Postglacial recolonization history of the European crabapple (*Malus sylvestris* Mill.), a wild contributor to the domesticated apple

A. CORNILLE,*†‡ T. GIRAUD,*†‡ C. BELLARD,*†‡ A. TELLIER,§ B. LE CAM,¶
M. J. M. SMULDERS,** J. KLEINSCHMIT,†† I. ROLDAN-RUIZ‡‡ and P. GLADIEUX*†‡§§

*CNRS, Laboratoire Ecologie Systématique et Évolution, UMR8079, Bâtiment 360, 91405, Orsay, France, †University of Paris Sud, 91405, Orsay, France, ‡AgroParisTech, 91405, Orsay, France, §Section of Population Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, 85354 Freising, Deutschland, ¶INRA, IRHS, PRES UNAM, SFR QUASAV, Rue G. Morel, F-49071, Beaucouzé, France, **Plant Research International, Wageningen UR Plant Breeding, PO Box 16, 6700 AA, Wageningen, The Netherlands, ††Department of Forest Genetic Resources, Northwest German Forest Research Institute, Professor-Oelkers Str. 6, 34346 Hann. Münden, Germany, ‡‡ILVO, Plant—Growth and Development, Caritasstraat 21, 9090, Melle, Belgium, §§Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720-3102, USA

Abstract

Understanding the way in which the climatic oscillations of the Quaternary Period have shaped the distribution and genetic structure of extant tree species provides insight into the processes driving species diversification, distribution and survival. Deciphering the genetic consequences of past climatic change is also critical for the conservation and sustainable management of forest and tree genetic resources, a timely endeavour as the Earth heads into a period of fast climate change. We used a combination of genetic data and ecological niche models to investigate the historical patterns of biogeographic range expansion of a wild fruit tree, the European crabapple (*Malus sylvestris*), a wild contributor to the domesticated apple. Both climatic predictions for the last glacial maximum and analyses of microsatellite variation indicated that *M. sylvestris* experienced range contraction and fragmentation. Bayesian clustering analyses revealed a clear pattern of genetic structure, with one genetic cluster spanning a large area in Western Europe and two other genetic clusters with a more limited distribution range in Eastern Europe, one around the Carpathian Mountains and the other restricted to the Balkan Peninsula. Approximate Bayesian computation appeared to be a powerful technique for inferring the history of these clusters, supporting a scenario of simultaneous differentiation of three separate glacial refugia. Admixture between these three populations was found in their suture zones. A weak isolation by distance pattern was detected within each population, indicating a high extent of historical gene flow for the European crabapple.

Keywords: approximate Bayesian computation, ecological niche modelling, fruit tree, intraspecies diversification, microsatellite

Received 4 July 2012; revision received 11 December 2012; accepted 13 December 2012

Introduction

The climatic oscillations of the Quaternary Period clearly affected the distribution and genetic structure of many species (Hewitt 2004; Lascoux *et al.* 2004). In Europe, many temperate species faced with the expansion of ice sheets over the northern part of the continent survived periods of glaciation in small ice-free spots in Southern and south-eastern Europe (called glacial refugia), in which conditions were more favourable (Hewitt 2004; Schmitt 2007). These range contractions were followed
by range expansions during the interglacial periods, with populations at the northern boundaries of refugia expanding into the large territories further north, which became increasingly hospitable as the climate rapidly warmed. Such shifts in distribution following rapid climate change are excellent models for understanding the mechanisms by which intra- and interspecific genetic differentiation arises and is maintained (Hewitt 1990, 1996, 2004; Petit et al. 2004; Schmitt 2007; Excoffier et al. 2009; Petit & Excoffier 2009). Identification of the geographical area that served as refugia that existed during the last glacial maximum (LGM) can also help set priorities for the conservation and management of genetic resources (Hampe & Petit 2005). Furthermore, understanding past population dynamics is also important to make predictions of the effects of current climate change on species distribution and survival (Petit et al. 2004, 2008; Nogués-Bravo et al. 2008; Provan & Bennett 2008).

Trees are fascinating models for investigating the impact of climatic fluctuations on species colonization, adaptation, speciation and extinction (Petit & Hampe 2006; Savolainen & Pyhájärvi 2007). Retrospective analyses based on paleoecological and genetic data have greatly increased our understanding of tree phylogeography (Hewitt 2000; Davis & Shaw 2001; Lascoux et al. 2004; Petit et al. 2004; Stewart et al. 2010). A combination of population genetic studies and explicit species distributions determined from fossil pollen and macrofossil evidence has revealed that trees recolonized higher latitudes during the Holocene, following expansion from southern glacial refugia (Brewer et al. 2002; Cheddadi et al. 2006; Magri et al. 2006), but also from small populations that persisted in more northern ‘cryptic’ refugia (Anderson et al. 2006; Parducci et al. 2012). However, additional studies of other species are required for more general conclusions to be drawn regarding the impact of historical contingencies and differences in climatic niches (from southern temperate to southern boreal) and mechanisms of gene dispersal on interspecific differences in postglacial histories. In particular, the postglacial evolutionary histories of wind-dispersed tree species growing in high-density populations have been extensively documented (e.g. Petit et al. 2002; Heuertz et al. 2006), but fewer studies have focused on low-density, animal-dispersed fruit trees (Palmé & Vendramin 2002; Oddou-Muratorio et al. 2004; Jolivet & Degen 2011). Animal-dispersed species would be expected to have stronger genetic structures and weaker dispersal capacities than wind-dispersed species, whereas low-density species would be expected to have higher dispersal capacities than high-density species (Vokemans & Hardy 2004; Hardy et al. 2006). It remains unclear how the balance between these two life history traits characteristics of fruit trees—animal mediated dispersal and a low density—as well as the importance of outcrossing shapes the genetic diversity and genetic structure of populations during post-LGM recolonization.

