Heredity (2010) 104, 431–437

www.nature.com/hdy

ORIGINAL ARTICLE

Linkage disequilibrium in wild French grapevine, *Vitis vinifera* L. subsp. *silvestris*

A Barnaud[1], V Laucou, P This, T Lacombe and A Doligez
INRA, UMR 1097 DIAPC, Equipe Génétique Vigne, Montpellier, France

Association mapping based on linkage disequilibrium (LD) can provide high resolution for whole-genome mapping of genes underlying phenotypic variation. This field has received considerable attention over the last decade. We present here the first characterization of LD in wild French grapevine, *Vitis vinifera* L. subsp. *silvestris*. To assess the pattern and extent of LD, we used a sample of 85 plants from southern France and 36 microsatellite markers distributed over 5 linkage groups. LD was evaluated with independence tests and multiallelic r^2, using both unphased genotypic data and reconstructed haplotypic data. LD decayed rapidly, with r^2 values decreasing to 0.1 within 2.7 cM for genotypic data and within 1.4 cM for haplotypic data. Compared to the results of a previous study on cultivated grapevine subsp. *sativa*, where significant LD was found up to 16.8 cM, LD in subsp. *silvestris* was no longer significant past 1.4 cM. LD was therefore 12 times further extended in cultivated than wild grapevine, even though LD in wild grapevine seemed to extend slightly further than in wild relatives of other crops. Domestication bottlenecks and vegetative propagation are the primary factors responsible for this difference between cultivated and wild grapevine. The rapid decay of LD observed in this study seems promising for future association mapping studies of functional variation in wild *V. vinifera* grapevine.
Heredity (2010) **104,** 431–437; doi:10.1038/hdy.2009.143; published online 21 October 2009

Keywords: LD; association mapping; domestication; SSR; genetic diversity; *Vitis vinifera*

Introduction

Linkage disequilibrium (LD) is the nonrandom association of alleles at different loci and its extent can vary according to factors including mating system, genetic drift, natural and artificial selection, recombination rate, and population size and structure (Flint-Garcia *et al*., 2003). In cultivated plants, LD has received considerable attention as a tool for association mapping (Mackay and Powell, 2007), whole-genome scans and the testing of specific candidate genes. Association mapping, also known as LD mapping, estimates the position of quantitative trait loci (QTL) or major genes, based on the correlation between a trait and a marker (Mackay and Powell, 2007). LD mapping relies on germplasm samples and as such does not require the development of specific crosses with large progeny sizes. This is an obvious benefit for the study of long-cycled perennial species. However, the type of results that can be obtained through LD mapping depends on LD extent in the population studied. Efficient genome-wide mapping requires long-range LD, because a lower number of markers is needed. In contrast, efficient testing of candidate genes requires short-range LD (ideally not even extending to neighboring genes) to ensure the identification of causal polymorphisms (Kruglyak, 1999; Mackay and Powell, 2007). Recent studies in *Arabidopsis* (Nordborg *et al*., 2005), maize (Remington *et al*., 2001; Tenaillon *et al*., 2001), soybean (Hyten *et al*., 2007) and barley (Kraakman *et al*., 2004; Morrell *et al*., 2005) have provided contrasting results for LD fine mapping of QTL, or major genes, underlying phenotypic variation in plants. In crops where collections of elite, landrace and wild plants exist, the study of different populations or genome regions with contrasting LD extent (that is, resulting from different histories and biological features) might therefore allow both genome-wide and fine mapping.

Vitis vinifera L. is widely cultivated around the world and is economically the most important fruit species (Vivier and Pretorius, 2002). Furthermore, because of its small genome size and its diploidy, grapevine is particularly attractive for genomic research from QTL mapping to candidate genes studies (This *et al*., 2006). *V. vinifera* is an outcrossing heterozygous and perennial species subdivided into two subspecies: *sativa* (or *vinifera)*, the cultivated subspecies, and *silvestris* (or *sylvestris*), the wild subspecies, which is thought to be the wild progenitor of cultivated grapevines (Levadoux, 1956; Zohary and Hopf, 2000).

