Nuclear and chloroplast DNA phylogeography reveals vicariance among European populations of the model species for the study of metal tolerance, *Arabidopsis halleri* (Brassicaceae)

Maxime Pauwels, Xavier Vekemans, Cécile Godé, Hélène Frérot, Vincent Castric and Pierre Saumitou-Laprade

Laboratoire de Génétique et Evolution des Populations Végétales, FRE CNRS 3268, Université de Lille-Lille1, F-59655 Villeneuve d'Ascq, France

Author for correspondence:
Maxime Pauwels
Tel: +33 3 20 33 62 38
Email: maxime.pauwels@univ-lille1.fr

Summary

- *Arabidopsis halleri* is a pseudometallophyte involved in numerous molecular studies of the adaptation to anthropogenic metal stress. In order to test the representativeness of genetic accessions commonly used in these studies, we investigated the *A. halleri* population genetic structure in Europe.
- Microsatellite and nucleotide polymorphisms from the nuclear and chloroplast genomes, respectively, were used to genotype 65 populations scattered over Europe.
- The large-scale population structure was characterized by a significant phylogeographic signal between two major genetic units. The localization of the phylogeographic break was assumed to result from vicariance between large populations isolated in southern and central Europe, on either side of ice sheets covering the Alps during the Quaternary ice ages. Genetic isolation was shown to be maintained in western Europe by the high summits of the Alps, whereas admixture was detected in the Carpathians.
- Considering the phylogeographic literature, our results suggest a distinct phylogeographic pattern for European species occurring in both mountain and lowland habitats. Considering the evolution of metal adaptation in *A. halleri*, it appears that recent adaptations to anthropogenic metal stress that have occurred within either phylogeographic unit should be regarded as independent events that potentially have involved the evolution of a variety of genetic mechanisms.

Introduction

Biological variation below the species level and the distribution of this variation in geographically delimited subspecific units are long-debated phenomena (Mallet, 2001; Winker, 2010). Subspecific units may involve either ecogeographic or phylogeographic units (Avise, 2000). Ecogeographic units are usually determined using the spatial distribution of nonmolecular, organismal traits, and are called either '(morphological) subspecies', when differential traits involve morphology (Wilson & Brown, 1953), or 'ecotypes', when phenotypic differentiation is assumed to result from adaptation to different environments (Turesson, 1922). By contrast, phylogeographic units are identified using molecular genotyping. They are geographic sets of conspecific populations in which all alleles at a specific locus are genealogically closer to each other than to any allele from other sets (Avise, 2000; Morrone, 2009). They are assumed to result from vicariance among populations that have been genetically isolated over long periods of time (Avise, 2000; Morrone, 2009).

Ecogeographic and phylogeographic concepts both suggest the existence of geographically structured units within species, among which barriers have sufficiently reduced genetic exchange to generate 'breaks' in population structure and favour genetic differentiation among isolated units. Barriers may be natural obstacles, such as mountains or rivers, or may be genetically determined when selection acts against the hybridization of divergent units, or a combination of both (Bierne *et al.*, 2011). In some circumstances, ecogeographic and phylogeographic breaks may have established simultaneously and may therefore be confounded. For example, ecologically based adaptive differentiation among units, generating exogenous barriers to gene flow, may influence the localization of tension zones and genetic breaks among evolutionary units (Bierne *et al.*, 2011). Thus, ecogeographic differentiation among populations in contrasted edaphic environments is sometimes assumed to favour reproductive isolation among divergent evolutionary units (Kruckeberg, 1986; Macnair & Gardner, 1998; Rajakaruna, 2004). However, although the setup of phylogeographic breaks must require long-lasting reduction in gene flow and is necessarily of ancient origin (Avise, 1989), breaks in ecogeographic structure may establish rapidly, that is, in a small number of generations, in particular when selection is involved in the reduction in gene flow (e.g. Macnair & Gardner,

1998), and may therefore be more recent (Zink & Barrowclough, 2008). In such cases, phylogeographic and ecogeographic breaks may be noncongruent. This suggests that phylogeography could give a 'temporal context (to) traditional ecogeographic perspectives' (Avise, 2000). Indeed, the comparison of spatial localizations of phylogeographic and ecogeographic breaks within a single species may be helpful to infer the spatial and temporal scales at which barriers to gene flow and population differentiation have established (Thorpe *et al.*, 1995; Zink & Barrowclough, 2008).