Most previous studies investigating the phylogeography of European tree species were based on chloroplast DNA sequences. An advantage is that maternally inherited markers often display stronger differentiation between populations than biparental markers, due to their smaller effective size and lack of recombination. However, it has repeatedly been argued that this lack of recombination together with selective pressures and interspecific organelle introgressions can bias biogeographic inferences (Ballard & Whitlock 2004). Conversely, microsatellites have proved powerful for the elucidation of species history and recent hybridization (Lu et al. 2001; Randi & Lucchini 2002). Although DNA sequence information can be more useful for dating ancient demographic events, studies using microsatellites should help to unravel recent demographic processes, which have received little attention to date, such as gene flow between recolonization wave fronts.

Postglacial recolonization routes have been inferred from patterns of population structure (Taberlet et al. 1998; Lascoux et al. 2004; Schmitt 2007). However, until recently, the lack of suitable and tractable statistical methods for distinguishing between different demographic scenarios and estimating demographic and population parameters (effective population size, divergence time between populations and the amount of gene flow between populations) (Richards et al. 2007) hindered the accurate and precise inference of complex scenarios. Approximate Bayesian computation (ABC) provides a robust framework, allowing powerful inferences of a species’ past demography (Beaumont et al. 2002; Bertorelle et al. 2010; Csilléry et al. 2010). In particular, it enables probabilistic comparisons of alternative demographic scenarios (Bertorelle et al. 2010). A new framework has recently been developed, called ecological niche models (ENM), which may further improve inferences of the postglacial history of species by a complementary approach to genetic inference (Carstens & Richards 2007; Richards et al. 2007; Waltari et al. 2007; Brown & Knowles 2012).

The European crabapple (Malus sylvestris), a woody fruit species occurring across Western and Central Europe (Larsen et al. 2006), is a tree of the Rosaceae family, mainly pollinated by bees and flies, that grows in low-density populations in natural habitats. A great diversity of wild animals feeds on apple fruits but their efficiencies as seed dispersal vectors are unknown (Larsen et al. 2006). Malus sylvestris has been identified as a major contributor to the genome of the cultivated...
considered to be endangered (Jacques et al. 2009). Efforts to decipher the genetic structure and demographic history of this tree species are thus timely (i) to mitigate the effects of climate change and prevent negative effects of habitat fragmentation on this endangered species and (ii) to increase the genetic resources available for apple breeding programs, through the identification of genes conferring resistance to pathogens or tolerance of diverse abiotic stress conditions. The genetic structure of the European crabapple has been studied only in limited areas (Coart et al. 2003; Larsen et al. 2006) and appears to be weak at fine scale, suggesting a high dispersal capacity. No study has yet investigated the phylogeography of this emblematic wild European species across its entire distribution range. We therefore used population genetic analyses and ENM to investigate the glacial refugia and postglacial recolonization history of M. sylvestris. For genetic analyses, we used a comprehensive set of M. sylvestris individuals sampled throughout Europe and 26 microsatellite markers. We addressed the following questions: (i) Did wild apples survive the last glacial period in a single refuge area in Europe or at multiple areas? Did all relict populations contribute equally to the postglacial recolonization of Europe by wild apples? (ii) Can we detect the genetic consequences of successive founder events during postglacial colonization, that is, does genetic diversity decline with increasing distance from refugia? (iii) Can we detect genetic patterns of isolation-by-distance (IBD) and obtain information about dispersal capacities? (iv) Did admixture occur between recolonizing populations during expansion of the postglacial range of this species? (v) Can we reconstruct habitats that were suitable for the European crabapple in the past by ENM methods?

Materials and methods

Sampling

We have previously demonstrated the existence of gene flow from the cultivated apple M. domestica to the European wild apple (Cornille et al. 2012). We therefore initially carried out a STRUCTURE 2.3.3 (Pritchard et al. 2000) analysis, including 40 reference M. domestica cultivars previously identified as displaying no introgression from the European crabapple (i.e. with membership coefficients >0.9 to the M. domestica gene pool), and the entire M. sylvestris data set comprising 837 individuals. We retained only the 381 individuals assigned with a value of >0.9 to the M. sylvestris gene pool for further analyses. Leaf material was collected at 37 sites (including 25 sites with at least four individuals) across Europe (Table S1 and Fig. S1, Supporting information), covering most of the geographical distribution of the European crabapple except Spain and Sweden (see Euforgen map of the European crabapple http://www.euforgen.org/fileadmin/www.euforgen.org/Documents/Maps/PDF/Malus_sylvestris.pdf). As the European crabapple is a scattered species (i.e. 1 individual/ha), we defined site as a group of close individuals (<20 km); the geographical coordinates for each site were defined as the average of geographical coordinates over all individuals from that site.

DNA extraction and microsatellite genotyping

DNA was extracted with the NucleoSpin® plant DNA extraction kit II (Macherey & Nagel, Düren, Germany). PCR amplification was performed with the Multiplex PCR kit (Qiagen). We used 26 microsatellites, distributed over the 17 chromosomes, typed in 10 different multiplex reactions as previously described (Paticchi et al. 2009; Cornille et al. 2012). We retained only multilocus genotypes with fewer than 25% missing data. We checked the suitability of the markers for population genetic analysis with ARLEQUIN (Excoffier & Lischer 2010). None of the 26 microsatellite markers deviated significantly from the neutral equilibrium model, as shown by the nonsignificant P-values obtained in Ewens–Watterson tests, and no pair of markers was in significant linkage disequilibrium (Raymond & Rousset 1995; Rousset 2008). The markers used may therefore be considered to be unlinked and to be evolving in a quasi-neutral manner.

