Linked versus unlinked markers: multilocus microsatellite haplotype sharing as a tool to estimate gene flow and introgression

Molecular Ecology

Koopman, W.J.M.; Li, Y.; Coart, E.; Weg, W.E.; Vosman, B.J. et al

https://doi.org/10.1111/j.1365-294X.2006.03137.x

Molecular Ecology (2007) **16**, 243–256 doi: 10.1111/j.1365-294X.2006.03137.x

Linked vs. unlinked markers: multilocus microsatellite haplotype-sharing as a tool to estimate gene flow and introgression

WIM J. M. KOOPMAN,*§ YINGHUI LI,†§ ELS COART,‡§ W. ERIC VAN DE WEG,* BEN VOSMAN,* ISABEL ROLDÁN-RUIZ‡ and MARINUS J. M. SMULDERS*

**Plant Research International, Wageningen UR, PO Box 16, 6700 AA Wageningen, The Netherlands, †The National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI)/Key Laboratory of Germplasm & Biotechnology (MOA), Institute of Crop Science, Chinese Academy of Agricultural Sciences, 100081 Beijing, China, ‡Unit Plant, Institute for Agricultural and Fisheries Research (ILVO), Caritasstraat 21, 9090 Melle, Belgium*

Abstract

We have explored the use of multilocus microsatellite haplotypes to study introgression from cultivated (*Malus domestica*) into wild apple (*Malus sylvestris*), and to study gene flow among remnant populations of *M. sylvestris*. A haplotype consisted of alleles at microsatellite loci along one chromosome. As destruction of haplotypes through recombination occurs much faster than loss of alleles due to genetic drift, the lifespan of a multilocus haplotype is much shorter than that of the underlying alleles. When different populations share the same haplotype, this may indicate recent gene flow between populations. Similarly, haplotypes shared between two species would be a strong signal for introgression. As the expected lifespan of a haplotype depends on the strength of the linkage, the length [in centiMorgans (cM)] of the haplotype shared contains information on the number of generations passed. This application of shared haplotypes is distinct from using haplotype-sharing to detect association between markers and a certain trait. We inferred haplotypes for four to eight microsatellite loci on Linkage Group 10 of apple from genotype data using the program PHASE, and then identified those haplotypes shared between populations and species. Compared with a Bayesian analysis of unlinked microsatellite loci using the program STRUCTURE, haplotype-sharing detected a partially different set of putative hybrids. Cultivated haplotypes present in *M. sylvestris* were short (< 1.5 cM), indicating that introgression had taken place many generations ago, except for two Belgian plants that contained a haplotype of 47.1 cM, indicating recent introgression. In the estimation of gene flow, F_{ST} based on unlinked loci indicated small (0.032–0.058) but statistically significant differentiation between some populations only. However, various *M. sylvestris* haplotypes were shared in nearly all pairwise comparisons of populations, and their length indicated recent gene flow. Hence, all Dutch populations should be considered as one conservation unit. The added value of using sharing of multilocus microsatellite haplotypes as a source of population genetic information is discussed.

Keywords: conservation, gene flow, haplotype, introgression

Received 29 March 2006; revision received 3 July 2006; accepted 21 August 2006

Introduction

Some serious limitations are encountered when unlinked neutral markers are used to study interspecific hybridization or gene flow among populations of endangered species. The hybridization signal, as deduced from the occurrence of alleles of the parental species at a combination of unlinked loci across the nuclear genome, becomes difficult to detect in advanced crosses (Pritchard *et al*. 2000; Beaumont *et al*. 2001). As a result, it is often impossible to determine whether a given allele is the result of hybridization or reflects allele frequency differences as they exist in the separate

Correspondence: M.J.M. Smulders, Fax: +31317418094; E-mail: rene.smulders@wur.nl
§These authors contributed equally to this study.

species. In gene flow studies of endangered species, levels of differentiation among populations are often low compared to the intrapopulation variation (Hedrick 1999). Unfortunately, it is difficult to assess the biological relevance of F_{ST} values below 0.05. The statistical significance of low F_{ST} values often reflects the number of samples and the associated sampling variance. In addition, the value does not provide a clue as to whether the level of gene flow is low but steady, or has decreased strongly in the recent past. Migration rates based on F_{ST} values are incorrect for situations in which populations have recently declined in size or experienced fragmentation, as mutation-drift equilibrium has not been achieved (Neigel 2002; Pearse & Crandell 2004).

We propose an alternative approach based on the analysis of multilocus haplotypes. The haplotypes consist of combinations of alleles at several loci on the same chromosome. It is expected that such haplotypes will be shared by related individuals, and passed on to successive generations until recombination breaks up the association. The chance that the association between alleles at different loci will be broken is proportional to the recombination distance [centiMorgans (cM)] between the loci. This implies that if gene flow between two populations is frequent, one would expect to find some of the same haplotypes in both populations; if the populations have been isolated from each other for some generations, the same alleles might still be found in both populations, but in different haplotype combinations. As the breakdown of haplotypes through recombination occurs much faster than loss of alleles due to genetic drift, the lifespan of a multilocus haplotype is much shorter than that of the underlying alleles. We can therefore interpret the occurrence of identical haplotypes in different populations as a strong signal for recent genetic exchange between populations. Similarly, the presence of identical haplotypes in plants of different species is a strong signal of interspecies hybridization. Moreover, as the expected lifespan of a given haplotype depends on the strength of the linkage among the loci of which it is composed, we can use a set of loci at varying (recombination) distances along a chromosome to obtain information on the number of generations passed. This application of shared haplotypes is distinct from using haplotype-sharing to detect association between marker and trait.

In this study, we demonstrate the usefulness of this approach by studying population differentiation among wild apple populations from Flanders (northern region of Belgium) and The Netherlands, and levels of introgression herein from apple cultivars. Wild apple [*Malus sylvestris* (L.) Mill.] has become one of the most endangered tree species in Europe due to destruction of suitable habitats and conversion of open-type forest into mature forest types, which are too dark for this light-demanding species (Stephan *et al*. 2003). In the densely populated region of Flanders and The Netherlands, only scattered trees and small populations can be found nowadays (Coart *et al*. 2003, 2006). This has probably affected patterns of inter-population gene flow, a factor of special relevance for small populations of an insect-pollinated, obligate outcrossing species such as *M. sylvestris*, in which the lack of compatible pollen may completely prevent seed production. This has been demonstrated in several other outcrossing plant species, such as *Maianthemum canadensis* (Worthen & Stiles 1988), *Maianthemum bifolium* (Honnay *et al*. 2006) and *Scirpus maritimus* (Charpentier *et al*. 2000). In addition, as landscape fragmentation leads to patches characterized by a relatively high ratio of edge to habitat (Hargis *et al*. 1998), the chances for hybridization with cultivated apple (*Malus domestica* Borkh.) may have increased steadily. Although Coart *et al*. (2003) demonstrated that the wild and the cultivated gene pools could still clearly be differentiated using simple sequence repeat (SSR) and amplified fragment length polymorphism markers, recent chloroplast DNA data suggest that interspecific hybridization might be more common than initially believed (Coart *et al*. 2006).

A well-designed study to assess the value of haplotypes of linked loci compared to traditional approaches based on the analysis of unlinked loci is possible in apple, as 15 years of research have led to a large number of markers and a dense linkage map of *M. domestica* (Cevik & King 2002; Liebhard *et al*. 2002; Vinatzer *et al*. 2004; Gao *et al*. 2005a, b; Silfverberg-Dilworth *et al*. 2006). Coart *et al*. (2003) showed that these markers also amplify in *M. sylvestris*. In this study, we compared two groups of eight apple microsatellite markers each: a first group with microsatellites mapped to different chromosomes and a second group consisting of microsatellites located on chromosome (i.e. linkage group) 10 of the *M. domestica* linkage map.