Arabidopsis halleri is a clonal, self-incompatible and highly outcrossing perennial weed that is closely related to the model species *A. thaliana*. Because it is able to both colonize zinc- and cadmium-polluted sites (metal tolerance) and concentrate large amounts of these metals in foliar tissues (metal hyperaccumulation), the species is a promising model to gain an insight into the genome-wide processes involved in ecologically important traits (Roosens *et al.*, 2008a,b; Verbruggen *et al.*, 2009; Bomblies & Weigel, 2010).

Across its distribution range, from lowland to subalpine zones, from Europe to the Far East (Al-Shehbaz & O'Kane, 2002), *A. halleri* can be divided into ecogeographic units in several ways. First, the species occurs on both metal-polluted and nonpolluted soils; in Europe, ecological and phenotypic differentiation for metal tolerance between metallicolous (M) and nonmetallicolous (NM) populations suggests the existence of two established edaphic ecotypes (Brej & Fabiszewski, 2006; Pauwels *et al.*, 2006; Meyer *et al.*, 2010). Second, the occurrence of morphological subspecies with distinct geographic distributions has been described (Al-Shehbaz & O'Kane, 2002; Kolník & Marhold, 2006; Koch *et al.*, 2008). It is well recognized, for example, that Asian populations belong to a distinct subspecies, *A. h.* ssp. *gemmifera* (Al-Shehbaz & O'Kane, 2002), probably isolated from European accessions from the Pleistocene (Koch *et al.*, 2008). Indeed, chorological data from the literature (Hoffmann, 2005; Koch & Matschinger, 2007) and online databases (AFE, http://www.luomus.fi/english/botany/afe/; GBIF Data Portal, http://data.gbif.org) suggest that European and Asian populations are disjunct. By contrast, in Europe, although no major gaps in the distribution are mentioned, the species is divided into several morphological subspecies (although the actual number of subspecies remains debated, and varies from two (*A. h.* ssp. *halleri* and *A. h.* ssp. *ovirensis*; Al-Shehbaz & O'Kane, 2002) to four (*A. h.* ssp. *dacica*, *A. h.* ssp. *halleri*, *A. h.* ssp. *ovirensis*, *A. h.* ssp. *tatrica*; Kolník & Marhold, 2006; Koch *et al.*, 2008)).

Interestingly, although the occurrence of several ecogeographic units suggests the existence of differentiated genetic pools, and although the belonging of any genetic accession to a particular pool could affect the final results, this is barely taken into account in studies of metal tolerance using *A. halleri* as a model. In this context, we analysed the population genetics of *A. halleri* in Europe, in which both edaphic ecotypes and up to four subspecies are mentioned, using a phylogeographic approach. In comparison with Pauwels *et al.* (2005), we surveyed the population structure covering the entire species range in Europe. Moreover, in contrast with Pauwels *et al.* (2008a,b), we surveyed

both the chloroplast and nuclear genome (cpDNA and nDNA, respectively) and combined data on allele frequencies from both genealogically ordered chloroplast haplotypes (chlorotypes) and unordered alleles at multiple independent nuclear microsatellite loci. Compared with chlorotype data, microsatellite data are typically more powerful to document ongoing demographic processes, such as contemporary gene exchange (e.g. gene frequency data are particularly efficient in revealing the hybrid nature of some individuals; Vallender *et al.*, 2007; Montarry *et al.*, 2010). To help interpret the phylogeographic patterns (Waltari *et al.*, 2007), we performed ecological niche modelling and compared the potential current species distributional areas with those occupied by the species during the Last Glacial Maximum (LGM, 15 000 ^{14}C years ago). We then compared the phylogeographic pattern in *A. halleri* with current knowledge about the phylogeographic history of European plant species, and with the spatial distribution of know ecogeographic units, in order to place their differentiation into an historical perspective.

Materials and Methods

Sampling

We sampled 65 locations (Supporting Information Table S1), 28 of which have already been described in a previous genetic study (Pauwels *et al.*, 2005). Sampling included M and NM locations scattered across the European species range. Most samples belonged to the native species range of *Arabidopsis halleri* (L.) O'Kane & Al-Shehbaz (syn. *Cardaminopsis halleri* (L.) Hayek), as defined in the *Atlas Florae Europaeae*.