Descriptive statistics

We tested for the occurrence of null alleles with MICRO-CHECKER 2.2.3 (Van Oosterhout et al. 2004). Allelic richness and private allele richness were calculated with ADZE (Szpiech et al. 2008), for sites (i.e. geographical locations) and populations (i.e. clusters inferred by TESS analyses including hybrids up to 0.55 membership coefficient in the given cluster), using sample sizes of N = 12 (six individuals × two chromosomes) and N = 200 (100 individuals × two chromosomes), corresponding to the smallest number of observations for sites and populations, respectively. Heterozygosity, Weir and Cockerham F-statistics and Hardy–Weinberg equilibrium were assessed with GENEPOP 4.0 (Raymond & Rousset 1995; Rousset 2008). Only sampling sites with at least four successfully genotyped specimens were included in site-specific computations (25 sites in total comprising N = 344 individuals) except for ADZE analysis in which 22 sites were included (i.e. geographical locations with at least four individuals and within which at least two individuals were successfully genotyped for each marker). The population-specific
computations excluded five individuals that were not clearly assigned to any population, leaving a sampling size of \( N = 376 \) individuals.

**Population subdivision, genetic variability and isolation by distance**

We used the individual-based Bayesian clustering methods implemented in **STRUCTURE** 2.3.3 (Pritchard et al. 2000) and **TESS** 2.3.1 (Chen et al. 2007) to investigate population subdivision. These methods are based on the use of Markov chain Monte Carlo (MCMC) simulations to infer the assignment of genotypes into \( K \) distinct clusters. The underlying algorithms attempt to minimize deviations from Hardy–Weinberg and linkage disequilibrium within each cluster. The clustering procedure of TESS also includes a spatial component, such that genotypes from geographically closer locations are considered more likely to belong to the same cluster. In TESS analyses, we used the conditional autoregressive (CAR) Gaussian model of admixture with a linear trend surface, setting the spatial interaction parameter \( \rho \) at 0.6. These parameters \( \rho \) and trend affect the weighting assigned to spatial proximity when clustering genotypes.

For both methods, ten independent analyses were carried out for each number of clusters \( K (2 \leq K \leq 6 \) for **TESS** and \( 1 \leq K \leq 6 \) for **STRUCTURE**), with 500 000 MCMC iterations after a burn-in of 50 000 steps. Outputs were processed with **CLUMPP** 1.1.2 (Jakobsson & Rosenberg 2007), for the identification of potentially distinct modes (i.e. clustering solutions) in replicated runs for each \( K \). A \( G' \)-statistic >80% was used to assign replicates to a common TESS mode. We determined the amount of additional information explained by increasing \( K \) using the \( \Delta K \) statistic (Evanno et al. 2005) for **STRUCTURE** analyses and the rate of change of the deviation index criterion (DIC) when increasing \( K \) for **TESS** analyses.

We checked for IBD patterns as previously described (Loiselle et al. 1995). A Mantel test with 10 000 random permutations was performed between the individual coefficient of relatedness \( F_{st} \) and the matrix of the natural logarithm of geographical distance. These analyses were performed with **SPAGEDi** 1.3 (Hardy & Vekemans 2002). Spatial patterns of genetic variability were visualized by mapping variation in allelic richness for 22 sites in total (i.e. geographical locations with at least four individuals and within which at least two individuals were successfully genotyped for each marker) across space with the interpolation kriging function in **ArcInfo** (ESRI, Redlands, CA, USA), using a spherical semivariogram model.

**Approximate Bayesian inference**

We found weak genetic differentiation in the European crabapple between the north-eastern and south-eastern populations identified with **TESS**. We therefore investigated whether the observed pattern of genetic diversity resulted from (1) the simultaneous expansion of three independent populations (i.e. western, south-eastern and north-eastern; referred to as W, SE and NE) or (2) two simultaneous expansions of the W and SE populations, followed by the divergence and expansion of the NE population derived from the SE population (i.e. resulting from a more recent colonization wave front). We used **ABCtoolbox** (Wegmann et al. 2010) to compare these two scenarios, with and without the occurrence of gene flow (Fig. 1). The juvenile period of *M. domestica* lasts 5–10 years and no data for this parameter are available for *M. sylvestris*. We therefore assumed a generation time of 7.5 years. We estimated the effective size of each population \( (N_W, N_{NE}, N_{SE}) \), the rate of migration between populations for each generation \( (m_W, m_{NE}, m_{SE}) \), the exponential growth rate of each population \( (G_W, G_{NE}, G_{SE}) \) and the divergence times \( (T_{EXP}, T_{NE-SE}) \). The boundaries of the uniform (or log-uniform) prior distributions (Table S2, Supporting information) were chosen based on preliminary analyses run with very large priors (not shown). The ‘populations’ used in these analyses are the main clusters identified by Bayesian clustering methods.

For all models, identical microsatellite data sets were simulated for 14 of our loci (Ch01h01, Ch01h10, Ch02c06, Ch02d08, Ch05f06, Ch01f02, Hi02c07, Ch02c09, Ch03d07, Ch04c07, Ch02b03b, MS06g03, Ch04e03 and Ch02g01) that had been reported to carry perfect repeats (Gianfranceschi et al. 1998; Liebhard et al. 2002; Silfverberg-Dilworth et al. 2006), increasing confidence in the simulated mutation model. We checked that **TESS** yielded the same pattern of population structure with these 14 markers as observed with the 26 markers. Using **CLUMPP** (Jakobsson & Rosenberg 2007), we showed that both sets of markers [14 vs. 26 single sequence repeat (SSR)] gave similar clustering patterns (data not shown, similarity index \( G' = 96\% \) for \( K = 3 \)). We generated \( 2 \times 10^9 \) genetic data sets from coalescent simulations, using model parameters drawn from prior distributions (Table S2, Supporting information) under the four previously specified scenarios. For each simulation, we calculated two summary statistics per population: \( H \), the mean heterozygosity across loci and \( K \), the mean number of alleles across loci. We also calculated pairwise \( F_{ST} \) (Weir & Cockerham 1984) and genetic distances \( (\delta u)^2 \) (Goldstein et al. 1995) between pairs of populations. We
conducted a preliminary principal component analysis (PCA) with R software (MASS package, function prcomp), based on 3000 simulated data sets for *M. sylvestris*, for the four models, to reveal visual correlations between the main model parameters and the summary statistics (Tellier et al. 2011).