V. vinifera subsp. *silvestris* was abundant from the Atlantic coast of Europe to Tajikistan and the western Himalayas, as well as in northern Africa until the nineteenth century (Lacombe *et al*., 2003). Extensive anthropogenic habitat alteration and introduction of downy mildew, powdery mildew and phylloxera have led to population loss and decline. Presently, wild plants are restricted to small isolated populations along riverbank forests (Arnold, 1998; Zohary and Hopf,

Correspondence: Dr A Barnaud, INRA, UMR 1097 DIAPC, 2 place Viala, Equipe Génétique Vigne, Montpellier 34000, France.
E-mail: adelinebarnaud@hotmail.com
[1]Current address: Department of Botany and Zoology, Centre for Invasion Biology, Stellenbosch University, Matieland, Republic of South Africa.
Received 15 December 2008; revised 16 September 2009; accepted 22 September 2009; published online 21 October 2009

2000). *V. vinifera* subsp. *silvestris* is mainly dioecious, which is expected to result in low LD due to the elevated rate of recombination in a dioecious species. However, the loss and decline of populations due to habitat reduction and disease introductions have presumably increased LD.

In a previous study, we observed significant genotypic LD in cultivated grapevine, *V. vinifera* subsp. *sativa* (Barnaud *et al*., 2006). The large extent of LD observed (up to 16.8 cM, only within linkage groups (LGs)) suggests that genome-wide QTL mapping strategies exploiting LD could be effective in grapevine. Contrary to mapping in humans, where a very high density of markers is necessary (Kruglyak, 1999) except for LD blocks (Stumpf and Goldstein, 2003), in grapevine as in barley (Kraakman *et al*., 2004), whole-genome scans are likely to be a viable approach with fewer markers.

The objective of this study is to provide a first characterization of the extent and pattern of LD in the wild grapevine *V. vinifera* subsp. *silvestris*, to investigate the population genetic history of *V. vinifera* and to evaluate opportunities and conditions for LD mapping. To assess LD in wild *V. vinifera*, we studied 85 plants from Southern France. We analyzed LD between 36 simple sequence repeat (SSR) markers located in 5 LGs, with inter-SSR distances between 0 and 35.7 cM. To compare LD between wild and cultivated *V. vinifera*, we used the same SSRs and LD analyses as in our previous study (Barnaud *et al*., 2006).

Methods

Plant material

A total of 85 wild plants were collected from three locations in the south of France: Aquitaine (A), Languedoc-Roussillon (LR) and Midi-Pyrénées (MP). The plants were chosen according to previous analyses (Lacombe *et al*., 2003) to maximize the genetic diversity by limiting relatedness among individuals and to minimize the expected genetic structure. The plant numbers and geographic origins of samples are given in the Supplementary Data Table.

Genotyping

To assess the pattern and amount of genetic diversity, we chose the 20 unlinked SSRs (first set) that were used to genotype 2300 identified cultivars of *V. vinifera* maintained at the INRA experimental station of Vassal (Hérault, France, http://www.montpellier.inra.fr/vassal/). They have known map locations and are distributed throughout the 19 LGs (Laucou *et al*., unpublished data).

To assess LD, we used 36 SSRs (second set) previously used to characterize LD in *V. vinifera* subsp. *sativa* (Barnaud *et al*., 2006). These SSRs are distributed in five genomic regions of 17.9, 23.6, 12.7, 35.7 and 35.4 cM on LGs 4, 10, 11, 15 and 17, respectively. These genomic regions were chosen because they displayed a high density of mapped SSRs in progeny MTP3140 (Doligez *et al*., 2006). Two SSRs (VVIN85 and VVIV24) used for cultivated grapevine were excluded from this study because they were not polymorphic (the frequency of one allele was $\geqslant 95\%$) in the wild populations sampled.

DNA extraction and genotyping for both sets of SSRs were performed following methods described by Adam-Blondon *et al*. (2004) and Barnaud *et al*. (2006). Fragments of several loci with nonoverlapping sizes were amplified in multiplex using differentially labeled SSR primers and separated on an ABI PRISM 3100 Prism (Applied Biosystems, Foster City, CA, USA). Alleles were assigned using Genotyper software 2.5 (Applied Biosystems). To avoid scoring errors, we performed double reading of the profiles and data were manually checked for the presence of null alleles.

Diversity and structure analyses

Diversity and genetic structure were assessed using the data of the 20 unlinked SSRs (first set). Observed heterozygosity and gene diversity (expected heterozygosity) corrected for small sample size (Nei, 1978) were calculated using Genetix 4.04 software (Belkhir *et al*., 1996–2004).