Genetic analysis

DNA extraction In each sample, leaves were collected from 6 to 70 individuals, separated by at least 3 m to avoid the sampling of clones (van Rossum *et al.*, 2004). Sample sizes more or less reflected population sizes, with nearly exhaustive sampling in populations with the smallest sample sizes. Leaves were immediately dried in silica gel before molecular analysis. DNA from each genotype was extracted from 15 to 20 mg of dry material using the Qiagen® Dneasy® kit, and PCR amplification was performed on 1 : 100 dilutions.

Genotyping We genotyped 1403 individual samples at six nuclear microsatellite loci. Markers included five previously published loci (ATH, GC16, LYR132, LYR133, LYR417; van Rossum *et al.*, 2004) and one additional locus (LYR104), whose primer sequences were kindly provided by Thomas Mitchell-Olds, Duke University, Durham (primer F, CTCCATCATCG ATCTCAGCA; primer R, GAGGCGAATGTAGTGGAAGG). Genotyping conditions were as described in van Rossum *et al.* (2004). We located four microsatellite loci (GC16, LYR132, LYR133, ATH) on an interspecific *A. halleri* × *Arabidopsis lyrata petraea* (L.) O'Kane & Al-Shehbaz linkage map (Willems *et al.*, 2007) and determined that they belonged to four distinct linkage groups (C. Godé, unpublished data). The chlorotype dataset

(1256 individual samples), which was fully re-analysed in the present article, brought together data from 13 polymorphic sites in three genomic regions (*trnK*, *trnC-trnD*, *psbC-trnS*) previously published in two separate papers (Pauwels *et al.*, 2005, 2008b).

Analysis of genetic diversity within samples For each sample, microsatellite data were used to calculate the observed (H_O) and expected (H_E) heterozygosity employing GENETIX 4.01 (Laboratoire Génome et Populations, CNRS UPR 9060, Université de Montpellier II, Montpellier, France), F_{IS} and mean allelic richness (A_{Sn}) employing FSTAT 2.9.3. (Goudet, 1995). The deviation of F_{IS} from Hardy–Weinberg equilibrium expectations was assessed by performing 7800 randomizations. Estimates of A_{Sn} were standardized to a standard sample size equal to that of the smallest sample with no missing data ($n = 5$, SLO8). cpDNA data were used to define cpDNA haplotypes or chlorotypes and phylogenetic relationships among chlorotypes (Pauwels *et al.*, 2005, 2008b). For each sample, the allelic or chlorotypic richness (A_{Sc}) was estimated using FSTAT 2.9.3. (Goudet, 1995). Estimates of A_{Sc} were standardized to the smallest sample size ($n = 6$, SLO8). Chloroplast gene diversity (H_S) and its sampling variance were estimated using Arlequin v3.1 (Excoffier *et al.*, 2005).

Bayesian clustering We estimated the number of putative clusters or gene pools K that best explained the pattern of genetic structure at microsatellite loci using Bayesian clustering methods. Previously, we have investigated the linkage disequilibrium between loci within samples using the exact test implemented in GENEPOP 3.4 (Raymond & Rousset, 1995). Structure 2.1 (Pritchard *et al.*, 2000) was used under the admixture model, assuming an identical parameter of individual admixture (alpha, α) across clusters and a uniform prior. For $K = 2$ to $K = 30$, we performed 10 independent runs for 10^5 iterations after a burn-in period of 5×10^4 with no prior information on the origin of individuals. As suggested by Evanno *et al.* (2005), for each K, we computed the ΔK statistics from the mean likelihood $\log_e(K)$ of replicates to decide on a possibly optimal K.