We assumed a generalized stepwise model of microsatellite evolution (Estoup et al. 2002). The mutation rate was allowed to vary across loci, with locus-specific mutation rates being drawn from a gamma distribution ($\lambda$, $\lambda/a$) in which $\lambda$ is the mutation rate per generation and $a$ is a shape parameter. We assumed a log-uniform prior distribution for $\lambda$ [0.00001, 0.02] and a uniform distribution for $a$ [1, 30].

We compared the four models by calculating their Bayes factors (Wegmann et al. 2010) and estimating their relative posterior probabilities, based on the 1% of simulated data sets most closely matching the observed data (i.e. 2000 simulated data sets). Once the best model had been chosen, we estimated demographic parameters under this scenario, using a general linear model (ABC-GLM) post-sampling regression adjustment for the 2000 retained simulations (Leuenberger & Wegmann 2010; Wegmann et al. 2010). We report the mode and 95% highest posterior density (HPD) interval for each model parameter estimate.

The performance of the method for discriminating between competing historical models was assessed by analysing test data sets (called pseudo-observed data sets) simulated with the same number of loci and individuals as for the observed data sets. We simulated 2000 such data sets for each competing model, using parameter values drawn from the same prior distributions as for the original analyses. We determined the relative posterior probabilities of competing models for each pseudo-observed data set, using the model choice procedure, as described above (Wegmann et al. 2010). Confidence in model choice was then estimated from the percentage of times that a given scenario did not have the highest posterior probability of the competing scenarios when it was actually the true scenario (type I error) and the percentage of times that a given scenario had the highest posterior probability when it was not the true scenario (type II error).

**Fig. 1** Demographic models compared in approximate Bayesian computations. Model a assumes that three populations [western (W), south-eastern (SE), north-eastern (NE)] diverged and underwent simultaneous exponential growth; bidirectional gene flow was assumed between all pairs of populations. Model b is identical to model a but without gene flow. Model c assumes that the W and SE populations expanded simultaneously and that the NE population subsequently diverged from the SE population, with all populations undergoing exponential growth; bidirectional gene flow was assumed between all population pairs. Model d is identical to model c but without gene flow. $N_x$: effective population size of population x, $m_{x-y}$: gene flow per generation from population x to population y, $T_{\text{EXP}}$: divergence time at the beginning of the Holocene, $T_{\text{NE-SE}}$: time at which the NE population split from the SE population.

The performance of the method for discriminating between competing historical models was assessed by analysing test data sets (called pseudo-observed data sets) simulated with the same number of loci and individuals as for the observed data sets. We simulated 2000 such data sets for each competing model, using parameter values drawn from the same prior distributions as for the original analyses. We determined the relative posterior probabilities of competing models for each pseudo-observed data set, using the model choice procedure, as described above (Wegmann et al. 2010). Confidence in model choice was then estimated from the percentage of times that a given scenario did not have the highest posterior probability of the competing scenarios when it was actually the true scenario (type I error) and the percentage of times that a given scenario had the highest posterior probability when it was not the true scenario (type II error).

Projections of *Malus sylvestris* distribution during LGM

We modelled the climatic niche of *M. sylvestris* from the current species distribution with the *biomod* package (v 1.1-7.00, 2011-08-03; Thuiller et al. 2009) downloaded from *Worldclim* data set v1.4 (http://www.worldclim.org/;
Hijmans et al. 2005) and records of the presence of *M. sylvestris* obtained for sampled individuals (*N* = 381). We estimated the pairwise correlation of values for the 19 bioclimatic variables by calculating Pearson’s correlation coefficients and retained only the variables that were not strongly correlated (i.e. with Pearson’s correlation coefficients < 0.75). We tested both sets of bioclimatic variables because we had no primer assumptions about species’ preferences, and we wanted to compare the respective LGM projections when taking into account autocorrelation and the overfitting of the data. We removed duplicated coordinate data points, resulting in 73 presences in total (Dataset S1, Supporting information) to evaluate the distribution at the LGM. These models make the assumptions that climate is one of the main factors driving species distribution and that the climatic niche of this species has remained largely unchanged in recent centuries (Text S1, Supporting information).

**Results**

**Population structure**

Summary statistics for genetic variability are shown in Tables S3 and S4 (Supporting information). For the 25 sites with at least four samples, the mean number of genotypes was 13.8 ± 8.4 (average ± standard deviation), allelic richness was 3.7 ± 0.4 (range: 2.6–4.4) and genetic diversity was 0.84 ± 0.14 (range: 0.55–0.89), on average, across markers. Heterozygote deficit, estimated over the whole data set, was highly significant (*P* < 0.001), but low (*F*<sub>ST</sub> = 0.03, with a mean of 0.03 ± 0.07 per site and per marker). The mean *F*<sub>STs</sub> across loci were small (*F*<sub>ST</sub> = 0.10, range: 0.008–0.280) but significant, for all pairs of sites (*P* < 0.02, Table S5, Supporting information).

