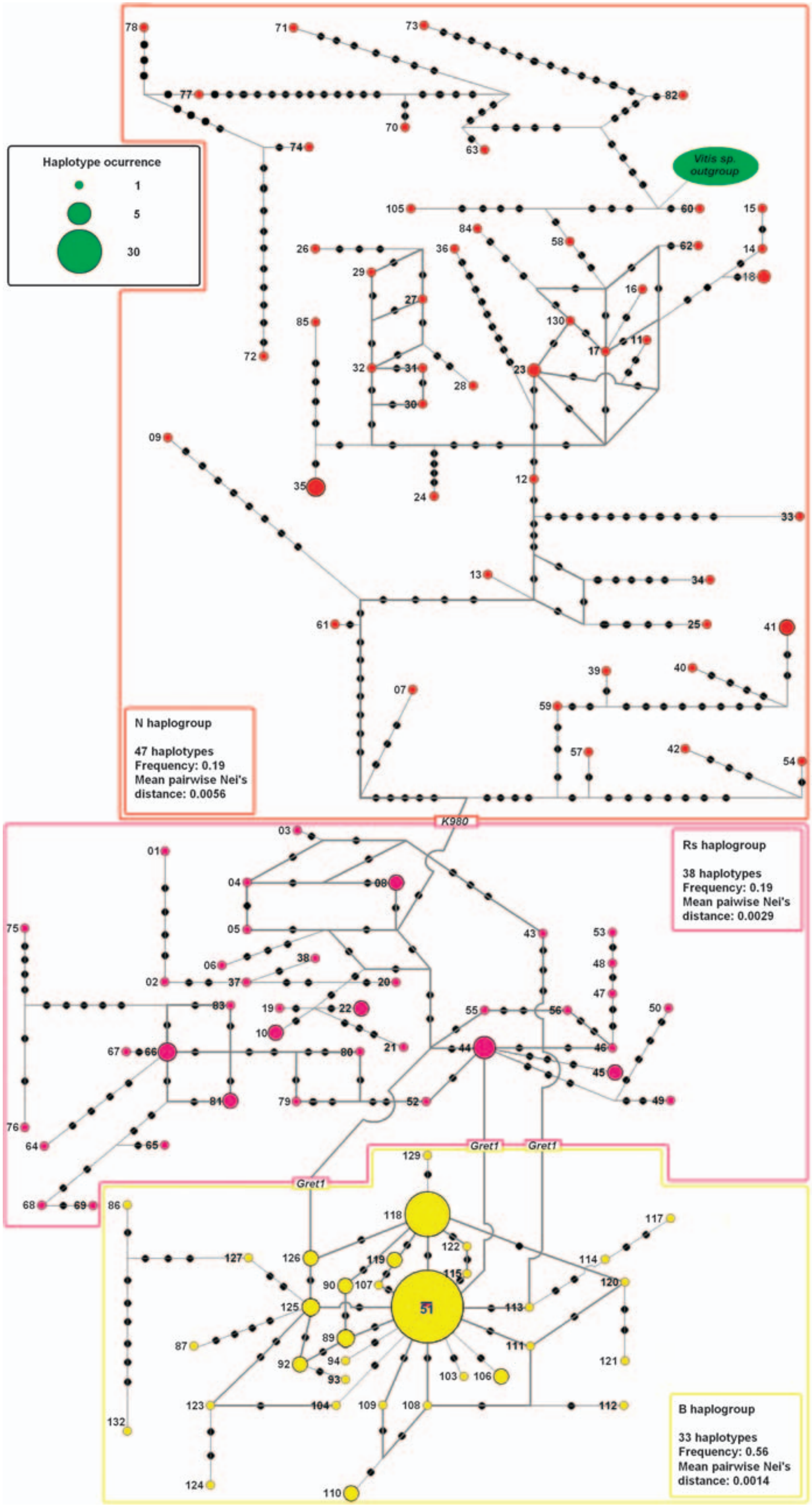