Structure 2.1 results were compared with Bayesian clustering results from TESS 2.0 (Francois *et al.*, 2006; Chen *et al.*, 2007). Compared with Structure 2.1, the algorithm implemented in TESS models spatial dependence in cluster membership using an interaction parameter ψ and therefore allows the testing of the robustness of genetic discontinuities in allele frequencies among clusters, which may be particularly relevant when sampling is unbalanced (Francois *et al.*, 2006). The admixture model was also assumed using the Monte Carlo Markov Chain computing method. The admixture parameter was set to unity. Four values of ψ were used, from $\psi = 0$ to $\psi = 0.9$ (step = 0.3). The algorithm was run 10 times for each value of K, with K ranging from 2 to 30, with total and burn-in numbers of sweeps of 50 000 and 10 000, respectively. The most likely number of clusters was identified using the smallest values of the Deviance Information Criterion (DIC) (Francois *et al.*, 2008).

Once the most likely K had been identified, we used CLUMPP v1.1 (Jakobsson & Rosenberg, 2007) to compute the average individual's membership coefficient to each cluster i (Q_{I_i}) over replicates. Then, the sample membership coefficients (Q_{S_i}) were computed averaging over individuals. Finally, DISTRUCT v1.1 (Rosenberg, 2004) was used to display the clustering results.

Comparison of spatial patterns of nDNA and cpDNA variation We paid particular attention to the congruence between nDNA and cpDNA datasets. First, we performed univariate or multivariate analyses of variance (ANOVA and MANOVA) (GLM procedure of SAS; SAS Institute, 2002; see Methods S1), depending on the K values (see the Results section) to test the effect of the chlorotype carried by an individual (considered as the categorical variable) on the corresponding Q_{I_i} values obtained from nuclear microsatellite data (as dependent variables). The MANOVA was followed by a canonical analysis in order to represent the distribution of chlorotypes (GLM procedure of SAS; Methods S1). Low-frequency chlorotypes (C, M and O) were excluded from this analysis. Second, we used FSTAT 2.9.3 (10 000 permutations) to perform Mantel tests between matrices of pairwise F_{ST} values obtained from SPAGeDi 1.2b (Hardy & Vekemans, 2002). In addition, we used SAS (CORR procedure; SAS Institute, 2002) to perform Pearson correlation tests among nDNA and cpDNA gene diversity or allelic richness estimates.

Patterns of genetic diversity within and among units Data on sample location, clustering results and chlorotype composition were used to assign samples to distinct population units. Mean estimates of A_{Sc}, A_{Sn}, H_E and F_{IS} were computed for each unit and among-unit comparisons were performed using FSTAT 2.9.3. The level of genetic differentiation among units was estimated through hierarchical F statistics using the HIERFSTAT package for R (Goudet, 2005). Tests for the effect of each hierarchical level on genetic structure were performed by randomization (5000 permutations) of the unit defining the level just below that of interest. The contribution of both allele sizes at microsatellite loci and molecular distances between chlorotypes to the level of genetic structure within and among units was investigated through the estimation of R_{ST} and N_{ST} statistics using SPAGeDi 1.2b (Hardy & Vekemans, 2002). Sample pairwise R_{ST} and N_{ST} values were compared with corresponding permutation values, performing 10 000 permutations of either allele sizes among alleles within microsatellite loci or rows and columns of the matrix of molecular distance among chlorotypes. Finally, the contribution of geographic distance to genetic structure within and among units at nuclear loci was assessed by comparing the matrices of log-transformed geographic distances among samples and linearized pairwise genetic distances ($F_{STn}/(1 - F_{STn})$) using a Mantel test with 10 000 permutations.

Ecological niche modelling

We inferred the current and past (21 000 yr before present (BP)) geographic distribution of *A. halleri* in continental Europe using the ecological niche modelling method implemented in MaxEnt 3.2.1 (Phillips *et al.*, 2006). Environmental layers representing 19 ecogeographic variables (EGVs, Table 3) over the area under study, at a resolution of 0.0417 decimal degree or 2.5 arc minutes