The results of TESS analyses are shown in Figs 2 and 3. For *K* = 2, the analyses revealed a clear west/east partitioning. The simulations for *K* = 3 split the eastern cluster into NE and SE clusters. A similar pattern was obtained for *K* = 4, whereas for *K* = 5, a fourth cluster was identified in Bosnia-Herzegovina. STRUCTURE analyses generated congruent clustering patterns (Figs S2 and S3, Supporting information).

In TESS analyses, DIC values decreased monotonically from *K* = 2 to *K* = 6. Thus, increasing the number of clusters continually improved the fit of the model to the data. However, DIC seemed to decrease more slowly after *K* = 3 (Fig. S4, Supporting information), suggesting that further increases in *K* provided little information. In STRUCTURE analyses, the mode of the Δ*K* statistic was observed at *K* = 2 (Δ*K* = 1670, *Pr* | ln *L* = 2683, Fig. S5, Supporting information), but Δ*K* was still high at *K* = 3 (Δ*K* = 480, *Pr* | ln *L* = 838), suggesting further improvement in the fit of the model. Based on the narrow geographical distribution of the clusters inferred at *K* > 3, and the Δ*K* and DIC values obtained, *K* = 3 was considered the most biologically relevant clustering solution for subsequent historical inference.

We used the TESS membership coefficient inferred at *K* = 3 to define the three populations used in subsequent analyses. Genotypes were assigned to a given population if their membership coefficient for that population exceeded 0.55. Five genotypes could not be assigned to any population and were not included in subsequent analyses; *N* = 376 individuals were thus retained for population-specific computations. The three populations are hereafter referred to as the ‘western’ (*W*, red, *N* = 213), the ‘north-eastern’ (*NE*, green, *N* = 90) and ‘south-eastern’ (*SE*, blue, *N* = 73) populations. The *W* population was relatively homogeneous, with 91% of genotypes having membership coefficients > 0.9 for that population (Figs 2 and 3). The *NE* and *SE* populations presented higher levels of admixture, with 31% (*N* = 28) and 26% (*N* = 19) of genotypes, respectively, having membership coefficients < 0.9.
Admixed genotypes in the NE population were assigned to both the W (N = 7) and SE (N = 14) populations, whereas admixture in the W and SE populations occurred only with the NE population (Figs 2 and 3).

Population-specific polymorphism summary statistics are shown in Table 1. Genetic differentiation was weakest between the NE and SE populations (FST = 0.04, P < 0.001), and higher between the W and NE (FST = 0.09, P < 0.001) or W and SE (FST = 0.12, P < 0.001) populations. The W population had a significantly lower genetic variability than the NE and SE populations, in terms of both genetic diversity (Hd) [Wilcoxon signed rank (WSR) tests: V = 271, P = 0.007 and V = 229, P = 0.008 respectively] and allelic richness (Ar) (WSR: V = 351, P < 0.0001 and V = 345, P < 0.0001, respectively). The W population contained a similar number of private alleles to the NE population (WSR: V = 262, P = 0.08), but significantly lower than that of the SE population (WSR: V = 341, P < 0.0001). The NE and SE populations displayed similar levels of Hd (WSR: V = 338, P = 0.7), but the SE population had a higher allelic richness (WSR: V = 320, P < 0.0001) and a larger number of private alleles (WSR: V = 338, P < 0.0001) (Table 1).

The map of interpolated allelic richness (Fig. 4) showed that genetic diversity decreased with increasing latitude. Within each of the three populations, genetic differentiation and geographical distance were significantly correlated (Table 1), consistent with an IBD model. The spatial Sp statistic can be used to quantify spatial structure; lower Sp values are associated with greater dispersal capacities and/or effective population sizes. Sp values were very low both overall and within each population (Table 1), lying close to 0 (Sp = 0.04, P < 0.0001). This suggests that each population had a high dispersal capacity and/or a large effective population size.

Models of population expansion and estimation of demographic parameters

We used ABC to determine statistically which of the following scenarios provided the most likely explanation for the existence of the three genetically differentiated populations of Malus sylvestris in Europe: (i) the simultaneous expansion of three independent populations (W, SE and NE) that had differentiated during the last period of glaciations, with (model a) or without...
(model b) gene flow, (ii) divergence and expansion of the W and SE populations, followed by the divergence and expansion of the NE population derived from the SE population (i.e. resulting from a more recent colonization wave front), with (model c) or without (model d) gene flow (Fig. 1). The correlations between model parameters and summary statistics are presented in Fig. S6 (Supporting information). The relative posterior probabilities calculated for each model provided the strongest statistical support for model a, suggesting that the three populations (W, SE and NE) diverged simultaneously, with all populations growing exponentially and bidirectional gene flow occurring between each pair of populations during expansion (Table 2; Bayes factor for model a = 4.37). The models with the NE population diverging from the SE population (models c and d) had the lowest relative posterior probabilities (Table 2). The migration rates per generation were estimated at $m_{W-SE} = 0.01$ [95% HPD: 0–0.21], $m_{W-NE} = 0.03$ [0–0.21], $m_{SE-NE} = 0.01$ [0–0.22], $m_{SE-W} = 1.01 \times 10^{-15}$ [0–0.17], $m_{NE-SE} = 1.01 \times 10^{-15}$ [0–0.21] and $m_{NE-W} = 0.05$ [0–0.27].

We obtained estimates of effective population sizes of $N_W = 40,883$ [524–828,327] for W, $N_{NE} = 20,691$ [505–565,916] for NE and $N_{SE} = 40,882$ [524–828,327] for SE. Using a generation time of 7.5 years, we estimated the population split to have occurred 303,016 years ago [120,963–545,649] (Fig. S7, Supporting information).