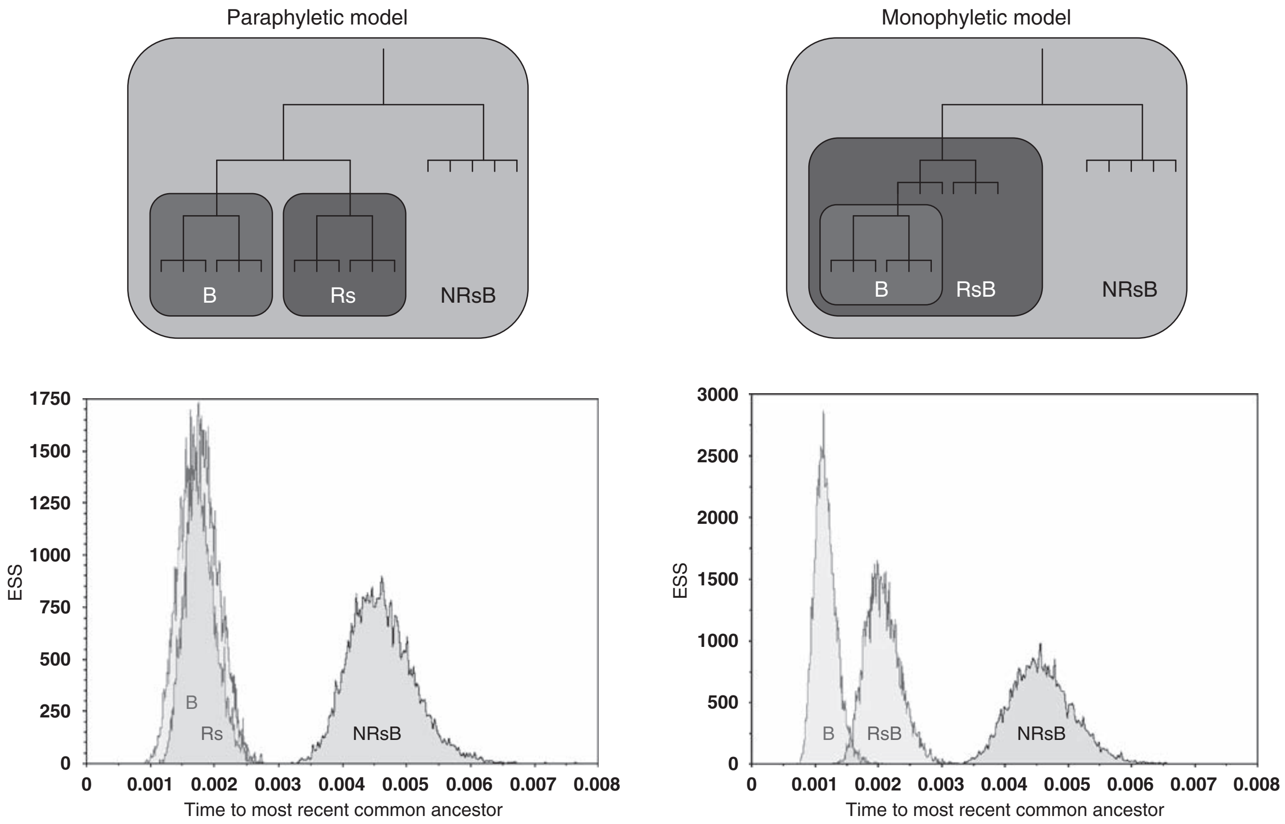