(4 km at the equator, 235 939 grid cells with data in total), were computed from publicly available databases. One variable was topographic and corresponded to the digital elevation model provided by the NASA Shuttle Radar Topographic Mission (available at http://csi.cgiar.org/index.asp). Other variables were climatic. They represented seasonal and extreme trends of temperature and precipitation, depicting either the current climate (WorldClim data set; Hijmans *et al.*, 2005) or the climate of the end of the LGM, as simulated by two alternative coupled climate models, the Community Climate System Model (CCSM; http://www.ccsm.ucar.edu) and the Model for Interdisciplinary Research on Climate (MIROC; http://www-pcmdi.llnl.gov/ipcc/model_documentation/MIROC3.2_hires.htm). Climate and palaeoclimate layers were taken from http://www.worldclim.org/past. Data on species occurrence in Europe included the geographic coordinates of 577 *A. halleri* locations sampled in the present study, mentioned in recent studies (Kolník & Marhold, 2006; Koch & Matschinger, 2007) or referenced in the Global Biodiversity Information Facility database (http://www.gbif.org). Duplicate presence records and locations corresponding to regions of recent introduction (northern France, Belgium, North Sea Coast of Lower Saxony in Germany; (Jalas & Suominen, 1994) were removed. Finally, 575 *A. halleri* accessions belonging to 348 grid cells were represented. EGV coefficients were computed from the current climate layers using the default parameter values; 25% of localities were used for model testing.

EGV coefficients were used to project the modelled ecological niche from current and palaeoclimate layers, resulting in predicted current and past geographic patterns of distribution. In both cases, we determined suitable grid cells, that is, cells with a significant output probability of occurrence, employing the 'average probability/suitability approach', which recommends the use of the average raw probability of occurrence over all grid cells of the area under study as the threshold to eliminate unsuitable grid cells (Liu *et al.*, 2005).

Results

Genetic analysis

Genetic diversity within samples All six microsatellite loci scored in this study were highly polymorphic (7–49 alleles per locus). Except for ATH and Lyr104, which displayed one allele of particularly large size, the variation in allele size appeared to be continuous. Moderate variation was observed among loci in H_T (coefficient of variation (CV) = 17.83%), overall F_{ST} (CV = 9.91%) and R_{ST} (CV = 30.69%), whereas stronger variation was observed in F_{IS} (CV = 59.82%).

In comparison with Pauwels *et al.* (2005), three additional chlorotypes (L, M, O) were identified, thus leading to a total of 14 chlorotypes, labelled from A to O (Fig. 1; Table S2). Chlorotypes M and O correspond to expected but missing chlorotypes in Pauwels *et al.* (2005). Chlorotype L derives from chlorotype G by the same mutation (A/T substitution) that links chlorotypes A and D. The homoplasic nature of the mutation was confirmed by sequencing (unpublished data). All chlorotypes are separated
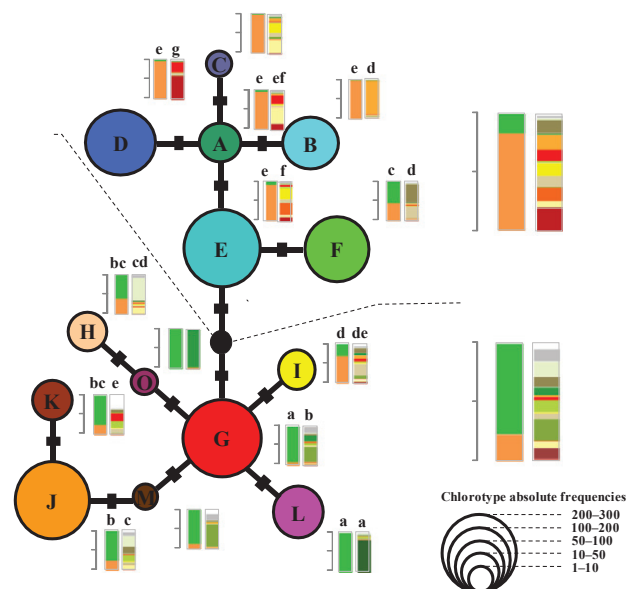

Fig. 1 Genealogical relationships among *Arabidopsis halleri* chlorotypes. Circle sizes are proportional to the absolute frequencies of chlorotypes. The dotted line separates the two main chlorotype lineages (see text in the Results section). In addition to chlorotypes and lineages, histograms represent the mean membership coefficients to each cluster identified by nuclear DNA markers at $K = 2$ (left) and $K = 15$ (right), calculated by averaging over individuals sharing the same chlorotype or chlorotype lineage. Lowercase letters indicate statistical categories as deduced, respectively, from multiple comparison Tukey tests following ANOVAs on Q_{l_i}, or on Q_{l_i} calculated from standardized canonical coefficients on axis 2 of the canonical analysis.

from neighbours by a single detected mutation event, with the notable exception of chlorotypes E and G, which are separated by two detected mutation events (Fig. 1). This higher genetic distance, confirmed by sequencing analysis (unpublished data), and the topology of the chlorotype network suggest the occurrence of two major lineages separating chlorotypes A–F from chlorotypes G–O.