Materials and methods

Plant material

We analysed a total of 159 wild apple [*Malus sylvestris* (L.) Mill.] trees (111 trees from six Dutch populations and 48 trees from two populations in Flanders), as well as 97 *Malus domestica* Borkh. genotypes (Table 1). The samples comprise virtually all known *M. sylvestris* trees in The Netherlands and two of the three largest Flemish populations (Fig. 1). The trees were identified and selected based on location ('old' forest which had been undisturbed for at least 100 years) as well as morphology (the presence of hairs on at least the lower leaf surface (Tutin *et al*. 1993)). Each population was sampled completely. The *M. domestica* samples were 19th and early 20th century cultivars from the CGN gene bank collection at PPO in Randwijk, The Netherlands. For the present study, the *M. domestica* samples were treated as one population. Most analyses were

Table 1 *Malus* accessions analysed in the present study

Species	Origin	Population	Code	No. of trees
M. sylvestris	The Netherlands	Drenthe	DR	34
		Veluwe	VE	6
		Winterswijk	WI	7
		Nijmegen	NY	14
		Sint Jansberg	JA	24
		Zeldersche Driessen	ZE	26
	Flanders (Belgium)	Voeren	WBVo	8
		Meerdaal	WBM	42
M. domestica	Cultivars (all analyses)		M.dom	41
	Additional cultivars (LD analyses only)		M.dom	56

Fig. 1 Geographical origin of the populations on the map of The Netherlands and Flanders. DR, Drenthe; VE, Veluwe; WI, Winterwijk; NY, Nijmegen; JA, Sint Jansberg; ZE, Zeldersche Driessen; WBVo, Voeren; WBM, Meerdaal. The first six populations are Dutch, the last two are Flemish.

performed with a random sample of 34 *M. domestica* cultivars plus seven reference cultivars (Cox, Discovery, Elstar, Fiesta, Golden Delicious, Jonathan, and Prima), which were included to calibrate marker sizes to other studies (Liebhard *et al*. 2002; Silfverberg-Dilworth *et al*. 2006), and to match the data generated in the two laboratories (Wageningen and Melle). These 'modern' cultivars are direct descendents of older cultivars and indeed shared some haplotypes (see below). In the linkage disequilibrium (LD) and haplotype analyses, 56 other cultivars were added to obtain a larger sample of cultivated haplotypes, and in this way, increase the probabilities of reconstructing shared haplotypes. Table S1 (Supplementary material) contains a list of all cultivars analysed.

DNA extractions

Young leaves were sampled in liquid nitrogen and stored at −80 °C. Before DNA extraction, the plant materials were freeze-dried and grinded in a ball mill. DNA was extracted using the DNeasy 96 Plant Mini Kit (QIAGEN) according to the manufacturer's instructions.

Microsatellite genotyping

PCR products were separated on an ABI PRISM 377 DNA Analyser or on a capillary AB 3130 DNA Analyser (Applied Biosystems). Allele sizes were estimated with GENOTYPER (Applied Biosystems), manually inspected, and transferred into a CONVERT-format spreadsheet (Glaubitz 2004).

Fifteen microsatellite loci were typed (Table S2, Supplementary material) to create the data sets of unlinked and linked loci. One set consisted of eight microsatellites located on various linkage groups, yielding data set DLG (i.e. different linkage groups) of unlinked markers. A second set contained eight microsatellites located on Linkage Group 10 of the *M. domestica* map (Silfverberg-Dilworth *et al*. 2006), yielding data set LG10 (used for linkage disequilibrium and haplotype analysis). Marker MS02a01 was present in both sets. The LG10 markers were selected at variable genetic distances, up to 47.1 cM apart. Figure 2c, d contain a schematic representation of the relative positions of the markers along LG10.

Data analysis

Linkage disequilibrium. We tested first, for each species, whether the linkage disequilibrium (LD) was indeed larger in the data set produced with linked markers (LG10) than in the data set with markers on different chromosomes (DLG). To obtain a greater accuracy, LG10 markers were also screened in 56 additional *M. domestica* genotypes (Table 1). The presence of LD was analysed using POPGENE 1.32 (Yeh & Wang 1999), as Burrow's composite measure of linkage disequilibrium between pairs of loci, and associated

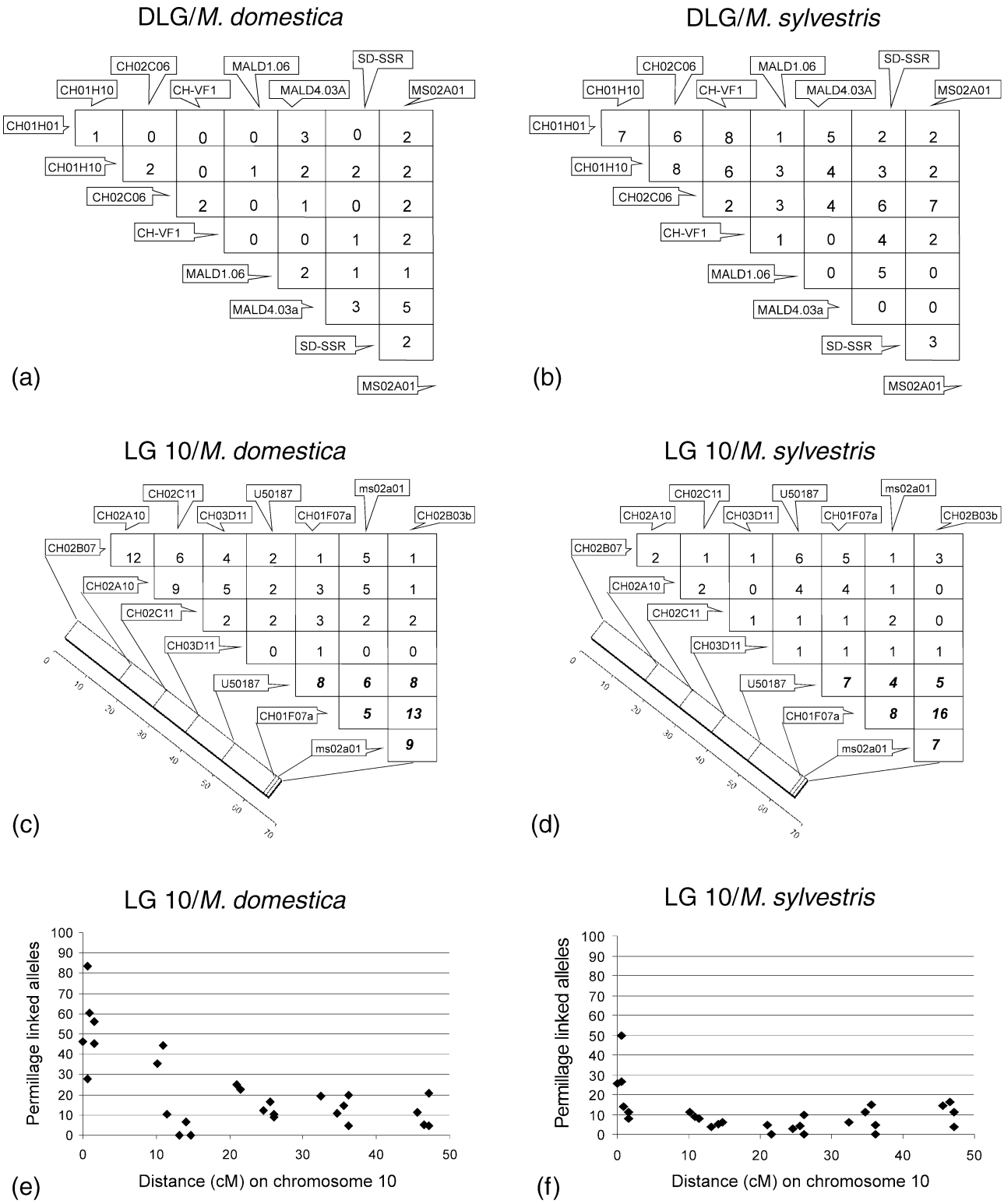

Fig. 2 Linkage disequilibrium in apple. Figure 2a–d — pairs of alleles at two loci in significant linkage disequilibrium: (a) in data set DLG (unlinked loci) for *Malus domestica*; (b) in data set DLG for *Malus sylvestris*; (c) in data set LG10 (linked loci) for *M. domestica*; and (d) in data set LG10 for *M. sylvestris*. The bar in 2c and 2d graphically depicts the relative position of the LG10 microsatellites on Linkage Group 10. Figure 2e–f — relationship between genetic distance (in cM) on Linkage Group 10 and the fraction of allele pairs in significant LD (in 0/00), across all pairs of LG10 loci.