To assess population structure, which may yield spurious LD between unlinked markers (Nei and Li, 1973; Pritchard and Przeworski, 2001), we used two complementary approaches: F-statistics and a Bayesian model-based clustering method. F_{st} values were computed among geographical populations using Genetix 4.04. Permutation procedures (10 000 permutations) were performed to test the significance of differences between values. As suggested for inferring population substructure at low levels of population differentiation (Latch *et al*., 2006), we used the Bayesian model-based clustering method of Pritchard *et al*. (2000) implemented in Structure 2.2 (http://pritch.bsd.uchicago.edu). This method assumes that each genotype in the sample may result from the admixture of an unknown number of differentiated ancestral populations. We used the basic admixture model with correlated allele frequencies, which is considered best in cases of low population structure (Falush *et al*., 2003). We assumed a number of populations (K) varying from 1 to 10, with 10 independent runs per K value. We used a burn-in period length of 10^7 followed by 1.5×10^6 MCMC steps, which allowed stability to be reached for statistical parameters and gave consistent results across runs in a pilot study. No *a priori* population information was used.

Linkage disequilibrium analyses

We performed the same analyses as in Barnaud *et al*. (2006) using the data of the 36 SSRs (second set). We assessed the extent and pattern of LD using both unphased genotypic data and haplotypic data reconstructed based on the assumption of coalescence. Although analyses of unphased genotypic data require no assumption about the genetic history or the genotype frequencies of the sample, they are expected to induce some loss in power compared with analyses of haplotypic data (Pritchard and Przeworski, 2001). LD was calculated after discarding, at each locus, alleles with a frequency lower than 5% in the total sample.

We first estimated the composite measure of LD (Δ), the sum of the intra- and intergametic disequilibria (Weir 1996) for all pairs of alleles for linked loci, using GDA V1.1 software (Lewis and Zaykin, 2002). We derived the genotypic multiallelic coefficient r^2_G from this composite measure (Barnaud *et al*., 2006). To test for

the significance of LD, we used Fisher's exact tests, as implemented in GDA. We also tested for loci in Hardy–Weinberg equilibrium using Genepop (Raymond and Rousset, 1995) (http://genepop.curtin.edu.au/genepop_op1.html). The variance components of LD within and among populations were computed following Otha (1982) using Linkdos software (Garniergere and Dillmann, 1992) (http://www.wbiomed.curtin.edu.au/genepop/linkdos.html).

We then inferred haplotypic data within each LG using a Bayesian method for reconstructing haplotypes from genotypic data (Stephens *et al*., 2001; Stephens and Donnelly, 2003), implemented in PHASE V2.1. The haplotypic multiallelic r^2_H correlation coefficient was directly estimated for all pairs of linked loci, with PowerMarker V3.25. To test for the significance of LD, we also performed Fisher's exact tests with the same software.

For the 630 independence tests, we applied the Bonferroni correction on an experiment-wise first-type error rate of 5% to adjust the critical probability for acceptance. The comparison-wise significance threshold was rounded up to 1×10^{-4}.

Results

Genetic diversity and population structure

All 20 loci of the first set of SSRs were polymorphic and revealed a total of 149 alleles. The average number of alleles per locus and per population are 6.2, 5.1 and 3.8 for A, LR and MP, respectively (Table 1). Observed and expected heterozygosity varied slightly among populations and ranged from 0.61 to 0.62 and from 0.62 to 0.64, respectively. For population A, the genetic diversity assessed with the 36 linked loci is slightly lower than expected (Table 1). We found a high percentage of rare alleles in both data sets: 55 and 48% for the first and the second data set, respectively. These rare alleles had an average frequency of 1%.

F_{st} statistics showed a weak structure in the sample. F_{st} was low but significant between A and LR ($F_{st} = 0.035$, $P < 0.01$) and between A and MP ($F_{st} = 0.022$, $P < 0.01$), but not significant between LR and MP ($F_{st} = 0.002$, $P > 0.05$). The F_{st} estimates involving MP should be interpreted cautiously, because this population contains only 12 individuals.

The Bayesian cluster analysis also supported partitioning of the genetic data into two geographic clusters, but the structure was weak. The estimated log probability of the data was highest for $K = 2$ and then decreased slowly with increasing K values. Only 20 and 11% of individuals were assigned to a cluster with a cluster membership over 80% for $K = 2$ and 3, respectively. Because a weak structure was revealed, we decided to perform LD analyses not only on the total sample, but also on the two subsamples A (50 individuals) and LR–MP (35 individuals), to assess the effect of this structure on LD.