Figure 2 Network analysis carried out on 63 polymorphic sites in three *VvMybA* genes using the median-joining method (Bandelt *et al*., 1999). The network shows all likely connections between the N, Rs and B haplogroups. The bold branches correspond to the torso of the network, the black dots correspond to a mutational step and coloured dots correspond to haplotypes, with size proportional to their frequency.



Paraphyletic model				Monophyletic model		
	Mean	ESS			Mean	ESS
t_{mrca} (NRsB)	0.00459	801.5		t_{mrca} (NRsB)	0.00474	1464.8
t_{mrca} (Rs)	0.00178	292.4		t_{mrca} (RsB)	0.00215	759.3
t_{mrca} (B)	0.00159	317.7		t_{mrca} (B)	0.00117	429.9
Log Pr(**D**\para)	-8147,6 (0.371)			*Log Pr(D\mono)*	-8140,5 (0.294)	
		Bayes Factor = 7.1				

Figure 3 Alternative scenarios for coalescent population growth analysis performed using Beast v1.4.8. *ESS* indicates effective sample size of the MCMC output, considered satisfying in a genealogy sampling when superior to 200 (Kuhner, 2009); t_{mrca}, time to the most recent common ancestor expressed in number of mutations per site since present time. *Log* is harmonic mean likelihood is calculated as in Drummond *et al*. (2006), Bayes Factor is calculated as in Kass and Raftery (1995) with modification from Suchard *et al*. (2001) and tends to favour the monophyletic model.

by *M*-maximization method (Schoen and Brown, 1993). This method involves sampling the maximum variability and limiting the overall redundancy; it thus biases the homothetic allele frequency representation between the grape natural diversity present at INRA Domaine de Vassal repository and the resulting core collection sample. Nonetheless, a core collection sample is beneficial for coalescent inference because it favours rare intermediary alleles and limits the overrepresentation of common ones, leading to a balanced phylogeny. This research strategy is thus suitable for studying phylogenetics and evolution in any species, where a core collection has been designed and where genotypic state is not very informative, such as in any highly hybridized plant with a long biological cycle. Furthermore, although the Domaine de Vassal collection of grape genetic resources is the largest in existence, numerous accessions consist of cultivars from traditional European wine growing countries such as Italy, Spain, Hungary, France and Portugal. This explains why the intermediary rare alleles in our sample are mainly found to originate from these regions.

Genetic diversity based on nuclear DNA sequence

The evolution of nuclear genes determining a quantitative trait in a vegetatively propagated crop is not only affected by germinal and somatic punctual mutation but also mixed by recombination and hybridization. The small number of generations and vegetative propagation imply limited recombination and have helped to maintain extensive linkage disequilibrium (LD; Barnaud *et al*., 2006), especially in the context of loci linked within a 43-kb region (Fournier-Level *et al*., 2009). Nonetheless, Schierup and Hein (2000) showed that recombination led to overestimation of the length of internal branches, regardless of its intensity, conflicting with the hypothesis of constant evolutionary rate along the coalescent phylogeny (Hudson, 1983). Taking as a prior the neutral

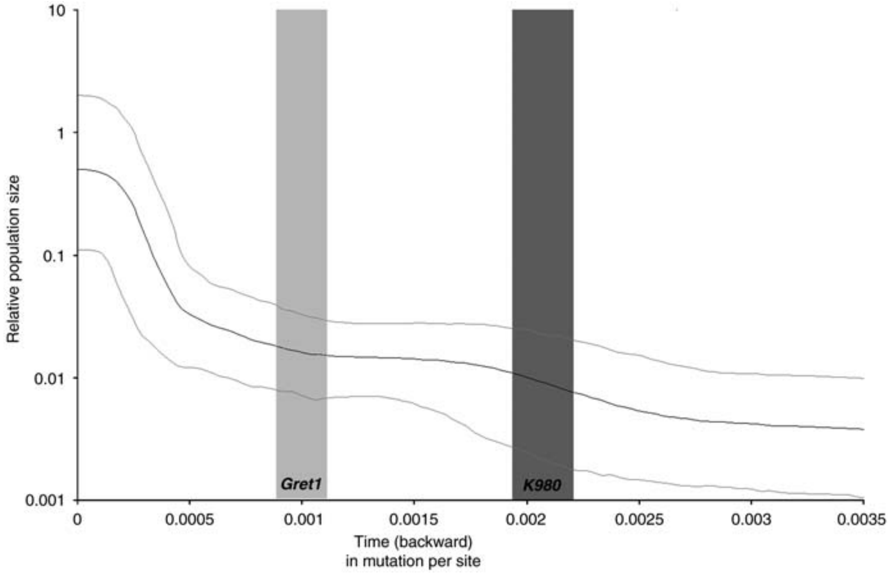


Figure 4 Effective population size evolution during the coalescent history of the *VvMybA* locus as inferred by the Bayesian Skyline method. The black line indicates the median of the simulation and grey lines indicate the standard deviation. The best fitting model shows initial exponential growth followed by uniform and new stronger exponential growth. Time is expressed backwards in number of mutations per site from the present time. Ages of the K980 mutation and *Gret1* insertion ± s.d. are indicated as vertical bars.

evolution model for DNA sequence mutation with limited recombination and under the molecular clock hypothesis (Kimura, 1983), the evolutionary history of the genes through the appearance of SNP mutations was followed. In this case, using the relaxed molecular clock model (Drummond and Rambaut, 2007), and supposing an equal number of non-detected recombinations within each haplogroup, the Bayesian Skyline reconstruction allowed a comparison between relative haplogroup dynamics.