χ^2 significance levels. Based on the results of the LD analyses (see Results section for details), data set LG10 was split into a data set with a group of loosely linked loci (Ch02b07, Ch02a10, Ch02c11, Ch03d11, spanning 32.4 cM; further referred to as data set LG10U), and one with a set of tightly linked loci (U50187, Ch01f07a, Ms02a01, Ch02b03b, spanning 1.5 cM; further referred to as data set LG10L).

Distinction of wild and cultivated apple using unlinked loci. First, a Bayesian model-based clustering approach as implemented in the software STRUCTURE (version 2) was applied (Pritchard *et al*. 2000; Falush *et al*. 2003). For analysis, a model with K populations is assumed, each of which is characterized by a set of allele frequencies at each locus. Individuals in the sample are assigned probabilistically to populations or jointly to two or more populations if their genotypes indicate that they are admixed. It is assumed not only that those loci are unlinked but also that within populations the loci are at Hardy–Weinberg equilibrium. As suggested by the authors, the data were firstly analysed using the admixture model. Alleles were considered correlated among populations. The analysis was run with K ranging from 1 to 15 and repeated 3 times for each value of K to check for consistency of results over different runs. Each run consisted of 1 000 000 burn-ins and 1 500 000 iterations. Even with this high number of iterations, it was difficult to decide on which K to choose. For several values of K (4–8) a similar estimate of log $\Pr(X/K)$ (i.e. the log likelihood of the model given that K clusters are present in the data) was obtained, described by Pritchard *et al*. (2000) as the often encountered 'more-or-less plateaus'. However, inspection of the results revealed that each inferred cluster consisted mainly of wild or mainly cultivated genotypes. In most cases, genotypes were also strongly assigned to clusters belonging to the wild or to the cultivated clusters. In this situation, where the genetic information is in rough agreement with the predefined populations, the STRUCTURE model that uses prior population information can be applied. Here we used cultivated apple samples as 'learning samples' to help us classify the individuals with unknown origin (trees sampled in the wild). K was set to 2 and Q (i.e. the admixture proportions for each individual) was inferred for all samples collected in the wild. This model normally improves the accuracy of the inference, as was shown by Beaumont *et al*. (2001). We applied this model, again with 1 000 000 burn-ins and 1 500 000 iterations, which resulted in consistent results for log $\Pr(X/2)$ over different runs, supporting our hypothesis of two distinct genetic units.

The results from both approaches (admixture and prior population information model) were very similar, with a slightly stronger assignment of genotypes to the wild or cultivated cluster in the latter model. This suggests that the assumed prior on the domestics is reasonable and we therefore only present results from this latter model.

We further ranked the values of $\hat{q}$ (assignment of genetic information to the cultivated cluster) and plotted these ranks against $\hat{q}$, illustrating the distribution of $\hat{q}$ among individuals. Inflexions and break-points in this plot were used to define groups of genuine wild genotypes and hybrids, as in Coart *et al*. (2006). For analysis of the population genetic structure among wild populations, the admixture model was applied with alleles considered correlated and for K set from 1 to 15. For each value of K, again three independent runs were done with 1 000 000 burn-ins and 1 500 000 iterations.

Second, the overall presence of genetic differentiation between *Malus domestica* and *Malus sylvestris* was tested using FSTAT (Goudet 1995) in its revised version 2.9.3.2. Prior to data analysis, the putative hybrids detected in the STRUCTURE analysis described above were removed from the data set. F_{ST} was calculated according to Weir & Cockerham (1984). The significance of the differentiation statistics was determined by bootstrapping over loci. The overall population differentiation was estimated as the log-likelihood G statistic of Goudet *et al*. (1996). Similar analyses were carried out to estimate population differentiation among *M. sylvestris* populations. Observed and expected heterozygosities and the effective number of alleles (N_e) were calculated with POPGENE 1.32.

Distinction of wild and cultivated apple using linked loci. STRUCTURE version 2 implemented a linkage model to deal with information derived from linked loci to infer population structure (Falush *et al*. 2003). This is essentially a generalization of the admixture model to deal with admixture linkage disequilibrium, i.e. the correlations that arise between linked markers in recently admixed populations. This linkage model was applied on data set LG10L containing four tightly linked SSR loci. Three independent runs were done for K ranging from 1 to 15, with 1 000 000 burn-ins (500 000 admixture burn-in length) and 1 500 000 iterations. Allele frequencies were assumed correlated among populations.

Haplotype-sharing analysis. Unlike *Arabidopsis*, which is nearly completely homozygous so that genotypes can be compared directly as haplotypes (e.g. Toomajian *et al*. 2006), apple is an obligatory outbreeder with a genetically controlled self-incompatibility system. Haplotype-based studies in outbreeders typically follow a two-step procedure: first, haplotypes are inferred from phase-unknown genotypes using a computational algorithm; in a second step, inferred haplotypes are fed into the multilocus analysis where they basically are treated as having been directly observed (Schouten *et al*. 2005). The latter step introduces some uncertainty (Schouten *et al*. 2005), and hence possible error (see Discussion). We inferred haplotypes using PHASE (Stephens *et al*. 2001) in its revised version (Stephens

& Donnelly 2003). We tested five sets of loci spanning an increasingly larger part of Linkage Group 10 (1.5, 14.7, 26.1, 36.2 and 47.1 cM, for sets containing 4–8 loci, respectively). For each set, 20 runs with different starting points were performed. Each run consisted of 500 final iterations, 100 burn-in iterations, and a thinning interval of 10. We employed the parent-independent mutation model for all loci. The run with the highest overall likelihood was saved as the final result. We then identified haplotypes that were shared between species and between populations.

Within each haplotype, each cM distance between two markers corresponds to a 1% chance that a marker allele at one locus was separated from a marker allele at another locus due to crossing-over (recombination) during meiosis in the cross between the commercial cultivars 'Fiesta' and 'Discovery' (Silfverberg-Dilworth *et al*. 2006), i.e. in a single generation. As no linkage map of *M. sylvestris* has been published to date, we have assumed that the genetic linkage of the microsatellite loci in this species is similar to that of the *M. domestica* cross used to produce a segregating population. The haplotype results in this study are consistent with this order of markers. Unfortunately, the estimation of average survival time (in number of generations) of a haplotype across generations is not a straightforward extrapolation of recombination frequency within a generation, as the decay of LD due to actual recombinations is influenced by historical population size, population structure, and the occurrence of selection (Toomajian *et al*. 2006). If we assume a constant recombination rate per generation, the probability P that a given haplotype did not change from its ancestor G generations ago is $P = (1 - r)\exp(G)$, with r = recombination and mutation rate (1 in Stephens *et al*. 1998). When we assume r is comparable to the map recombination distance (cM/100), then 50% of the original haplotypes will theoretically be lost after 46 generations for the set of 4 loci (spanning 1.5 cM), 5 generations for the set of 5 loci (14.7 cM), 3 generations for the set of 6 loci (26.1 cM), and 2 generations for the set of 7 (36.2 cM) or 8 loci (47.1 cM). This gradual decrease in survival time for the different 'haplotype lengths' was used to reconstruct patterns of gene flow between *M. domestica* and *M. sylvestris*, and among the different *M. sylvestris* populations. We analysed the occurrence and proportion of shared haplotypes between the cultivar group and the wild populations, and between pairs of wild populations, respectively. In the former case, nearly all shared haplotypes had a (much) higher frequency in *M. domestica*, and they were therefore assumed to indicate introgression into *M. sylvestris*.

Comparison of results obtained with unlinked SSR loci and sharing of haplotypes. To compare the results on hybrid identification, we determined the proportion of genotypes carrying 'cultivated haplotypes' for each of the groups defined by the STRUCTURE analysis of unlinked markers: 'pure' *M. sylvestris* trees and hybrids with *M. domestica*. 'Cultivated haplotypes' were defined as haplotypes that occur at least once among *M. domestica* accessions.

To compare the results on levels of population differentiation in *M. sylvestris*, the pairwise F_{ST} and its significance (based on unlinked markers) was compared to two measures of haplotype-sharing between populations: the longest haplotype shared and the average haplotype length shared. Average haplotype length was calculated as the sum of the length (in cM) of haplotypes from population A that were found in population B divided by the number of plants in population A (this measure depends on the size of population A, and is therefore not symmetric).