Linkage disequilibrium

We estimated genotypic LD, r^2_G for all pairs of loci within LGs. Figure 1a shows the relationship between r^2_G and distance (cM) for the total sample. As expected, r^2_G declined with distance, down to 0.1 within 2.7 cM. There were few differences among LGs, with LG 4 showing the highest r^2_G value (0.48) (data not shown).

We performed 630 Fisher's exact tests to investigate the significance of genotypic disequilibrium between all loci. A total of 7, 2 and 2 significant associations (Figure 2) were recorded for the total data set, subsample A and

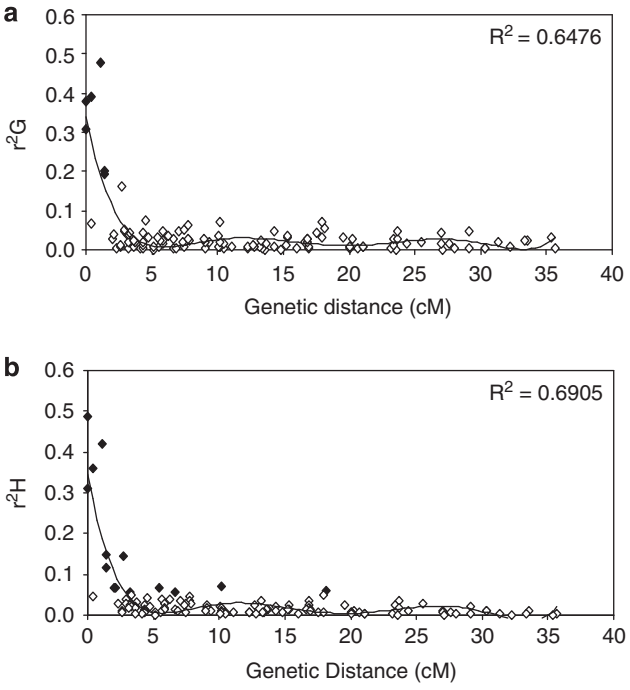

Figure 1 Multiallelic correlation coefficient as a function of distance (cM) in the total sample of *Vitis vinifera silvestris*: (**a**) Genotypic (r^2_G), (**b**) haplotypic (r^2_H). Filled diamonds, significant LD; open diamonds, nonsignificant LD. Black lines represent associated regression curves: polynomial of order six functions better fit our data.

Table 1 Number of plants per population, expected heterozygosity (corrected Nei index) with standard deviation in parentheses, observed heterozygosity with standard deviation in parentheses and mean number of alleles per locus

Populations	*No. of plants*	H_{exp}[a]	H_{obs}[a]	*MNA*[a]	H_{exp}[b]	*MNA*[b]
A	50	0.64 (0.12)	0.62 (0.14)	6.20	0.53 (0.18)	5.18
LR	24	0.62 (0.10)	0.61 (0.13)	5.10	0.60 (0.17)	4.84
MP	11	0.62 (0.10)	0.62 (0.10)	3.80	0.59 (0.18)	3.86
Total sample	85	0.65 (0.10)	0.65 (0.13)	7.45	0.58 (0.17)	6.42

Abbreviations: A, Aquitaine; H_{exp}, expected heterozygosity; H_{obs}, observed heterozygosity; LR, Languedoc-Roussillon; MNA, mean number of alleles; MP, Midi-Pyrénées.
[a]Diversity statistics assessed using the data of the 20 unlinked SSRs (first set).
[b]Diversity statistics assessed using the data of the 36 SSRs (second set).

subsample LR–MP, respectively. All involved locus pairs were genetically linked except one in the total sample (VVIH01 on LG 4 and VVP77 on LG 10). Significant genotypic LD extended up to 1.4, 1.4 and 0 cM in the total sample, and the A and LR–MP subsamples, respectively. The subsamples presented fewer loci in disequilibrium than the total sample. We also tested for Hardy–Weinberg equilibrium at each locus. A total of 11, 10 and 8 loci showed significant heterozygote deficiency for the total sample, subsample A and subsample LR–MP, respectively. No locus showed homozygote deficiency. Among the loci showing significant heterozygote deficiency, 6, 4 and 2 were also in LD for the total data set, A and LR–MP subsamples, respectively.

Following Otha (1982), we then estimated the spatial components of LD. For the majority of loci with significant genotypic LD, the within-population LD (D_{is}) is higher than the among-population LD (D_{st}) except for two pairs of loci out of the three pairs with significant LD only in the total sample. For instance, the pair of nongenetically linked loci, VVIP77/VVIH01 presented a D_{is} (0.005) lower than D_{st} (0.081). For these two pairs of loci, LD seems thus to result mainly from a structure effect.