The high diversity of grapes is hardly coherent with the reduced number of generations since domestication, believed to be below 100 (J-M Boursiquot, personal communication). However, grape, as the major and oldest fruit crop, has a huge population size planted over multiple locations, compared with low-numbered and isolated wild relatives. Consequently, cultivated grapes have to be seen as a complex meta-population with big effective population size and generation overlap that can easily lead to fast evolution rates. Another origin of such high DNA polymorphism is either hybridization with wild relatives which just recently started to be investigated (Di Vecchi-Staraz *et al*., 2009) or the somatic mutations induced by massive vegetative propagation through cuttings. Since domestication, cultivated grape as evolved in sympatry with its wild relative, *Vitis vinifera* L. *subsp. silvestris*, and hybridization phenomenon is likely to have occurred. An extended genetic diversity survey of wild grapes would provide precious information (Lacombe *et al*., 2003), allowing to estimate the gene flow between wild and cultivated grapes, such analysis is been now undertaken. Nonetheless, the phylloxera crisis at the end of the nineteenth century led to a drastic reduction in population size (Levadoux, 1956) and the capture of genetic diversity realized nowadays may not reflect the diversity present in the ancestral gene pool.

Somatic mutation could have also contributed to shaping diversity (Antolin and Strobeck, 1985; Moncada *et al*., 2006). An illustration of this phenomenon is the large clonal diversity of grape varieties, relative to their age and diffusion (43 distinct clones of Pinot Noir and 26 of Cabernet-franc certified in France; ENTAV-ITV, 2007). Furthermore, most of the antique cultivars have a complete series of coloured mutants such as Pinot or Grenache noir, gris and blanc. The determinants of such variation corresponding to major genomic re-arrangements, such as *Gret1* instability (Lijavetzky *et al*., 2006; Walker *et al*., 2007), and possibly maintained in chimeric states (Hocquigny *et al*., 2004) have not been investigated in this study. Multiple *Gret1* insertion/excision or recombination events may explain the contrasting pattern of selection in the Rs haplogroup. The purpose of this study was rather to identify the founder mutational events that have lead to present-day diversity and focus on sexually inherited genetic diversity.

Selection for white allele and reduction of anthocyanin content

Both the K980 mutation and the *Gret1* insertion significantly led to positive selection or generated a founder effect of the mutant allele during grapevine evolution ($D_{\mathrm{Tajima}} < -1$; Table 1). Considering the three genes separately, the results for *VvMybA*1 and *VvMybA*2, demonstrated as functional (Kobayashi *et al*., 2002; Walker *et al*., 2007), showed a reduced level of haplotypic diversity compared with *VvMybA*3, which was reported as truncated and putatively non-functional (Walker *et al*., 2007). This last gene may interact with the other two expressed *VvMybAs*, but only *VvMybA*1 and *VvMybA*2 are submitted to a functional constraint. This may illustrate the pseudogenization phenomenon of certain duplicated isogenes (Moore and Purugganan, 2005).

The N clade showed a D_{Tajima} around 1 and a mean pairwise genetic distance superior to what was found in the B clade, resulting in a pattern of structured diversity, coherent with an older origin and continuous selective constraint for functionality (Figure 2). The B clade showed a pattern of polymorphism that agrees with a recent selective sweep and/or strong exponential growth. We concluded that the N haplogroup appeared under balanced selection ($D_{\mathrm{Tajima}} = 1$) contrasting with the B haplogroup appearing highly selected or having undergone exponential growth ($D_{\mathrm{Tajima}} < -1$). The Rs haplogroup showed a contrasted pattern of diversity with balanced frequency polymorphisms.

Starting from a highly coloured wild ancestor, the *VvMybA*s gene cluster showed progressive emergence of a non-functional allele following the selection of reduced efficiency alleles. The issue of knowing whether the observed diversity reduction at a single locus is due to a founder effect or due to a targeted selective sweep is challenging, as both phenomena result in the same pattern locally. Nonetheless, as the selective sweeps concerned one or two haplogroups out of three within the same population, we may conclude that we effectively detected positive selection followed by the rapid increase in frequency of the selected haplotypes. Furthermore, it is not clear whether an extensive marking along chromosome 2 alone would allow us to distinguish between demography and selection phenomena, owing to the extensive pattern of LD (Barnaud *et al*., 2006).