Averaged over locus, sample estimates of nDNA and cpDNA allelic richness (A_{Sn} and A_{Sc}) varied in the ranges 1.96–4.20 and 1–4.89, respectively, observed heterozygosity (H_O) from 0.27 to 0.63, gene diversities (H_E and H_S) from 0.32 to 0.76 and 0 to 0.89, respectively, and inbreeding coefficient (F_{IS}) from −0.05 (not significant) to 0.38 ($P \le 0.001$) (Table S3).

Bayesian clustering Among a total of 915 tests for linkage disequilibrium between pairs of loci within samples, 11 were significant at $P \le 0.01$ and 49 at $P \le 0.05$. None of the global tests for pairs of loci across samples were significant (Fisher's method). We concluded that the analysed loci were sufficiently independent for the application of Bayesian methods for the analysis of population structure.

From Structure 2.1 results, the likelihood $\log_e(K)$ averaged over replicates increased with K, reached a plateau and decreased slightly only after $K = 15$, suggesting that $K = 15$ is the probable true number of clusters (Fig. S1). At $K = 15$, Q_{l_i} values were generally high and individuals from a given sample were consistently
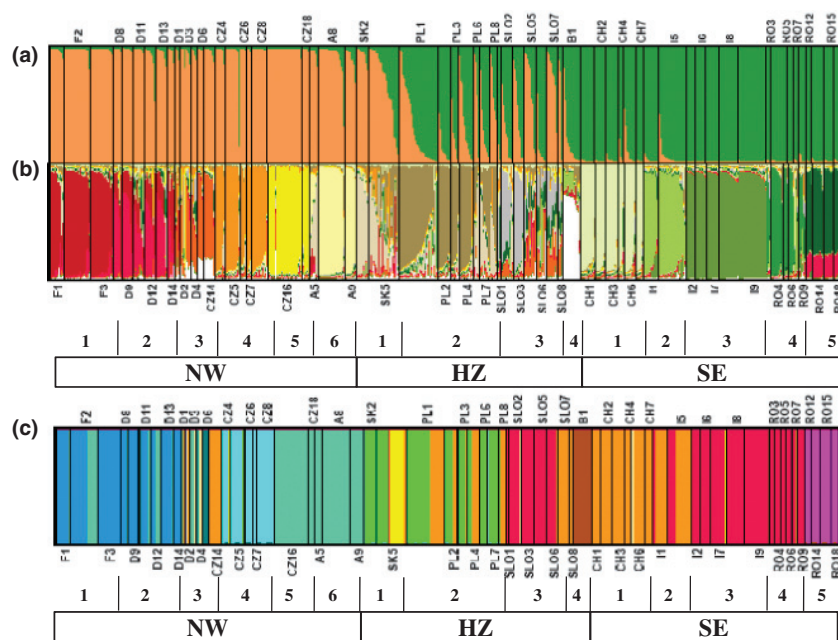
Fig. 2 (a,b) Bayesian clustering results of *Arabidopsis halleri* nuclear DNA markers. Each individual is represented by a vertical line partitioned into K segments corresponding to its membership coefficients at $K = 2$ (a) and $K = 15$ (b). Each colour represents a different gene pool. Vertical black lines separate individuals from different samples. The names of samples and corresponding subunits and units are given (see Table 1 and Supporting Information Table S1 for details). (c) Histograms representing chlorotype relative frequencies in samples. Histogram widths are proportional to sample sizes and each colour represents a different chlorotype. NW, north-western unit; HZ, hybrid zone unit; SE, south-eastern unit.

assigned to a common gene pool, leading to high mean values (Q_{S_i}) with low SE (Fig. 2b). Moreover, geographically proximate samples were often found to belong to the same cluster, suggesting a general island pattern of population structure.