Haplotypic LD analyses were based on reconstructed haplotypes within each LG. As for genotypic LD, the estimated r^2_H declined with distance, down to 0.1 within 2.7 cM (Figure 1b) and showed only minor differences among LGs (data not shown). r^2_H and r^2_G values were highly correlated (Spearman's correlation $r^2 = 0.68$, $P < 0.0001$).

As shown by Fisher's exact tests, significant haplotypic LD extended further than genotypic LD up to 7.1 cM for the total sample and 2.9 cM (A) and 3.2 cM (LR–MP), for the subsamples. There were slightly more significant haplotypic than genotypic LD between locus pairs. These differences are probably due to a larger power of independence test with haplotypes than with genotypes. All the linked locus pairs with significant genotypic LD also showed significant haplotypic LD.

Impact of population structure on LD estimates

To determine the cause of the differences in the numbers of locus pairs with significant LD between the total sample and subsamples, we resampled. We created 50 random samples of 50 individuals each, drawn from the total sample. We then performed one independent Fisher's exact test on each sample and averaged the *P*-values obtained over the 50 samples. A total of six significant associations were recorded. The six locus pairs involved also showed significant LD in the total sample.

Discussion

In natural populations, LD is expected to decay rapidly. In this study, r^2 values decreased to 0.1 within 2.7 cM/

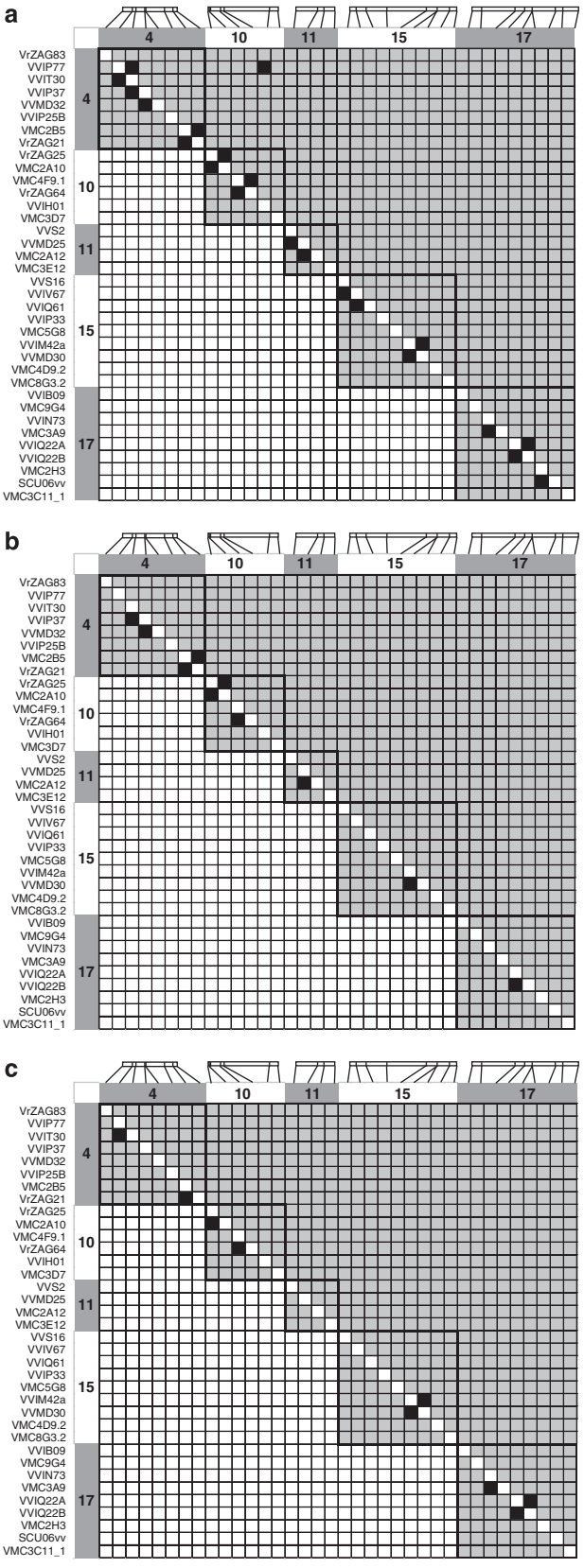

Figure 2 Linkage disequilibrium maps. Loci are sorted according to their map order. The upper right half of each matrix presents the results of the genotypic Fisher's exact tests, and the lower left half the results of the haplotypic exact tests within linkage groups only. Black squares, significant associations ($P < 1 \times 10^{-4}$); gray squares, nonsignificant associations; white squares, untested associations. Thick black lines delineate linked locus pairs. (**a**) Total sample, (**b**) A subsample and (**c**) LR–MP subsample. The numbers alongside the maps are for linkage groups.