At the genotype level, the crop evolution has favoured an excess of heterozygous genotypes (N/B), with the combination of functional and highly constrained N alleles and the emergence of non-functional alleles that diffused extremely rapidly. Almost all the common black varieties (Pinot, Merlot, Syrah, etc.) are heterozygous at the *VvMybA*s loci and postdate the emergence of the white alleles, revealing a putatively advantageous cryptic presence in most black grape varieties. On the whole data set, the nine genotypes homozygous for N macrohaplotypes corresponded to highly coloured wine varieties (Fournier-Level *et al*., 2009). These two features are reminders that anthocyanin pigment is deleterious at high concentration and that its synthesis also has a huge metabolic cost (Dixon *et al*., 2001).

Demography and dynamics of the diversity at the grape colour loci

Although the differences in diversity dynamics between N and B haplogroups are easily detected (Figure 2 and Table 1), the case of grape is particularly complex because of our poor understanding of its molecular evolution (mutation rate and effective population size), while good records of its history are available. As a domesticated crop, grapevine underwent various diffusion events, starting with the domestication itself, and continuing with expansion to new regions and new breeding practices. The radiation pattern around the macrohaplotypes 11, 16, 17, 23 and 130 (Figure 2) and the initial exponential growth (Figure 4) reflected the fast expansion of grape. Again the star-like pattern around macrohaplotype 51, the most likely founder of the B haplogroup, reflected a later exponential expansion. Given the low pairwise distance between macrohaplotypes of the B clade, the *Gret*1 insertion is believed to have only occurred recently, from macrohaplotype 44 of the Rs haplogroup. This recent and extremely rapid diffusion of the white grape is coherent with a late spread of viticulture after the decline of the Roman Empire (Schenk, 1992). Nonetheless, we tested models of an independent emergence of K980 and Gret1 and these alternative coalescent models performed systematically worse. Given the low frequency and the inconsistency of the Rec haplogroup (mean pairwise Nei's distance of 0.0113), the hypothesis of independent emergence of Gret1 outside the B haplogroup is not the most likely.

Although the *Gret1* insertion has been considered as the main factor determining grape colour (Kobayashi *et al*., 2005; This *et al*., 2007), the K980 mutation of *VvMybA*2 appears very relevant in the *VvMybA* diversification process. This mutation corresponded to a diversification node and both the significance of the Tajima tests for *VvMybA*2 in the Rs haplogroup and the significantly specific presence of Rs macrohaplotypes in nearly all coloured table grapes ($\chi^2 = 1.0$, $P < 0.001$) highlight the important distinction between the N and Rs haplogroups. Although it was not followed by a drastic change in effective population size, the K980 mutation led to a significant founder effect (decreased *VvMybA*2 diversity in the Rs haplogroup, Figure 4 and Table 1).

Domestication is likely to be a protracted process (Tanno and Willcox, 2006; Fuller, 2007; Allaby *et al*., 2008) that occurred in multiple locations (Molina-Cano *et al*., 2005; Özkan *et al*., 2005). This process of domestication can be seen as a gradual succession of selection steps leading to the 'elite' crop cultivated nowadays. Grapes, a perennial plant maintained through vegetative propagation, apparently kept the fingerprint of these past steps. Finally, the two mutations under study appeared to have only occurred in the last third of the past history of cultivated grape (Figure 4). The origin and precise date of appearance of the key colour mutations will hardly be assessed through molecular genetics and may also have appeared in the wild compartment. Nonetheless, its selection and fast diffusion concerned the cultivated compartment and greatly contributed to shape the berry colour diversity.