We also checked that the power of the analysis was sufficiently high to discriminate between the competing models: for model a against all three other models, the type I error rate and the mean type II error rate were 0. Overall, ABC analyses provided clear strong support

### Table 1 Genetic polymorphism and spatial pattern of differentiation within each population in *Malus sylvestris*. Five genotypes could not be assigned to any population, and thus, analyses were conducted on a total of $N = 376$ individuals

<table>
<thead>
<tr>
<th>Microsatellite polymorphism</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W (red)</td>
</tr>
<tr>
<td>$N$</td>
<td>213</td>
</tr>
<tr>
<td>$H_O$</td>
<td>0.69</td>
</tr>
<tr>
<td>$H_E$</td>
<td>0.74</td>
</tr>
<tr>
<td>$F_{TS}$</td>
<td>0.07***</td>
</tr>
<tr>
<td>$A_r$</td>
<td>11.5</td>
</tr>
<tr>
<td>$A_p$</td>
<td>1.3</td>
</tr>
</tbody>
</table>

### Spatial pattern (logarithm)

<table>
<thead>
<tr>
<th></th>
<th>Sp</th>
<th>Mean[Ln(dist)]</th>
<th>$b$</th>
<th>$F_{ij}$</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.005</td>
<td>1.3</td>
<td>$-0.006^{***}$</td>
<td>0.03</td>
<td>0.008</td>
</tr>
</tbody>
</table>

$N$: sample size of each population; $H_O$ and $H_E$: observed and expected heterozygosities; $F_{TS}$: inbreeding coefficient; $A_r$: mean allelic richness for loci, corrected by the rarefaction method, estimated for a sample size of 100; $A_p$: number of private alleles, corrected by the rarefaction method, estimated for a sample size of 100; Sp: Sp parameter; mean[Ln(dist)]: mean of the logarithm of the geographical distance between genotypes; $b$: regression slope between $F_{ij}$ and the logarithm of geographical distance; $F_{ij}$: mean $F_{ij}$ between individuals from the first distance class; $r^2$: squared correlation coefficient between the logarithm of geographical distance and $F_{ij}$.

***$P < 0.001$. (model b) gene flow, (ii) divergence and expansion of the W and SE populations, followed by the divergence and expansion of the NE population derived from the SE population (i.e. resulting from a more recent colonization wave front), with (model c) or without (model d) gene flow (Fig. 1). The correlations between model parameters and summary statistics are presented in Fig. S6 (Supporting information). The relative posterior probabilities calculated for each model provided the strongest statistical support for model a, suggesting that the three populations (W, SE and NE) diverged simultaneously, with all populations growing exponentially and bidirectional gene flow occurring between each pair of populations during expansion (Table 2; Bayes factor for model a = 4.37). The models with the NE population diverging from the SE population (models c and d) had the lowest relative posterior probabilities (Table 2). The migration rates per generation were estimated at $m_{W-SE} = 0.01$ [95% HPD: 0–0.21], $m_{W-NE} = 0.03$ [0–0.21], $m_{SE-NE} = 0.01$ [0–0.22], $m_{SE-W} = 1.01 \times 10^{-15}$ [0–0.17], $m_{NE-SE} = 1.01 \times 10^{-15}$ [0–0.21] and $m_{NE-W} = 0.05$ [0–0.27].

We obtained estimates of effective population sizes of $N_W = 40,883$ [524–828,327] for W, $N_{NE} = 20,691$ [505–565,916] for NE and $N_{SE} = 40,882$ [524–828,327] for SE. Using a generation time of 7.5 years, we estimated the population split to have occurred 303,016 years ago [120,963–545,649] (Fig. S7, Supporting information).

We also checked that the power of the analysis was sufficiently high to discriminate between the competing models: for model a against all three other models, the type I error rate and the mean type II error rate were 0. Overall, ABC analyses provided clear strong support
for gene flow between recolonizing refugia and for the simultaneous divergence of the three populations.

**Ecological niche modelling**

Model performance, as assessed from the AUC (Area Under the receiver operating characteristic Curve), was high for all six algorithms (Table S6, Supporting information; AUC = 0.98 ± 0.01), indicating that all six algorithms fit the data well (Monserud & Leemans 1992; Fieldings & Bell 1997; Allouche et al. 2006). With thresholds maximizing the True Skill Statistic (TSS), *M. sylvestris* had a good TSS value of 0.79 ± 0.04. We ran ENM with both sets of past climate data, CCSM2 and MIROC, but only MIROC gave consistent results across Europe. We therefore present projections based on MIROC data (Figs 5b and S8b, Supporting information). The projection onto current climate layers identified a putative suitable climate area essentially located in Western Europe for *M. sylvestris* (Figs 5a and S8a, Supporting information). The MIROC model predicted that the areas suitable for this species during the LGM were limited to lower latitudes than those considered suitable today and were more fragmented, with in particular low probability of contact between Eastern and Western Europe, in agreement with the genetic data (Figs 5 and S8, Supporting information). The climatic model suggested that populations of the European crabapple may have been maintained in areas further north than the typical glacial refugia (Hewitt 2004), with possible continuity between the populations from Western Europe, Italy and the Balkans. The predicted distribution, however, does not show a refugial distribution in isolated places as expected if the species survived in nunataks in northern latitudes.

**Discussion**

Paleodistribution modelling and genetic data allowed inferences on the phylogeography of the European crabapple, an endangered species and valued genetic resource for apple breeding. The distribution predicted based on climatic data was overall consistent with population genetic analyses, altogether suggesting population and range contractions of *Malus sylvestris* during the last glaciation with a fragmentation between Eastern and Western Europe followed by postglacial recolonization of Europe.