351–583 kb for genotypic data and haplotypic data (based on an average estimate of 130–216 kb/cM for the correspondence between genetic and physical distances in grapevine; Adam-Blondon *et al*., 2005). Differences observed among samples were not necessarily due to a lack of statistical power in small subsamples. The spatial components of genotypic LD suggest that geographic structure provides the most plausible explanation for both the higher number of loci with significant LD in the total sample than in subsamples and for significant LD between unlinked loci (in the absence of rare epistatic interactions).

LD appears to be relatively high in the dioecious *V. vinifera* L. subsp. *silvestris* compared with other wild species, even if these studies relied on sequence data. LD declined very rapidly in *Helianthus annuus*, the self-incompatible wild sunflower (Liu and Burke, 2006), *Solanum peruvianum* and *S. chilense*, the self-incompatible wild tomatoes (Arunyawat *et al*., 2007), reaching negligible levels within 200, 150 and 750 bp, respectively (see also *Persea americana*, the outcrossing wild avocado; Chen *et al*., 2008). LD in our data set extends even further than in wild barley (Morrell *et al*., 2005) and in wild soybean (Hyten *et al*., 2007). Despite their high selfing rate, these two species showed a rapid decay within 300 bp and 36–77 kb, respectively. However, the comparison of our results with these other studies should be considered with caution, because LD between SSRs usually extends further than LD between SNPs (Flint-Garcia *et al*., 2003; Stich *et al*., 2006). Furthermore, our wild sample is relatively local compared with the large extent of the geographic area of the subspecies. A more global sample might yield lower physical LD and larger diversity, even though it would probably be more structured.

Pattern of LD in *V. vinifera* L. subsp. *silvestris* and insight in the evolutionary dynamics of grapevine

Why is LD extensively observed in wild grapevine despite the effective rate of recombination in this dioecious species? The extent of LD in wild grapevine results from the interplay of many factors. It can potentially be linked to population history, particularly recent bottlenecks. At the end of the nineteenth century, wild grapevine underwent a drastic reduction in diversity, owing to the disease-causing agents mildew and phylloxera (This *et al*., 2006), as shown by the relative genetic diversity deficit observed in wild grapevine compared to cultivated grapevine (Aradhya *et al*., 2003; Grassi *et al*., 2003). Our results also reveal such a deficit. Using the data set of 36 linked loci, genetic diversity (H_{exp}) was lower in wild grapevine than in the cultivated grapevine (core collection, Barnaud *et al*., 2006): 0.58 and 0.72, respectively. Thus, a whole-genome LD increase, associated with the decline of wild grapevine populations, is likely to have already occurred recently. Even though we are yet to observe isolation by distance (Wahlund effect) in wild adult plants, LD will increase further, through an elevated level of consanguinity linked to the low gene dispersal and small population size of wild grapevine (Di Vecchi-Staraz *et al*., 2009).

In a previous study, we reported LD analysis in a core collection of 141 grapevine cultivars (Barnaud *et al*., 2006). Although the sample sizes were different, LD significance results are comparable as the same SSR set was used, and no decrease in the power of Fisher's exact tests was noticed for sample sizes decreasing from 99 to 42 in Barnaud *et al*. (2006), or from 85 to 50 in this study. In the wild data set, significant genotypic LD was found up to 1.4 cM, whereas in the cultivated data set it was twelve times extended (16.8 cM), both in the total core collection (141 accessions) and in the subsample of wine and wine-table cultivars (99 accessions). Furthermore, comparisons are possible because we used exactly the same measure of LD (Pritchard and Przeworski, 2001). The rate of decay of genotypic LD was greater in the wild grapevine data set than in the cultivated one, with r^2_G values declining to around 0.1 within 2.7 and 5 cM, respectively. This difference in LD extent between wild and cultivated grapevine is even more noticeable as LD in a more global wild sample is expected to be even lower than in a local sample, due to less inbreeding.