Geographical origin and use, an interpretation of mutated gene emergence

Grape (*Vitis vinifera subsp. sativa*) is an outcrossing, vegetatively propagated, perennial crop (Sefc *et al*., 2000; Aradhya *et al*., 2003). These features imply interleaved generations, allowing for a highly overlapping parent–offspring relationship. Old cultivars multiplied vegetatively for many years coexist and hybridize with brand new breeds obtained from genetically distant parents. Nonetheless, we tried to link the conclusion drawn at the macrohaplotype scale with knowledge at the grape genotype scale. We correlated the pattern of genetic diversity with the traditional structure of grape agro-morphological diversity, most common use (wine vs table; Levadoux, 1956), geographic origin and structure assignation to Eastern or Western populations (Le Cunff *et al*., 2008).

The presence of the N macrohaplotypes forming a radial pattern (macrohaplotypes 11, 16, 17, 23 and 130, Figures 1 and 2) in old wine type cultivars and their closeness to the *Vitis sp.* outgroup reinforced the hypothesis of an ancestral origin of the N haplogroup. In our sample, most of the grapes carrying either of the last N macrohaplotypes appeared to come from the Iberian Peninsula, suggesting that Spanish and Portuguese grapes have kept the fingerprint of the transition from the N to the Rs haplogroup. This observation may either be due to an isolation of these varieties in a region considered as a refuge zone (Olalde *et al*., 2002), with limited gene flow from the East, or due to hybridization with non-domesticated endemic *Vitis vinifera* L. *subsp. silvestris* (Arroyo-Garcia *et al*., 2006). Furthermore, the Rs macrohaplotypes closer to the N haplogroup were essentially found in Central European genotypes, emphasized as the geographical source of different colour variation. Contrary to the coloured wine grapes, which appeared to randomly carry either N or Rs macrohaplotypes, the near systematic presence of Rs macrohaplotypes in all coloured table grapes supported the hypothesis of a more recent and eastern breed of the table cultivars, compared with the wine cultivars. Interestingly, these findings tend to favour the hypothesis that the western grapes have conserved the ancestral macrohaplotypes, whereas the eastern grapes are the result of continuous and intensive breeding practices that have led to the loss of antique diversity, even that endemic to these regions (Levadoux, 1956). This evidence favours the correlation between lower selective pressure for the wine uses in the western region and an increased selection for table use in the eastern region (Arroyo-Garcia *et al*., 2006; Le Cunff *et al*., 2008). Furthermore, selection process for table grapes is more straightforward because of easier quality trait evaluation. Finally, molecular differences between the preferred presence of Rs haplotypes in table grape and N haplotypes in wine grape enlightened a phenomenon of selective adaptation to a particular use.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Stéphane De Mita, Anne-Céline Thuillet and Bouchaïd Khadahari for careful reading and fruitful comments. This work has been supported by INRA and Région Languedoc-Roussillon, the ANR CoreGrapeGen project and the European project FLAVO.

References

Adam-Blondon AF, Roux C, Claux D, Butterlin G, Merdinoglu D, This P (2004). Mapping 245 SSR markers on the *Vitis vinifera* genome: a tool for grape genetics. *Theor Appl Genet* **109**: 1017–1027.

Allaby RG, Fuller DQ, Brown TA (2008). The genetic expectations of a protracted model for the origins of domesticated crops. *Proc Natl Acad Sci USA* **105**: 13982–13986.

Antolin MF, Strobeck C (1985). The population genetics of somatic mutation in plants. *Am Nat* **126**: 52.

Aradhya MK, Dangl GS, Prins BH, Boursiquot JM, Walker MA, Meredith CP *et al.* (2003). Genetic structure and differentiation in cultivated grape, *Vitis vinifera* L. *Genet Res* **81**: 179–192.

Arroyo-Garcia R, Ruiz-Garcia L, Bolling L, Ocete R, Lopez MA, Arnold C *et al.* (2006). Multiple origins of cultivated grapevine (*Vitis vinifera* L. *ssp sativa*) based on chloroplast DNA polymorphisms). *Mol Ecol* **15**: 3707–3714.

Avise JC (1994). *Molecular Markers, Natural History and Evolution*. New York, 511pp.

Bandelt HJ, Forster P, Rohl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.

Barnaud A, Lacombe T, Doligez A (2006). Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L. *Theor Appl Genet* **112**: 708–716.

Boni MF, Posada D, Feldman MW (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**: 1035–1047.

Brown TA, Jones MK, Powell W, Allaby RG (2009). The complex origins of domesticated crops in the Fertile Crescent. *Trends Ecol Evol* **24**: 103–109.