Results

General characteristics of the SSR loci analysed

Levels of observed and expected heterozygosity estimated for the different loci were similar in *Malus sylvestris* and *Malus domestica*. One exception was locus Mald4.03a, for which the observed and expected heterozygosities were much lower in wild than in cultivated apple. For all loci positive or slightly negative F_{IS} values were estimated in both groups of samples (Table S2, Supplementary material).

Extent of linkage disequilibrium across the M. domestica *and the* M. sylvestris *genomes*

A total of 37% of allele pairs at two loci were in significant LD for data set DLG (loci on different linkage groups) in *M. domestica*, and three times more (117%) for data set LG10. For the much larger sample of *M. sylvestris*, the overall fraction of allele pairs in LD was similar for both sets of loci: 94% for DLG and 86% for LG10. These differences between *M. domestica* and *M. sylvestris* may reflect the relatively closer relationships among cultivars than among *M. sylvestris* genotypes. The pedigree relationships among the cultivars from the gene bank collection are not very well known but they are probably quite extensive, just as nowadays. For example, the modern cultivars 'Elstar', 'Fiesta', 'Gala' and 'Prima' have 'Golden Delicious' as parent or grandparent, while 'Elstar', 'Fiesta' and 'Gala' are also related through 'Cox'.

Figure 2(a–d) show, for all pairwise comparisons of loci, the number of allele combinations that were in significant LD. Linkage disequilibrium on Linkage Group 10 was due, for the major part, to the four loci that are located within 1.5 cM of each other: Ch02b03b, Ms02a01, Ch01f07a, and U50187. Among these loci, LD of the pair Ch02B03b/Ch01F07a stands out for both *M. domestica* and *M. sylvestris*. Among the remaining LG10 loci, the pairs Ch02A10/Ch02B07 and Ch02A10/Ch02C11 show relatively high LD in *M. domestica*, but not in *M. sylvestris*.

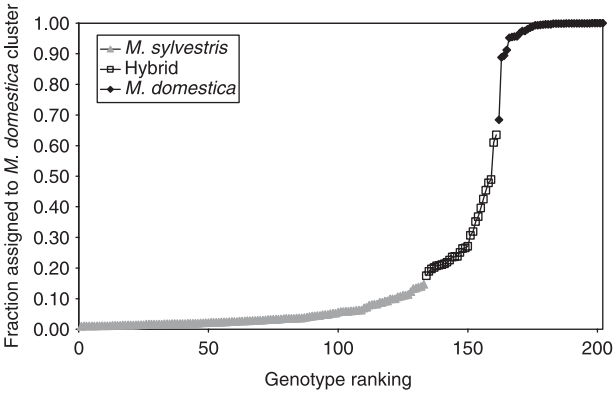

Fig. 3 Distinction of wild and cultivated apple, expressed as the proportion of a given genotype assigned to the cultivated cluster in a STRUCTURE analysis of the combined data sets DLG + LG10U. This assignment proportions have been ranked and the ranks are plotted against the assigned proportions. ▲: *Malus sylvestris*; □: putative hybrid genotypes; ◆: *Malus domestica* cultivars.

When the fraction pairs of alleles in LD were plotted against the genetic distance between pairs of alleles along the linkage group (Fig. 2e–f), a strong decrease in LD was observed with increasing recombination distance, as can be expected, but with some differences between *M. domestica* and *M. sylvestris*: (i) the amount of LD among the four closely linked loci is generally higher for *M. domestica* than for *M. sylvestris*; (ii) the relatively high LD found in *M. domestica* for locus pairs Ch02A10/Ch02B07 and Ch02A10/Ch02C11 at ~10 cM distance is not present in *M. sylvestris*; and (iii) the level of LD for pairs of less closely linked loci is relatively even for both species, but slightly higher for *M. domestica* (0–30%) than for *M. sylvestris* (0–20%).

Based on these results, we subdivided data set LG10 into a data set with a group of loosely linked loci (Ch02b07, Ch02a10, Ch02c11, Ch03d11; data set LG10U), and one with a set of tightly linked loci (U50187, Ch01f07a, Ms02a01, Ch02b03b; data set LG10L). Species and population differentiation were examined using DLG plus LG10U markers.

Interspecific relationships and population differentiation parameters using unlinked loci

The STRUCTURE results of the relationship between *M. domestica* and *M. sylvestris* using 12 SSR loci (sets DLG + LG10U) showed a clear assignment of most samples to either the 'cultivated' or the 'wild' cluster (Fig. 3). All but four *M. domestica* genotypes (Prima, Discovery, Groninger Peppeling and Zigeunerin) were assigned to the cultivated cluster for more than 95%. All genotypes sampled in nature as *M. sylvestris* had an assignment proportion to the cultivated cluster of less than 65%. The distribution of assignment proportions to the cultivated gene pool in Fig. 3 follows an exponential curve. Small gaps were observed between the values 0.15 and 0.17 and between 0.65 and 0.68. Based on this information, we tentatively distinguished two groups of genotypes within the noncultivated samples: 'truly wild' genotypes (assignment proportions to the cultivated gene pool < 0.15), and 'putative hybrids' (assignment between 0.17 and 0.65). The distribution of the putative hybrids across populations is fairly even; however, the numbers are low (Table S3, Supplementary material).

Between-species F_{ST}, based on these 12 markers and after elimination of 'hybrids', was 0.10 ($P < 0.001$), indicating a fair amount of interspecific differentiation. F_{IS} values indicate a slight deficit of heterozygotes for most *M. sylvestris* populations. *M. domestica* showed the smallest deficit. The number of loci not in HW equilibrium was small in all cases (Table S4, Supplementary material).

STRUCTURE analyses of population structure among *M. sylvestris* populations showed that differences in Pr(K) were small and that the individuals were evenly assigned to all populations (results not shown). These observations are in agreement with a lack of clear population structure in the *M. sylvestris* complete data set. Similar conclusions can be derived from F_{ST} estimates. The overall F_{ST} among *M. sylvestris* populations, based on 12 markers (set DLG + LG10U), indicated a low but significant overall population differentiation ($F_{ST} = 0.03$, $P < 0.001$). Pairwise F_{ST} values (Table 2) showed that statistically significant differentiation was restricted to comparisons involving populations DR and WBM, while all but one of the other population comparisons were nonsignificant. Moreover, these pairwise differences were only significant at the 5% level.

For comparison, STRUCTURE was also run with the linkage model on the four linked SSR loci of data set LG10L. However, the differences in Pr(K) were small and the assignment of individuals to the inferred groups was roughly even (results not shown). As expected, information on four microsatellites, representing only one set of linked markers, is not sufficient for STRUCTURE to discriminate between the two apple species.

Interspecific relationships and population differentiation based on shared haplotypes

In the first step of haplotype-sharing analysis, the PHASE program inferred haplotypes for 4–8 loci along LG10. For example, 82 different 8-loci haplotypes were inferred among the 97 *M. domestica* plants analysed (representing a total of 194 chromosomes or haplotypes). Among the 159 *M. sylvestris* plants analysed (318 chromosomes), 206 different haplotypes were inferred. On average therefore, in *M. domestica* the same haplotype was found in 2.37 (194/82) accessions, and in *M. sylvestris* in 1.54 accessions. Twenty-one 8-loci haplotypes were found in both species, while 61 were private for *M. domestica* and 185 were private for *M. sylvestris*.

Table 2 Population differentiation in *Malus sylvestris*

	DR	VE	WI	NY	JA	ZE	WBVo	WBM
DR	8	6	5	6	8	6	4	5
VE	0.016	6	No	6	8	7	No	No
WI	0.037	0.02	No	6	6	6	No	5
NY	0.023	0.033	0.013	6	7	7	No	6
JA	0.042*	0.038	0.015	−0.007	7	6	No	7
ZE	0.037*	0.039	0.018	−0.006	0.029	8	5	5
WBVo	0.056*	0.061	0.066	0.041	0.042	0.056*	5	6
WBM	0.047*	0.058*	0.043*	0.026	0.033*	0.032*	0.033	8

Below diagonal: pairwise F_{ST} values and significance of the log-likelihood G statistic as measures of differentiation between *M. sylvestris* populations (based on the 12 microsatellite loci of sets DLG + LG10U). *indicates significant population differentiation at the 5% level after standard Bonferroni correction. Diagonal and above: longest LG10 haplotype that was shared within (diagonal) and between (above diagonal) populations. Population designations are according to Table 1. No, no alleles shared between plants within populations (diagonal) or between populations (above the diagonal).