By contrast, the modal value of the distribution of ΔK was located at $K = 2$, suggesting $K = 2$ as an alternative real number of clusters (Fig. S1), revealing a discrepancy in K determination using either the $\log_e(K)$ or ΔK model choice criterion. This was also observed by Evanno *et al.* (2005), who analysed datasets simulated under hierarchical models of population structure. In such cases, ΔK and $\log_e(K)$ were shown to reveal the number of sets of populations and the total number of populations, respectively. $K = 2$ was particularly expected from data simulated under a contact zone hierarchical model, when migration among two structured sets of populations was restricted to peripheral populations. In our case, $K = 2$ distinguished two geographically continuous and nonoverlapping sets of samples, within which Q_{I_i} values were high and consistent among individuals (Fig. 2a). Samples from France, Germany, the Czech Republic and Austria were highly assigned to a similar gene pool, whereas samples from Romania, Italy and Switzerland were assigned to another. In comparison, Q_{I_i} values were lower and heterogeneous in geographically intermediate samples from Poland, Slovakia and Slovenia, suggesting admixture.

Using TESS, only the lower level of population structure was revealed ($K = 15$; Fig. S1). However, Structure 2.1 and TESS gave highly congruent results at both levels of genetic structure (data not shown). Interestingly, TESS results were hardly affected by variation in ψ from 0 to 0.9. Thus, genetic discontinuities among samples, including those at a small geographic scale, for example, among Austrian and Italian samples from the eastern Alps (at $K = 2$ and $K = 15$) and among south German and Czech samples from the Bohemian Forest ($K = 15$ only), were maintained even when the clustering of neighbouring samples was

strongly favoured ($\psi = 0.9$). TESS results thus confirmed that clustering results from Structure 2.1 were not a result of patchy sampling along continuous clines (Francois *et al.*, 2006; Guillot & Santos, 2009).

Comparison of spatial patterns of nDNA and cpDNA variation We observed a strong geographic structure in chlorotype frequencies (Fig. 2c). Most chlorotypes were locally very abundant, if not fixed, and rare or absent in other geographic areas. Consequently, chlorotype mixing was generally low within samples. The chlorotypes belonging to the G–O lineage, which were either rare or absent in Pauwels *et al.* (2005), were quite frequent in most additional samples, that is, at lower latitudes, in Switzerland, Italy, Slovenia and Romania. The co-occurrence of chlorotypes from either the A–F or G–O lineage at comparable frequencies was only detected in Polish and Slovakian samples (F with J and F with I, respectively).

Direct observation of spatial distributions revealed striking correspondence between the nDNA and cpDNA population structure (Fig. 2). In particular, clustering results in the eastern Alps and Bohemian Forest (see the 'Genetic analysis: Bayesian clustering' paragraph of the Results section) were confirmed by cpDNA data: Austrian and Italian samples were associated with chlorotypes from distinct lineages (E vs mostly G and J, respectively); south German and Czech samples shared no chlorotype (Fig. 2c). Congruence between nDNA and cpDNA patterns of genetic structure was confirmed statistically (Methods S1). First, chlorotypes were shown to be highly predictive of individual nuclear genotypic composition (Q_{I_i}) at both levels of genetic structure (ANOVA for $K = 2$, one independent variable: $F_{10,1235} = 259.46$, $P < 0.0001$; MANOVA for $K = 15$, 14 independent variables: Roy's greatest root = 9.45, $P < 0.0001$; Fig. S2). Second, the Mantel test comparing matrices of pairwise genetic distance (F_{ST}) revealed a significant correlation between nDNA

Table 1 Mean estimates of statistics (± SD) describing genetic variation and genetic structure in *Arabidopsis halleri*