Cadle-Davidson MM, Owens CL (2008). Genomic amplification of the Gret1 retroelement in white-fruited accessions of wild Vitis and interspecific hybrids. *Theor Appl Genet* **116**: 1079–1094.

de Alencar Figueiredo LF, Calatayud C, Dupuits C, Billot C, Rami JF, Brunel D *et al.* (2008). Phylogeographic evidence of crop neodiversity in Sorghum. *Genetics* **179**: 997–1008.

Di Vecchi-Staraz M, Laucou V, Bruno G, Lacombe T, Gerber S (2009). Low level of pollen-mediated gene flow from cultivated to wild grapevine: consequences for the evolution of the endangered subspecies *Vitis vinifera* L. *subsp silvestris*. *J Hered* **100**: 66–75.

Dixon P, Weinig C, Schmitt J (2001). Susceptibility to UV damage in *Impatiens capensis* (Balsaminaceae): testing for opportunity costs to shade-avoidance and population differentiation. *Am J Bot* **88**: 1401–1408.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**: 1307–1320.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006). Relaxed phylogenetics and dating with confidence. *Plos Biol* **4**: 699–710.

Drummond AJ, Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 214.

Ehrenreich IM, Purugganan MD (2006). The molecular genetic basis of plant adaptation. *Am J Bot* **93**: 953–962.

ENTAV-ITV (2007). *Catalogues des variétés et clones de vignes cultivés en France*, 2nd edn. ENTAV-ITV: Montpellier, France, 455pp.

Felsenstein J (1981). Evolutionary trees from DNA-sequences—a maximum-likelihood approach. *J Mol Evol* **17**: 368–376.

Fournier-Level A, Le Cunff L, Gomez C, Doligez A, Ageorges A *et al.* (2009). Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. *ssp sativa*) berry: a QTL to QTN integrated study. *Genetics* **183**; e-pub ahead of print 31 August 2009.

Fuller DQ (2007). Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the old world. *Ann Bot* **100**: 903–924.

Gantenbein B, Fet V, Gantenbein-Ritter IA, Balloux F (2005). Evidence for recombination in scorpion mitochondrial DNA (Scorpiones: Buthidae). *Proc R Soc Lond B Biol Sci* **272**: 697–704.

Goor A (1966). The history of the grapevine in the Holy Land. *Economic Botany* **20**: 46–64.

Hasegawa M, Kishino H, Yano T (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**: 160–174.

Hocquigny S, Pelsy F, Dumas V, Kindt S, Heloir MC, Merdinoglu D (2004). Diversification within grapevine cultivars goes through chimeric states. *Genome* **47**: 579–589.

Hudson RR (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* **23**: 183–201.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–465.

Kass RE, Raftery AE (1995). Bayes factors. *J Am Stat Assoc* **90**: 773–795.

Kimura M (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press: Cambridge, 388pp.

Kobayashi S, Ishimaru M, Hiraoka K, Honda C (2002). Myb-related genes of the Kyoho grape (*Vitis labruscana*) regulate anthocyanin biosynthesis. *Planta* **215**: 924–933.

Kobayashi S, Goto-Yamamoto N, Hirochika H (2005). Association of *VvMybA*1 gene expression with anthocyanin production in grape (*Vitis vinifera*) skin-color mutants. *J Jap Soc Hortic Sci* **74**: 196–203.

Kuhner MK (2009). Coalescent genealogy samplers: windows into population history. *Trends Ecol Evol* **24**: 86–93.

Lacombe T, Laucou V, Di Vecchi M, Bordenave L, Bourse T, Siret R *et al.* (2003). Inventory and characterization of *Vitis vinifera ssp.silvestris* in France. *Proceedings of the 8th International Conference on Grape Genetics and Breeding* **1 and 2**: 553–557.

Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon AF *et al.* (2008). Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. *subsp sativa*. *BMC Plant Biol* **8**.

Levadoux L (1956). *Les Populations sauvages et cultivées de Vitis vinifera L. Annales de l'amélioration des plantes*, Vol 1. INRA: Paris, France, pp 59–119.

Lijavetzky D, Ruiz-Garcia L, Cabezas JA, De Andres MT, Bravo G, Ibanez A *et al.* (2006). Molecular genetics of berry colour variation in table grape. *Mol Genet Genomics* **276**: 427–435.