**Glacial refugia for Malus sylvestris**

*Malus sylvestris* displayed a clear geographical pattern of population structure, with three differentiated populations in Europe: (i) a western population (W) spanning a huge area from France to Norway, and an eastern group subdivided into (ii) a north-eastern (NE) population around the Carpathian Mountains and (iii) a south-eastern (SE) population located at the north-east edge of the Balkan Peninsula. The strong differentiation between the three populations and the decreasing allelic richness at increasing latitudes suggest that the European crabapple contracted its range to southern glacial refugia, one of which probably was in the Iberian Peninsula or in the south of France, and another one in the Balkans.

The NE population growing around the Carpathian Mountains displayed a low level of genetic differentiation from the SE population, but had a high level of allelic richness. There were two possible origins for this population: a refugium from north-eastern Europe that came into secondary contact with the SE population and a wave front from the Balkan refugium during the recolonization of Europe. ABC analyses evaluating the fit of various demographic models to microsatellite data showed that the most strongly supported scenario was the simultaneous divergence of the three populations, with gene flow between all population pairs. The NE population, therefore, probably originated from a glacial refugium rather than during the postglacial recolonization of Europe.

The geographical pattern of population structure uncovered in *M. sylvestris* is consistent with those found in other animal, plant and fungal taxa (Hewitt 2004; Lascoux et al. 2004; Schmitt 2007; Vercken et al. 2010) and, in particular, with the patterns commonly found in temperate forest trees (Petit et al. 2002; Heuertz et al. 2004, 2006; Magri et al. 2006). Indeed, the existence of a large western European population and differentiated eastern populations in the Balkans Peninsula has been reported in *Quercus* sp., *Abies alba* and *Fraxinus excelsior*. In some other tree species, such as *Alnus glutinosa* (King & Feris 1998) and the common beech *Fagus sylvatica* (Magri et al. 2006), the existence of additional refugia in the eastern part of Europe, particularly in the Carpathian Mountains, has been suggested on the basis of the high level of genetic diversity in cpDNA and pollen fossil records (Magri et al. 2006). Our microsatellite

<table>
<thead>
<tr>
<th>Model</th>
<th>$p$</th>
<th>BF</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.81</td>
<td>4.37</td>
</tr>
<tr>
<td>b</td>
<td>0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>c</td>
<td>2.3e-12</td>
<td>2.3e-12</td>
</tr>
<tr>
<td>d</td>
<td>1.0297e-56</td>
<td>1.03e-56</td>
</tr>
</tbody>
</table>

Table 2 Relative posterior probabilities ($p$) and Bayes factor (BF) for the four historical models compared by approximate Bayesian computations. The models are described in Fig. 1.
markers clearly identified a distinct genetic cluster around the Carpathian Mountains in the European crabapple. A similar situation has been suggested for the common ash *Fraxinus excelsior* (Heuertz et al. 2004), but no clear origin of the Carpathian population could be established due to a lack of samples in this area. Our ABC analyses indicated that the NE population originated at the same time as the other recolonizing populations, thus suggesting that the Carpathian Mountains may have acted as a glacial refugium for temperate

Fig. 5 Ensemble forecasting of the six different algorithms predicting the current (a) and last glacial maximum (LGM) (b) suitable climate area distribution for *Malus sylvestris*. The probabilities (×1000) of a suitable habitat are given in the legend.
forest tree species, as for other temperate species, including mammals, reptiles and amphibians (Provan & Bennett 2008; Stewart et al. 2009). The existence of such a ‘northern’ glacial refugium is consistent with the results of ENM, suggesting that *M. sylvestris* may have survived at high latitudes. In some tree species, such as *Fagus sylvatica*, it has been suggested that such a northern refugium may have served as the main source population for the recolonization of Eastern Europe after the LGM, whereas a south-eastern population spread over a much more limited area during postglacial recolonization (Hu et al. 2008).

For ash, silver fir, oaks and beech, Italy has been identified as a possible additional glacial refugium, isolated from other lineages by the Alpine barrier (Taberlet et al. 1998). We detected no footprint of an Italian refugium in *M. sylvestris*. This may be due to the lack of pure *M. sylvestris* from this region. Our samples from Italy indeed were all introgressed by the cultivated apple *M. domestica*, which may be linked to the introduction of the cultivated apple in Europe by the Romans in Italy, probably 3000 years ago (Mabberley et al. 2001). For the W population, which probably expanded from an Iberian refugium, the level of allelic richness was high, but significantly lower than that in the eastern populations. These differences in genetic variability may be due to much more severe climatic episodes (i.e. arid and cold) during the Quaternary Period in this region than in other parts of Europe (Petit et al. 2003). Tree populations that survived successive ice ages in the Iberian Peninsula were restricted to a few small suitable areas and were thus smaller than those in other parts of Europe.

**Suture zones and recolonization fronts for the European crabapple**

There were two main waves of recolonization by *M. sylvestris* from the glacial refugia in Europe: Western Europe, right into the north, was probably recolonized by populations from the Iberian Peninsula or the south of France, whereas the population from the Carpathian Mountains spread, albeit to a lesser extent, northwards in Eastern Europe. The population from the Balkan refugium does not seem to have recolonized large areas.

Our microsatellite markers indicate some admixture in the three populations. The W population contained the smallest number of admixed individuals and displayed the highest level of genetic differentiation from the other populations. This pattern suggests higher levels of recent genetic exchange between the SE and NE populations. The lack of samples from Central Europe may have resulted in an underestimation of the number of individuals with admixed ancestry in the western and eastern populations, but ABC analyses provided a high level of confidence for the choice of the model assuming gene flow between populations.

The biogeographic scenario uncovered in our study, with two main recolonization fronts in Western and Eastern Europe, has been demonstrated for many other temperate tree species (Heuertz et al. 2004; Lascoux et al. 2004). Suture zones in Central Europe, as detected for the European crabapple, are also typical of other temperate tree species (Petit et al. 2002; Heuertz et al. 2004, 2006; Magri et al. 2006; Liepelt et al. 2009). However, a clear suture between SE and NE populations and evidence for admixture have not been reported before. These findings demonstrate the utility of nuclear microsatellite markers for retracing the ancient demographic history of populations and the extent of admixture in phylogeographic studies.