Domestication of plants leads to a reduction in genetic diversity, thus gene history in domesticated plants involves bottlenecks (Vigouroux *et al*., 2005) that generate LD (Mackay and Powell, 2007). As expected, wild grapevine presented fewer significant LD associations between loci and a higher rate of LD decay with distance than cultivated grapevine. A similar pattern was found between wild and cultivated sunflower (*H. annuus*) (Liu and Burke, 2006) and in *Glycine soja*, the wild ancestor of soybean, LD did not extend past 100 kb, whereas in three cultivated *G. max* populations (landraces, North American ancestors and elite cultivars), LD extended from 90 to 574 kb (Hyten *et al*., 2007).

Domestication of grapevine probably took place in the Near East and Western Mediterranean regions (Arroyo-García *et al*., 2006). Domestication led to major changes in the mating system of the grapevine, with the selection of hermaphrodism and the development of vegetative propagation (Zohary and Hopf 2000). The greater LD observed in cultivated grapevine is consistent with a decrease in effective population size owing to the domestication bottleneck associated with the selection of hermaphrodite genotypes and the long-term process of selection of suitable genotypes producing larger and sweeter berries of attractive colors (This *et al*., 2006). Furthermore, vegetative propagation tends to limit recombination events, favoring the maintenance of the LD resulting from domestication. On the basis of historical records, many of the current grapevine varieties can be traced back hundreds of years (Bowers *et al*., 1999). According to these historical records, grapevine cultivars appear to be separated from their wild relatives by a low number of sexual generations, not higher than 80 (Arroyo-García *et al*., 2006). This low number of generations can explain the maintenance of such a large difference (LD twelve times extended) between cultivated and wild grapevines, although our data set was not designed to accurately resolve the demographic history associated with domestication. Comparisons between observed LD patterns with simulated data, obtained under scenarios with fluctuating demography and recombination rates, might be an interesting tool to resolve the relative importance of these parameters in grapevine evolution, as shown in other species (Bataillon *et al*., 2006).

Potential for association mapping analyses

The differences in LD extent between wild and cultivated grapevine might be of particular interest for application in LD mapping. The rate of decay of LD affects the resolution of LD mapping and the density of markers required to identify phenotype–genotype associations (Jorde, 2000; Buckler and Thornsberry, 2002; Nordborg *et al*., 2002; Kraakman *et al*., 2004). A whole-genome search for QTLs could be achieved with reasonable marker densities in populations of cultivated grapevine but the resulting confidence interval of QTL locations would be large. Therefore, QTL location could be refined by studying populations of wild grapevine with much more limited LD. Such a two-step procedure would rely on the assumptions that allelic diversity involved in phenotypic variation exists in both wild and cultivated populations and that the present results are representative of the whole-genome LD.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This research was funded by INRA, except prospecting and molecular analysis of wild *V. vinifera* with 20 SSRs, which were funded by the Ministry of Ecology through a grant of the BRG (Bureau des Ressources Génétiques). We are grateful to B Van Vuuren, D Spear and G Perdersen for revising the English version of this paper. We thank the editor and two anonymous reviewers for valuable suggestions.

References

Adam-Blondon A-F, 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.

Adam-Blondon A-F, Bernole A, Faes G, Lamoureux D, Pateyron S, Grando MS *et al.* (2005). Construction and characterization of BAC libraries from major grapevine cultivars. *Theor Appl Genet* **110**: 1363–1371.

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

Arnold C (1998). Situation de la vigne sauvage *Vitis vinifera* ssp. silvestris en Europe. *Vitis* **37**: 159–170.

Arroyo-García R, Ruiz-García L, Bolling L, Ocete R, López 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.

Arunyawat U, Stephan W, Städler T (2007). Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol Biol Evol* **24**: 2310–2322.

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

Bataillon T, Mailund T, Thorlacius S, Steingrimsson E, Rafnar T, Halldorsson MM *et al.* (2006). The effective size of the Icelandic population and the prospects for LD mapping: inference from unphased microsatellite markers. *Eur J Hum Genet* **14**: 1044–1053.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (1996-2004). *Genetix 4.05, logiciel sous Windows TM pour la génétique des populations*. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II: Montpellier, France.

Bowers J, Boursiquot JM, This P, Chu K, Johansson H, Meredith C (1999). Historical genetics: the parentage of Chardonnay, Gamay, and other wine grapes of northeastern France. *Science* **285**: 1562–1565.