Martin DP, Posada D, Crandall KA, Williamson C (2005a). A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21**: 98–102.

Martin DP, Williamson C, Posada D (2005b). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**: 260–262.

Mazza G (1995). Anthocyanins in grapes and grapes products. *Crit Rev Food Sci Nutr* **35**: 341–371.

Maynard Smith J (1992). Analyzing the mosaic structure of genes. *J Mol Evol* **34**: 126–129.

Mc Govern PE (2003). *Ancient Wine the Search for the Origins of Viniculture*, 2nd edn. Princeton University Press: New Jersey, 400pp.

Molina-Cano JL, Russell JR, Moralejo MA, Escacena JL, Arias G, Powell W (2005). Chloroplast DNA microsatellite analysis supports a polyphyletic origin for barley. *Theor Appl Genet* **110**: 613–619.

Moncada X, Pelsy F, Merdinoglu D, Hinrichsen P (2006). Genetic diversity and geographical dispersal in grapevine clones revealed by microsatellite markers. *Genome* **49**: 1459–1472.

Moore RC, Purugganan MD (2005). The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol* **8**: 122–128.

Nei M, Tajima F (1987). Problems arising in phylogenetic inference from restriction-site data. *Mol Biol Evol* **4**: 320–323.

Niu TH (2004). Algorithms for inferring haplotypes. *Genet Epidemiol* **27**: 334–347.

Olalde M, Herran A, Espinel S, Goicoechea PG (2002). White oaks phylogeography in the Iberian Peninsula. *For Ecol Manage* **156**: 89–102.

Olmo HP (1976). Grapes. In: *Evolution of Crop Plants*. Longman: London, pp 294–298.

Özkan H, Brandolini A, Pozzi C, Effgen S, Wunder J, Salamini F (2005). A reconsideration of the domestication geography of tetraploid wheats. *Theor Appl Genet* **110**: 1052–1060.

Palaisa K, Morgante M, Tingey S, Rafalski A (2004). Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc Natl Acad Sci USA* **101**: 9885–9890.

Posada D, Crandall KA (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Posada D, Crandall KA (2002). The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* **54**: 396–402.

Qin ZHS, Niu TH, Liu JS (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* **71**: 1242–1247.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.

Sanderson MJ, Wojciechowski MF, Hu JM, Khan TS, Brady SG (2000). Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Mol Biol Evol* **17**: 782–797.

Schenk W (1992). Viticulture in Franconia along the River Main: human and natural influences since AD 700. *J Wine Res* **3**: 185–203.

Schierup MH, Hein J (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**: 879–891.

Schoen DJ, Brown AHD (1993). Conservation of allelic richness in wild crop relatives is aided by assessment of genetic-markers. *Proc Natl Acad Sci USA* **90**: 10623–10627.

Sefc KM, Lopes MS, Lefort F, Botta R, Roubelakis-Angelakis KA, Ibanez J *et al*. (2000). Microsatellite variability in grapevine cultivars from different European regions and evaluation of assignment testing to assess the geographic origin of cultivars. *Theor Appl Genet* **100**: 498–505.

Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S *et al.* (2008). Deletion in a gene associated with grain size increased yields during rice domestication. *Nat Gent* **40**: 1023–1028.

Stephens M, Scheet P (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76**: 449–462.

Suchard MA, Weiss RE, Sinsheimer JS (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol* **18**: 1001–1013.

Tajima F (1989). Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

Tanno K, Willcox G (2006). How fast was wild wheat domesticated? *Science* **311**: 1886.

Tenaillon MI, Tiffin PL (2008). The quest for adaptive evolution: a theoretical challenge in a maze of data. *Curr Opin Plant Biol* **11**: 110–115.

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

This P, Lacombe T, Cadle-Davidson M, Owens CL (2007). Wine grape (*Vitis vinifera* L.) color associates with allelic variation in the domestication gene VvMybA1. *Theor Appl Genet* **114**: 723–730.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS *et al.* (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39**: 31–40.

Unwin PTH (1991). *Wine and the Vine: an Historical Geography of Viticulture and the Wine Trade*. Routeledge: New York.

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

Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**: 256–276.

Zohary D, Hopf M (2000). *Domestication of Plants in the Old World*. Oxford University Press: London.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)