In the second step, pairs of plants were identified that shared the same haplotype. Figure 4 shows how many haplotypes were shared within and between groups, for haplotypes consisting of 4, 5, 6, 7, and 8 loci, expressed as percentage of shared haplotypes, summed over all plants in a given population. Overall, the proportion of shared haplotypes decreased rapidly with haplotype length, consistent with a shorter survival times of 'longer' haplotypes (spanning larger genetic distance).

The first row of panels shows haplotype-sharing within *M. domestica* (grey bars) and of *M. domestica* with *M. sylvestris* populations (black bars). Note that the scale of this row is different from all other rows, due to the fact that as much as 70% of the 4-loci haplotypes (grey bar in the first panel) occurred in more than one cultivar. Haplotypes that also occur in *M. sylvestris* (all black bars in this row) were relatively rare, and nearly completely limited to haplotypes of 4 and 5 loci. In *M. sylvestris* populations (all other rows) haplotype-sharing within populations was always 50% or less. The difference can partly be due to differences in population size, the *M. domestica* sample being much larger than any of the *M. sylvestris* populations. Note that we sampled all trees in the *M. sylvestris* populations, the small and unequal sample sizes reflect the endangered situation of the species.

The first bar in the other rows of panels (DR-WBM) of Fig. 4 indicates haplotypes in *M. sylvestris* populations that also occurred in *M. domestica*, expressed as a percentage of all haplotypes in *M. sylvestris* populations. The differences in sample size will affect these data to some extent, but some patterns are visible across population sizes. These haplotypes may have originated from *M. domestica* and have been introgressed into *M. sylvestris* populations, which is supported by the fact that they are more frequent in *M. domestica* than in *M. sylvestris*. This influence of *M. domestica* in *M. sylvestris* varied from completely absent to a little above 10%. The *M. domestica* haplotypes in WI, NY, and WBVo covered 4–5 loci, which may be indicative of old introgression events. The haplotypes in DR, JA, and ZE covered up to 6 loci. In WBM, two haplotypes of 8 loci occurred in two different plants. These plants could therefore be first generation hybrids, but also offspring of hybrids in which LG10 has not yet been destroyed by recombination.

The proportion of shared haplotypes among the *M. sylvestris* populations was rather variable (Fig. 4). The smallest population, VE, shared up to 50% of the shortest haplotypes with other populations. Longer haplotypes were also shared, indicating recent gene flow among VE and the other populations. A similar picture, but with smaller proportions of shared haplotypes, emerged for other Dutch populations. The two populations from Flanders shared a fair amount of haplotypes with each other but only few haplotypes with most of the Dutch populations. Although differences in sample size exist, the effect is possibly not so large, as there is no direct relationship between sample size and degree of haplotype-sharing.

The 6–8 loci haplotypes have a predicted half-life of only 2–3 generations therefore sharing of such long *M. sylvestris*-specific haplotypes between populations may indicate recent gene flow between these populations or with a third population (that may not exist any more). This is a likely case for populations that are geographically close to each other (NY with JA and ZE). Also DR shared some long haplotypes with JA, and WBM with WBVo. Conversely, the 4-locus haplotype has a long half-life of 46 generations, and therefore absence of shared 4-loci haplotypes may be taken as an indication of effective isolation for a long period of time. Absence of sharing of haplotypes was only found in the smallest populations (WBVo, VE, and WI), although the chance of finding shared haplotypes here is the lowest due to sample size.

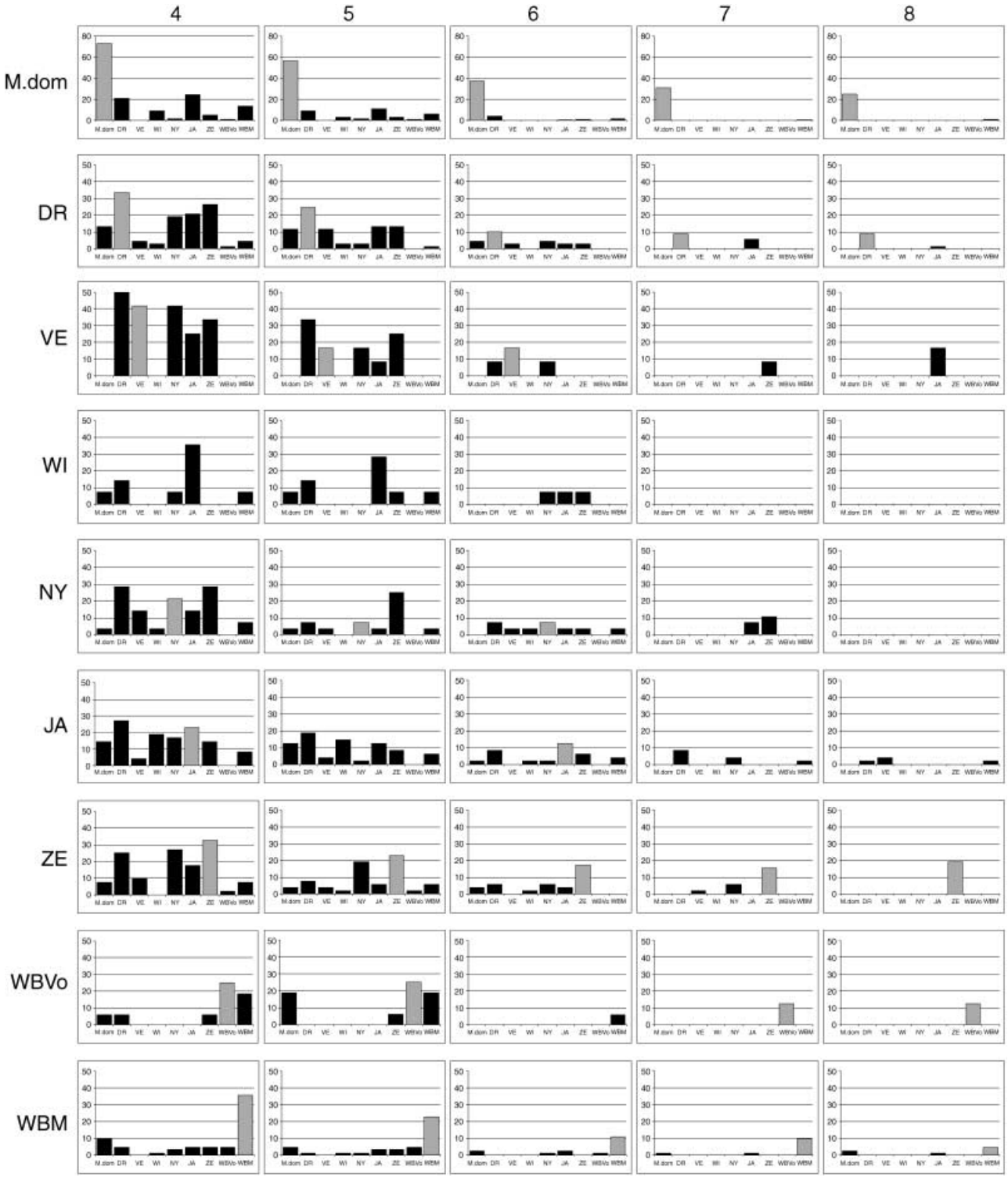

Fig. 4 Shared haplotypes among *Malus domestica* and the various wild *Malus sylvestris* populations, for haplotypes consisting of 4, 5, 6, 7, and 8 loci on LG10, expressed as percentage of shared haplotypes, summed over all plants in a given population. In each panel, the shaded bar represent haplotype-sharing within the same population, filled bars indicate haplotypes shared with each of the other populations. Note the deviating scale in the first row (M.dom, *M. domestica*). Abbreviations for *M. sylvestris* populations are as in Fig. 1.

Table 3 Comparison of the analyses with unlinked loci (DLG + LG10U) and analyses based on LG10 haplotypes (data set LG10L). Distinction of *Malus sylvestris* and putative hybrid plants based on the STRUCTURE classification vs. that based on the presence/absence of cultivar haplotypes

Classification of *M. sylvestris* and putative hybrid plants based on STRUCTURE groups	Presence of *Malus domestica* haplotypes inferred by PHASE		
	Total no. of plants	Yes (hybrid)	No (*M. sylvestris*)
Total no. of plants	161	24	137
Hybrid	28	7	21
M. sylvestris	133	17*	116

*of these 17 plants, three were homozygous for the cultivar haplotype.