Buckler ES, Thornsberry JM (2002). Plant molecular diversity and applications to genomics. *Curr Opin Plant Biol* **5**: 107–111.

Chen H, Morrell PL, de la Cruz M, Clegg MT (2008). Nucleotide diversity and linkage disequilibrium in wild avocado (*Persea americana* Mill.). *J Hered* **99**: 382–389.

Di Vecchi-Staraz M, Laucou V, Bruno G, Lacombe T, Gerber S, Bourse T *et al.* (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.

Doligez A, Adam-Blondon A-F, Cipriani G, Di Gaspero G, Laucou V, Merdinoglu D *et al.* (2006). An integrated SSR map of grapevine based on five mapping populations. *Theor Appl Genet* **113**: 369–382.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003). Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54**: 357–374.

Garniergere P, Dillmann C (1992). A computer-program for testing pairwise linkage disequilibria in subdivided populations. *J Hered* **83**: 239.

Grassi F, Imazio S, Failla O, Scienza A, Ocete Rubio R, Lopez MA *et al.* (2003). Genetic isolation and diffusion of wild grapevine Italian and Spanish populations as estimated by nuclear and chloroplast SSR analysis. *Plant Biol* **5**: 608–614.

Hyten DL, Choi I, Song Q, Shoemaker RC, Nelson RL, Costa JM *et al.* (2007). Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**: 1937–1944.

Jorde LB (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res* **10**: 1435–1444.

Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* **168**: 435–446.

Kruglyak L (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**: 139–144.

Lacombe T, Laucou V, Di Vecchi M, Bordenave L, Bourse T, Siret R *et al.* (2003). Contribution à la caractérisation et à la protection *in situ* des populations de *Vitis vinifera* L. ssp. *silvestris* (Gmelin) Hegi, en France. Quatrième colloque national du BRG, La Châtre 14–16 Octobre 2002. *Les Actes du BRG* **4**: 381–404.

Latch E, Dharmarajan G, Glaubitz J, Rhodes O (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conserv Genet* **7**: 295–302.

Levadoux L (1956). Les populations sauvages et cultivées de *Vitis vinifera* L. *Ann Amelior Plantes* **6**: 59–117.

Lewis PO, Zaykin D (2002). *Genetic Data Analysis (GDA): Computer Program for the Analysis of Allelic Data, Version 1.1*. Available at http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php.

Liu A, Burke JM (2006). Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics* **173**: 321–330.

Mackay I, Powell W (2007). Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* **12**: 57–63.

Morrell PL, Toleno DM, Lundy KE, Clegg MT (2005). Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. *Proc Natl Acad Sci USA* **102**: 2442–2447.

Nei M, Li WH (1973). Linkage disequilibrium in subdivided populations. *Genetics* **75**: 213–219.
Nei M (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583–590.
Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J *et al.* (2002). The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet* **30**: 190–193.
Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H *et al.* (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3**: e196.
Otha T (1982). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc Natl Acad Sci USA* **79**: 1940–1944.
Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
Pritchard JK, Przeworski M (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**: 1–14.
Raymond M, Rousset F (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Heredity* **86**: 248–249.
Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J *et al.* (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* **98**: 11479–11484.
Stephens M, Smith NJ, Donnelly P (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**: 978–989.
Stephens M, Donnelly P (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73**: 1162–1169.
Stich B, Maurer HP, Melchinger AE, Frisch M, Heckenberger M, Rouppe van der Voort J *et al.* (2006). Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. *Mol Breeding* **17**: 217–226.
Stumpf MPH, Goldstein DB (2003). Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr Biol* **13**: 1–8.
Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc Natl Acad Sci USA* **98**: 9161–9166.
This P, Lacombe T, Thomas MR (2006). Historical origins and genetic diversity of wine grapes. *Trends Genet* **22**: 511–519.
Vigouroux Y, Mitchell S, Matsuoka Y, Hamblin M, Kresovich S, Smith JSC *et al.* (2005). An analysis of genetic diversity across the maize genome using microsatellites. *Genetics* **169**: 1617–1630.
Vivier MA, Pretorius IS (2002). Genetically tailored grapevines for the wine industry. *Trends Biotechnol* **20**: 472–478.
Weir SB (1996). *Genetic Data Analysis II*. Sinauer Associates, Inc.: Sunderland, pp 91–138.
Zohary D, Hopf M (2000). *Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe, and the Nile Valley*, 3rd edn. Oxford University: New York.

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