**Historical gene flow and dispersal capacity in the European crabapple**

The European crabapple is dispersed by animals, but the weak spatial genetic structure within each population at the European scale, weak IBD patterns and low Sp values suggest that this species may be dispersed over large distances. These results are consistent with those previously reported for scattered temperate tree species (Oddou-Muratorio et al. 2004; Oddou-Muratorio & Klein 2008). The high dispersal capacity of the European crabapple and its current wide distributional range extending into the northern-most parts of Europe characterize this species as a rapid colonizer (Svenning & Skov 2007), consistent with its high pioneering capacity (Larsen et al. 2006).

**Ecological niche models**

Ecological niche models projections onto current climate layers resulted in a current suitable habitat for the European crabapple covering most of Europe, except southern Spain and Italy, consistent with the Euforgen map of actual distribution (http://www.euforgen.org/fileadmin/www.euforgen.org/Documents/Maps/PDF/Malus_sylvestris.pdf), although this extends further east than the region studied in this study. Predictions showed that the distribution of the European crabapple, like those of many other European tree species (Svenning et al. 2008), was affected by climate changes during the Quaternary Period, with a contraction and fragmentation of populations. The modelling results suggested possible refugia in the European crabapple in the northeast, southeast and west, concordant with the distribution of the three genetic clusters and ABC results.
Despite the overall concordance between ENM and genetic data, the predicted past distribution extends northwards up within the extent of the glaciers. This raises questions about the accuracy of ENM for paleoreconstructions and their interpretation, including limitations concerning the available predictor variables (Araújo & Guisan 2006). Indeed, in these models, the species niche is defined in terms of temperature and precipitation data only. Although climate is one of the main factors underlying species distribution, other factors will need to be taken into account in the future (Pearson & Dawson 2003), including dispersal processes (using hybrid approaches), biotic interactions (e.g. competitors and pollinators) and stages of succession in forests. One of the main issues of niche modelling is the assumption of niche conservatism. If the species niche shifts over time, the species may not respond to climate change in a predictable way (Jezkova et al. 2011). In addition, calibrating the climatic niche of species under current conditions and projecting them to non-analogous conditions in the past would lead to spurious response curves and therefore to naive projection (Thuiller et al. 2004).

**From population structure to conservation strategies**

Our inferences suggest high rates of historical gene flow between European crabapple populations, but current migration rates may be different, especially due to forest fragmentation. As a wild relative and contributor to the cultivated apple (Cornille et al. 2012), the European crabapple is a target for conservation and sustainable management programs for genetic resources. Knowledge of the existence and location of differentiated populations and of admixture zones is essential to guide conservation programs (Pautasso 2009). We report here the south of France, the Balkans and the Carpathian Mountains as hotspots of genetic diversity. In addition the admixture zones correspond to ‘melting pots’ of genetic diversity (Petit et al. 2003; Liepelt et al. 2009; Jay et al. 2012) that may include recombinant genotypes well suited to new environmental conditions and capable of facing up to global warming. The genetic uniqueness of southern ‘rear-edge’ populations, such as the populations from the South of France, is thus also of key importance for long-term conservation purposes (Petit & Hampe 2006).

**Acknowledgements**

We thank the Région Ile de France (PICRI), IDEEV and SBF (Société Botanique de France) for funding, Plateforme de Génotypage GENTYANE INRA UMR 1095 and Pauline Lasserre for assistance with genotyping. We thank all those who helped with sampling (Table S1, Supporting information). We thank Xavier Vekemans, Bruno Fady and three anonymous referees for very useful suggestions.

**References**


Hardy OJ, Vekemans X (2002) SPAgeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes, 2, 618–620.
© 2013 Blackwell Publishing Ltd


Data accessibility

Microsatellites data are published on the DRYAD repository with sampling locations, the ecological niche models data and the TESS input used for analyses and ABCtoolbox inputs, doi: 10.5061/dryad.sn1 m7.

Supporting information

Additional supporting information may be found in the online version of this article.

Dataset S1 X/Y coordinates of presences of Malus sylvestris in Europe used for ENM.

Text S1 ENM methodology used to project paleodistribution of Malus sylvestris.

Fig. S1 Sampling of the different Malus sylvestris sites through Europe.

Fig. S2 Bayesian clustering results of Malus sylvestris in Europe (N = 381) using the program STRUCTURE from K = 2 to K = 6.

Fig. S3 Maps of mean membership probabilities per site from the STRUCTURE analysis for Malus sylvestris assuming 2 to 5 clusters.

Fig. S4 Estimated number of populations in Malus sylvestris from TESS analyses using the DIC.

Fig. S5 Estimated number of populations in Malus sylvestris from STRUCTURE analyses using the ΔK.

Fig. S6 PCA on 3,000 simulations for Malus sylvestris.

Fig. S7 Marginal posterior distributions of demographic and historical parameters estimated by Simcoal2.

Fig. S8 Ensemble forecast using Malus sylvestris presence records and pseudo-absences projected onto the map of Europe and Western Russia using 19 bioclimatic variables.

Table S1 Description of the Malus sylvestris accessions analysed with their geographical origins and providers, and acknowledgement.

Table S2 Prior distributions used in approximate Bayesian computations.

Table S3 Summary statistics for each Malus sylvestris sampling site with at least four individuals.

Table S4 Summary statistics for the 26 microsatellite loci in Malus sylvestris.

Table S5 Pairwise genetic differentiation (FST) among the 25 sites.

Table S6 AUC Index for ENM ran with eight and 19 bioclimatic variables for each of the six models and each repetition.