Comparison of analyses using linked and unlinked markers

The *M. sylvestris* accessions can be divided into two groups according to the presence or absence of cultivated (*M. domestica*) haplotypes on LG10, or on the basis of the STRUCTURE results using unlinked markers. In Table 3, both methods are compared. Only seven plants were considered as hybrid by both methods, while an additional 21 plants were identified by STRUCTURE and 17 by the presence of cultivar haplotypes. Strikingly, three accessions homozygous for cultivar haplotypes were assigned to the wild group using unlinked markers. It may be that these plants had inherited the cultivar-specific LG10 haplotype from a *M. domestica* introgression some generations ago, but not a number of cultivar-specific alleles on the other chromosomes, or that these have been lost in successive generations. Since the STRUCTURE results are based on markers on various chromosomes, and the haplotype analysis on data from LG10, this may reflect the complicated inheritance of an *M. domestica* introgression in further generations.

Table 2 presents, next to the pairwise F_{ST} values, also the longest shared haplotype between pairs of populations. The general patterns are comparable, i.e. the lower the F_{ST} value, the longer the haplotype shared. However, some differences may be biologically relevant. Population WBM was significantly differentiated from five of the six Dutch populations according to the *F* statistics but did share long (6–7 loci) haplotypes with most of them. Population DR was also significantly differentiated from most other populations according to F_{ST} but did share long haplotypes (DR and JA even shared an 8-locus haplotype). Conversely, the pairwise F_{ST} values of WBVo were not significant, probably due to small population size; but with the haplotype analysis, it turned out that this population shared the least haplotypes in the whole study, suggesting that this is the most isolated population.

Table 4 Average length of LG10 haplotypes shared within and between populations of *Malus sylvestris*. The average length (in cM) is calculated as the total length of haplotypes found shared in the same (diagonal) or with other populations divided by the number of haplotypes in the population shared from. Note that the table is not symmetric, i.e. if two plants in the 'shared with' population share the same haplotype, it counts for two in the 'shared from' population. In bold: values in the lower triangle that are below 1.5 cM/plant (= on average one shortest haplotype per plant)

	Shared with							
	DR	VE	WI	NY	JA	ZE	WBVo	WBM
Shared from								
DR	6.83	2.06	0.43	1.39	3.91	2.48	0.02	0.26
VE	6.1	4.73	0	3.78	9.33	6.82	0	0
WI	2.1	**0**	0	1.97	5.12	1.86	0	1.05
NY	2.19	1.09	0.99	2.08	2.75	7.08	0	0.99
JA	4.9	2.58	2.44	1.73	3.42	2.03	0	1.86
ZE	2.05	**0.85**	**0.5**	4.18	**1.46**	10.6	0.28	0.88
WBVo	**0.09**	**0**	**0**	**0**	**0**	**0.92**	9.56	3.47
WBM	**0.23**	**0**	**0.18**	**0.35**	**1.06**	**0.54**	0.84	6.22

A slightly different statistic is the average length of haplotypes shared expressed in cM, which was calculated both within and among populations (Table 4). Some populations had very high levels of shared haplotypes within the population, whereas other populations with comparable current population sizes had quite different levels of haplotype-sharing (e.g. compare 3.42 cM for JA with 10.6 cM for ZE). Between-population haplotype-sharing was of course much lower, although VE-JA and VE-ZE shared 9.33 cM and 6.82 cM, which clearly indicates that these populations were in contact through a considerable level of gene flow.

Discussion

In this study, we have compared the use of microsatellite markers spread across different linkage groups with a new approach that uses microsatellite markers on one linkage group. Using markers in LD, we have statistically inferred the underlying multilocus haplotypes and followed the occurrence of shared haplotypes across populations as a tool to detect recent introgressions and gene flow that occurred up to a few generations ago. The multilocus haplotypes ranged from 1.5 to 47.1 cM, which is much longer than in typical association studies.

Linkage disequilibrium

For both *Malus domestica* and *Malus sylvestris,* the amount of LD rapidly decreased with increasing distance between

the loci along linkage group (LG) 10. The high LD found in *M. domestica* along LG10 most likely reflects the breeders' simultaneous selection on several traits located on this chromosome. The effect is apparent from the higher LD on the majority of the *M. domestica* locus combinations, but most obvious for locus pairs Ch02A10-Ch02B07 and Ch02A10-Ch02C11. In close proximity to these loci, resistances against scab (Calenge *et al*. 2004; Tartarini *et al*. 2004) and powdery mildew (Calenge & Durel 2006) are located, as well as multiple NBS containing disease resistance gene analogs (Baldi *et al*. 2004; Calenge *et al*. 2005). It is not known whether *M. sylvestris* is polymorphic for these traits. The locus pair CH01F07a-CH02B03b shows an increased LD in both *M. domestica* and *M. sylvestris*. It is located close to a QTL for fruit firmness (King *et al*. 2000; Maliepaard *et al*. 2001), and to an ACO gene associated with shelf life (Costa *et al*. 2005), two traits which are of interest for an apple breeder but which are unlikely to be selected for in *M. sylvestris*. Rather, a more likely explanation for LD observed is that the two loci are only 0.6 cM apart. Note that positive selection due to breeding would increase the frequency of the selected haplotypes among cultivars (and possibly also increase their length through 'genetic drag'), and that more frequent cultivar haplotypes would be transmitted more frequently to *M. sylvestris*, but their presence in *M. sylvestris* populations would still be indicative of gene flow, as we do not take into consideration the 'identity' of the cultivated haplotype.

Heterozygosity

An alternative explanation of the increased LD in the cultivated apple is that most selection and breeding effort is done within a small group of genitors, so that the same haplotypes occur in many closely related cultivars. Indeed, in apple breeding and selection procedures, favourable phenotypes generally come from mass selection in older cultivars, or through a combination of 'plus' alleles from different parents on different homologous chromosomes in modern breeding practice. This will increase the level of heterozygosity. Indeed, the level of observed heterozygosity in cultivars (0.648 for 8 markers, 0.708 for 12 markers) is higher than that in *M. sylvestris* (0.554 and 0.656, respectively). For one marker (Mald1.06 A), the observed heterozygosity is even higher than expected (Table S2, Supplementary material).

For *M. domestica*, the general statistics showed a relatively high heterozygosity, while the haplotype reconstruction showed that more than 57% (110/194) of the haplotypes were shared among *M. domestica* cultivars. Thus, we detect more alleles as well as higher heterozygosity in fewer haplotypes compared to *M. sylvestris*. Indeed, cultivars do share haplotypes, but are rarely homozygous for one haplotype (e.g. only 2 out of 97 cultivars were homozygous for the 5-loci LG10 haplotype, which is half of the rate in *M. sylvestris* (5/111)). This may be due to selection by breeders in *M. domestica*, or a certain degree of loss of diversity in the *M. sylvestris* populations, or a combination of both.

In *M. sylvestris*, the haplotype-sharing analysis is generally consistent with heterozygosity levels. For example, in population WI, with the highest observed heterozygosity, no plants homozygous for LG10 haplotypes were found and no haplotypes were shared between plants within the population. Population WBVo, with the lowest observed heterozygosity, has one plant homozygous for LG10 haplotypes and a high proportion of haplotypes that are shared between plants within the population.

Haplotype inference uncertainty

Inferring haplotypes from genotypes using statistical algorithms ('phasing') leads to uncertainty (Kimmel & Shamir 2005; Schouten *et al*. 2005) which may affect the accuracy of using shared haplotypes as measure for gene flow or introgression. Here we used the most likely haplotype pair for each genotype as inferred by PHASE, but this ignores the uncertainties due to measurement errors, population haplotype frequencies (many equifrequent haplotypes increase the uncertainty), and the fact that populations may not be in Hardy–Weinberg equilibrium, which leads to biased estimates towards the null in association mapping (discussed in Kraft *et al*. 2005). The PHASE analysis did produce the correct haplotypes for the modern cultivars in our data set (van de Weg *et al*. unpublished data). However, as the cultivars are on average more related (57.7% shared 8-loci haplotypes in *M. domestica* compared to 35.2% in *M. sylvestris*), this does not imply that also all *M. sylvestris* haplotypes are correct. In fact, we observed that haplotype switching did occur sometimes between the inferred 8-loci haplotype and shorter haplotypes in the same pair of *M. sylvestris* plants. Kimmel & Shamir (2005) compared the accurateness of PHASE with their program GERBIL. Switch error rate for PHASE was slightly lower (at the cost of much longer calculation times) and varied among data sets that differed in number and density of SNP markers and in number of samples from 1 to 12% (the latter value in a set with small sample size and high (8%) missing data rate; the missing data rate in our microsatellite data set was 3.77%). Note that, as the number of hypothetical haplotypes is very high, most switch errors will infer nonexisting haplotypes. Therefore, the net effect of errors will be to reduce haplotype-sharing between pairs of plants, and hence to reduce the theoretical power of the analysis, but this is unlikely to lead to false inference of shared haplotypes. The haplotype-sharing analysis as we performed here can be extended to include several inferred haplotypes per plant with their likelihoods, but it remains to be determined

whether the increased complexity of the analyses results in an increased power of detecting shared haplotypes.

Hybridization between wild and cultivated apple

In the STRUCTURE analysis of unlinked microsatellite markers, the cultivars and the sampled *M. sylvestris* plants formed two distinct groups, showing that the *M. sylvestris* morphotypes are also genetically different from *M. domestica*. This is consistent with Coart *et al*. (2003) results. However, 17% of the *M. sylvestris* genotypes had intermediate probabilities in the STRUCTURE analysis and were identified as putative hybrids. Seven of these plants did share *M. domestica* haplotypes on LG10, but the majority (21) did not. At the same time, the haplotype analysis identified 17 additional *M. sylvestris* plants with cultivar LG10 haplotypes, which had not been identified as hybrids using unlinked microsatellite markers. The difference was not due to allele frequency differences between populations, as the three larger populations contained putative hybrids according to STRUCTURE as well as plants with cultivar haplotypes.

At first sight, this comparison does not tell us which of the haplotype analysis or the unlinked markers analysis is the best. If, however, we realize that the introgression signal varies across chromosomes (as is observed in hybrid zones, Rieseberg *et al*. 1999) and among plants (due to segregation), then sampling more of the genome will necessarily lead to the identification of more plants with an introgression signal, even if small remnant cultivated DNA chunks will still remain undetected. It is promising that the haplotype analysis that focuses on one linkage group identified the same groups and a comparable number of additional hybrid plants as the one that uses markers across linkage groups.

An original application of a shared haplotype analysis is the dating of the introgression events. We interpreted the STRUCTURE analysis with unlinked markers as indicating a possible recent origin for most of the *M. sylvestris*/*domestica* hybrids (some plants have a probability consistent with a BC1). However, we did not find recently shared haplotypes (i.e. haplotypes longer than 4 markers or 1.5 cM, Fig. 2) except for a very few cases, notably two plants with an 8-loci haplotype (spanning 47.1 cM) in population WBM. The latter plants very likely were derived from a recent introgression. It is possible that the cultivars that were the source of introgression were not included in this study, and thus haplotypes found in *M. sylvestris* were not identified as cultivated haplotypes. Alternatively, the introgression events may be very old, as the half-life of the 4-locus haplotypes is estimated at 46 generations. The generation time of *M. sylvestris* is unknown. As *M. domestica* seedlings took about 10 years to first flowering prior to breeding efforts aimed at improving earliness, and Austerlitz *et al*. (2000) used a factor of 1.5–3 between age of first flowering and the corresponding generation time of a hypothetical tree species with no overlapping generations, we can estimate the generation time of *M. sylvestris* at 15–30 years. In that case, the 1.5 cM haplotype could be 700–1400 years old, which goes back into mediaeval times.

Population structure within M. sylvestris

Although the population differentiation among the eight *M. sylvestris* populations was statistically significant (with $F_{ST} = 0.03$), 97% of the variation was present within the individuals trees and within the populations. Pairwise F_{ST} values indicated that DR and WBM were significantly different from most of the other populations, while all other pairs were not, although some of the nonsignificant F_{ST} values were in the same range. This may be due to the lower number of plants in these small populations. Given these results, the Dutch and Flemish populations can be considered as separate but closely related groups, which is consistent with their geographical origin (see Fig. 1). Within The Netherlands, population DR seems to contain some unique variation that may warrant a separate position from the other Dutch populations. Again, this position is consistent with geographical origin. Our results do not support any other separation of populations within The Netherlands or Belgium.

The results from the haplotype analyses contained evidence of recent gene flow among all Dutch populations, including DR, and between the Belgian population WBM and the Dutch populations as well. They do not support the small but significant genetic differentiation of DR and WBM, and suggest that all Dutch populations should be treated as one conservation unit. At the same time, the genetic isolation of WBVo relative to the Dutch populations was clear through the absence of haplotype-sharing, even though F_{ST} values were high but not statistically significant for this small population.

Compared to general measures of allelic variation, heterozygosity, and inbreeding, the haplotype reconstruction has the advantage that it enables a reconstruction of heterozygosity and inbreeding on a time scale. For example, populations JA and ZE have comparable coefficients of allelic richness and heterozygosity, with ZE being slightly more variable than JA. The haplotype reconstructions however, revealed a high level of haplotype-sharing between plants of the ZE population (on average 10.6 cM per plant, which is the highest value for *M. sylvestris*). Within population JA it was only 3.4 cM. This large difference may indicate that the history of these populations is different, with JA having been a much larger population than ZE during the period when the current trees were generated, or a larger effective population size.

Conclusions

Our analysis of sharing of inferred multilocus microsatellite haplotypes from linked loci yields general information on population structure and population differentiation that is largely comparable with the results obtained with unlinked loci, such as *F* statistics and Bayesian analysis. The strength of the method may be in generating information on the temporal scale of processes, which is important when population sizes have varied in the past (as is often the case in human-influenced and fragmented landscapes), and when the current population does not resemble that from which the remaining plants have been derived (as is often the case in long-lived plants such as trees). In addition, identifying a small number, or even one long haplotype shared between two populations indicates gene flow in a qualitative way, even when F_{ST} analysis, based on quantitative differences in allele frequencies, is not conclusive because of low sample sizes. In our study, the haplotype-sharing information on the apple populations led to a slightly different view of conservation units for *Malus sylvestris* than based on traditional unlinked marker F_{ST} analysis alone, and indicated different population history for populations that are now comparable in size. As for introgression, multilocus haplotypes made it possible to distinguish past and recent hybridization events. These examples demonstrate the usefulness and added value of sharing of multilocus haplotypes from linked microsatellite markers as a source of population genetic information.

Acknowledgements

The study was funded by The Netherlands' Ministry of Agriculture, Nature, and Food Quality (DLO Program 382), the Ministry of the Flemish Community, Forest and Green Areas Administration and the Belgian Federal Science policy (http://www.belspo.be/belspo/fedra/proj.asp?l = en&COD = EV/28). We thank Hennie Ketelaar, Chris Rövekamp, Bert Maes, and René van Loon for collecting the material, SBB for supporting the collection and sustaining the collected material in the gene bank 'Bronnen voor Nieuwe Natuur', and Nancy Mergan and Sabine Van Glabeke for excellent laboratory assistance. YL was supported by an IAC fellowship of The Netherlands' Ministry of Agriculture, Nature, and Food Quality.

Supplementary material

The supplementary material is available from
http://www.blackwellpublishing.com/products/journals/suppmat/MEC/MEC3137/MEC3137sm.htm

Table S1 List of *Malus domestica* cultivars included in the analyses.

Table S2 Microsatellite markers used and diversity detected. Note that marker MS02a01 on LG10 is part of both data sets. CentiMorgans (cM) is the genetic distance from the first marker on Linkage Group 10 according to the genetic map in Silfverberg-Dilworth *et al*. (2006). Identical letters in Column 'Multiplex PCR' indicates that the corresponding primer pairs were amplified in a single PCR reaction.

Table S3 Number of hybrid genotypes in the *Malus sylvestris* populations, according to the STRUCTURE results. Putative hybrids are trees with 0.17–0.65 probability of assignment to the cultivated gene pool, as shown in Fig. 3. Population designations are according to Table 1.

Table S4 General measures of allelic variation, heterozygosity, and inbreeding based on eight DLG markers. Non-HW = number of loci showing a significant departure from Hardy–Weinberg equilibrium at the 5% level. Population designations are according to Table 1.

References

Austerlitz F, Mariette S, Machon N, Gouyon PH, Godelle B (2000) Effects of colonization processes on genetic diversity: differences between annual plants and tree species. *Genetics*, **154**, 1309–1321.

Baldi P, Patocchi A, Zini E *et al.* (2004) Cloning and linkage mapping of resistance gene homologues in apple. *Theoretical and Applied Genetics*, **109**, 231–239.

Beaumont M, Barrett EM, Gottelli D *et al.* (2001) Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, **10**, 319–336.

Calenge F, Durel CE (2006) Both stable and unstable QTLs for resistance to powdery mildew are detected in apple after four years of field assessments. *Molecular Breeding*, **17**, 329–339.

Calenge F, Faure A, Goerre M *et al.* (2004) Quantitative trait loci (QTL) analysis reveals both broad-spectrum and isolate-specific QTL for scab resistance in an apple progeny challenged with eight isolates of *Venturia inaequalis*. *Phytopathology*, **94**, 370–379.

Calenge F, Van der Linden CG, Van de Weg E *et al.* (2005) Resistance gene analogues identified through the NBS-profiling method map close to major genes and QTL for disease resistance in apple. *Theoretical and Applied Genetics*, **110**, 660–668.

Cevik V, King J (2002) High-resolution genetic analysis of the Sd-1 aphid resistance locus in *Malus* spp. *Theoretical and Applied Genetics*, **105**, 346–354.

Charpentier A, Grillas P, Thompson JD (2000) The effects of population size limitation on fecundity in mosaic populations of the clonal macrophyte *Scirpus maritimus* (Cyperaceae). *American Journal of Botany*, **87**, 502–507.

Coart E, Van Glabeke S, De Loose M, Larsen AS, Roldán-Ruiz I (2006) Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*M. domestica* Borkh.). *Molecular Ecology*, **15**, 2171–2181.

Coart E, Vekemans X, Smulders MJM *et al.* (2003) Genetic variation in the endangered wild apple [*Malus sylvestris* (L.) Mill.] in Belgium as revealed by amplified fragment length polymorphism and microsatellite markers. *Molecular Ecology*, **12**, 845–857.

Costa F, Stella S, Van de Weg WE *et al.* (2005) Role of the genes *Md-ACO1* and *Md-ACS1* in ethylene production and shelf life of apple (*Malus domestica* Borkh.). *Euphytica*, **141**, 181–190.

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.

Gao ZS, Van de Weg WE, Schaart JG *et al.* (2005a) Genomic cloning and linkage mapping of the *Mal d 1* (PR-10) gene family in apple (*Malus domestica*). *Theoretical and Applied Genetics*, **111**, 171–183.

Gao ZS, Van de Weg WE, Schaart JG *et al.* (2005b) Genomic characterization and linkage mapping of the apple allergen genes *Mal d 2* (thaumatin-like protein) and *Mal d 4* (profilin). *Theoretical and Applied Genetics*, **111**, 1087–1097.

Glaubitz JC (2004) CONVERT: a user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes*, **4**, 309–310.

Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices. (Version 2.9.3.2).

Goudet J, Raymond M, Demeeus T, Rousset F (1996) Testing differentiation in diploid populations. *Genetics*, **144**, 1933–1940.

Hargis CD, Bissonette JA, David JL (1998) The behavior of landscape metrics commonly used in the study of habitat fragmentation. *Landscape Ecology*, **13**, 167–186.

Hedrick PW (1999) Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution*, **53**, 313–318.

Honnay O, Jacquemyn H, Roldán-Ruiz I, Hermy M (2006) Consequences of prolonged clonal growth on local and regional genetic structure and fruiting success of the forest perennial *Maianthemum bifolium*. *Oikos*, **112**, 21–30.

Kimmel G, Shamir R (2005) GERBIL: genotype resolution and block identification using likelihood. *Proceedings of the National Academy of Sciences, USA*, **102**, 158–162.

King GJ, Maliepaard C, Lynn JR *et al.* (2000) Quantitative genetic analysis and comparison of physical and sensory descriptors relating to fruit flesh firmness in apple (*Malus pumila* Mill.). *Theoretical and Applied Genetics*, **100**, 1074–1084.

Kraft P, Cox DG, Paynter RA, Hunter D, De Vivo I (2005) Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. *Genetic Epidemiology*, **28**, 261–272.

Liebhard R, Gianfranceschi L, Koller B *et al.* (2002) Development and characterisation of 140 new microsatellites in apple (*Malus* × *domestica* Borkh.). *Molecular Breeding*, **10**, 217–241.

Maliepaard C, Silanpää MJ, Van Ooijen JW, Jansen RC, Arjas E (2001) Bayesian versus frequentist analysis of multiple quantitative trait loci with an application to an outbred apple cross. *Theoretical and Applied Genetics*, **103**, 1243–1253.

Neigel JE (2002) Is F_{ST} obsolete? *Conservation Genetics*, **3**, 167–173.

Pearse DE, Crandell KA (2004) Beyond F_{ST}: analysis of population genetic data for conservation. *Conservation Genetics*, **5**, 585–602.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rieseberg LH, J.Whitton K, Gardner (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**, 713–727.

Schouten MT, Williams CKI, Haley CS (2005) The impact of using related individuals for haplotype reconstruction in population studies. *Genetics*, **171**, 1321–1330.

Silfverberg-Dilworth E, Matasci CL, Van de Weg WE *et al.* (2006) Microsatellite markers spanning the apple (*Malus* × *domestica* Borkh.) genome. *Tree Genetics and Genomes*, in press.

Stephan BR, Wagner I, Kleinschmit J (2003) EUFORGEN Technical guidelines for genetic conservation and use for wild apple and pear (*Malus sylvestris* and *Pyrus pyraster*). International Plant Genetic Resources Institute, Rome, Italy. http://www.ipgri.cgiar.org/publications/pubfile.asp?ID_PUB=922.

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction. *American Journal of Human Genetics*, **73**, 1162–1169.

Stephens JC, Reich DE, Goldstein DB *et al.* (1998) Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *American Journal of Human Genetics*, **62**, 1507–1515.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Tartarini S, Gennari F, Pratesi D *et al.* (2004) Characterisation and genetic mapping of a major scab resistance gene from the old Italian apple cultivar 'Durello di Forlì'. *Acta Horticulturae*, **663**, 129–133.

Toomajian C, Hu TT, Aranzana MJ *et al.* (2006) A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biology*, **4**, e137.

Tutin TG, Heywood VH, Burges NA *et al.* (1993) *Flora Europaea*, Vol. 2, pp. 66–67. Cambridge University Press, Cambridge, UK.

Vinatzer BA, Patocchi A, Tartarini S *et al.* (2004) Isolation of two microsatellite markers from BAC clones of the *Vf* scab resistance region and molecular characterization of scab-resistant accessions in *Malus* germplasm. *Plant Breeding*, **123**, 321–326.

Weir BS, Cockerham CC (1984) Estimating *F* statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Worthen WB, Stiles EW (1988) Pollen-limited fruit set in isolated patches of *Maianthemum canadense* Desf. in New Jersey. *Bulletin of the Torrey Botanical Club*, **155**, 299–305.

Yeh FC, Wang R-C (1999) POPGENE, version 3.31. Department of Renewable Resources, University of Alberta, Alberta, Canada.

Wim Koopman in interested in the use of genetic data to resolve phylogenetic and forensic issues. Yinghui Li is researcher at the Institute of Crop Science, Chinese Academy of Agricultural Sciences. Her main research interests are genetic diversity, evolution and core collection construction and conservation in soybean. Els Coart is a postdoc working in the research group led by Isabel Roldán-Ruiz, who is senior scientist at ILVO-Plant. Their main research interests are (agro)biodiversity, the relationship between crops and their wild relatives and the genome organization of disease resistance and reproduction-related genes in different species. Marinus Smulders and Ben Vosman are senior researchers at Plant Research International who share a common interest in the use of various marker methods to study neutral and functional (agro)biodiversity.