(Sub)units	Samples (*n*)	Nuclear diversity					Chloroplast diversity		
		A_{Sn}	H_O	H_E	F_{IS}	F_{STn}	A_{Sc}	H_S	F_{STc}
NW1	F1–3 (3)	2.81 ± 0.03	0.5 ± 0.01	0.53 ± 0.01	0.06 ± 0.01	0.019	1.57 ± 0.51	0.21 ± 0.24	0.258
NW2	D8–14 (6)	3.25 ± 0.31	0.45 ± 0.06	0.57 ± 0.05	0.21 ± 0.1	0.077	2.35 ± 0.54	0.4 ± 0.21	0.146
NW3	D1–6; CZ14 (6)	3.35 ± 0.54	0.54 ± 0.06	0.61 ± 0.07	0.13 ± 0.08	0.146	2.63 ± 1.47	0.47 ± 0.34	0.554
NW4	CZ4–8 (5)	3.42 ± 0.18	0.51 ± 0.03	0.61 ± 0.01	0.16 ± 0.06	0.011	1.28 ± 0.38	0.06 ± 0.08	0.008
NW5	CZ16–18 (2)	2.36 ± 0.03	0.41 ± 0.01	0.43 ± 0	0.06 ± 0.03	0.122	1.1 ± 0.14	0.02 ± 0.02	−0.057
NW6	A5–9 (3)	2.02 ± 0.06	0.31 ± 0.03	0.34 ± 0.02	0.08 ± 0.07	0.255	1 ± 0	0 ± 0	0.000
Differences among NW subunits		**	*	**	0.09	ns	*	*	ns
NW	**(25)**	**3.04 ± 0.58**	**0.47 ± 0.09**	**0.54 ± 0.11**	**0.14 ± 0.09**	**0.298**	**1.85 ± 0.99**	**0.25 ± 0.29**	**0.742**
SE1	CH1–7 (6)	2.36 ± 0.16	0.37 ± 0.05	0.42 ± 0.04	0.11 ± 0.11	0.298	1.64 ± 0.5	0.23 ± 0.19	0.107
SE2	I2; I6–9 (5)	2.21 ± 0.15	0.35 ± 0.06	0.37 ± 0.04	0.05 ± 0.05	0.265	1.38 ± 0.39	0.07 ± 0.06	0.007
SE3	I1; I5 (2)	3.31 ± 0.05	0.53 ± 0.01	0.53 ± 0.04	0 ± 0.06	0.063	2.32 ± 0.45	0.47 ± 0.02	−0.014
SE4	RO3–9 (6)	3.29 ± 0.45	0.46 ± 0.09	0.54 ± 0.09	0.15 ± 0.1	0.238	1.44 ± 0.49	0.11 ± 0.12	−0.023
SE5	RO12–18 (4)	3.34 ± 0.11	0.56 ± 0.04	0.56 ± 0.03	0 ± 0.05	0.372	1.58 ± 0.4	0.13 ± 0.09	−0.036
Differences among SE subunits		**	*	0.070	0.070	ns	ns	ns	ns
SE	**(23)**	**2.82 ± 0.58**	**0.44 ± 0.1**	**0.47 ± 0.1**	**0.08 ± 0.1**	**0.284**	**1.58 ± 0.48**	**0.17 ± 0.16**	**0.706**
HZ1	SK2; SK5 (2)	3.23 ± 0.4	0.5 ± 0.05	0.58 ± 0.04	0.13 ± 0.02	0.096	2.21 ± 0.4	0.43 ± 0.16	0.334
HZ2	PL1–8 (7)	3.12 ± 0.51	0.5 ± 0.08	0.57 ± 0.09	0.13 ± 0.08	0.152	1.93 ± 0.46	0.37 ± 0.19	0.152
HZ3	SLO1–8 (7)	3.4 ± 0.44	0.54 ± 0.08	0.57 ± 0.08	0.06 ± 0.05	0.161	1.96 ± 0.5	0.25 ± 0.17	0.615
HZ4	B1 (1)	2.29	0.35	0.39	0.09	−	1.62	0.13	−
Differences among HZ subunits		*	ns	ns	ns	ns	0.07	ns	ns
HZ	**(17)**	**3.2 ± 0.5**	**0.51 ± 0.08**	**0.56 ± 0.09**	**0.1 ± 0.07**	**0.158**	**1.96 ± 0.44**	**0.31 ± 0.18**	**0.562**
	Total	3 ± 0.57	0.47 ± 0.09	0.52 ± 0.1	0.11 ± 0.09	0.33	1.78 ± 0.72	0.24 ± 0.23	0.74
Differences among units		0.10	*	*	0.06	*	0.10	ns	0.11

Black text indicates comparisons made between samples within subunits, bold text indicates comparisons made between different subunits, and the differences between units are shown at the bottom of the table.

*, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$; ns, not significant.

NW, north-western unit; SE, south-eastern unit; HZ, hybrid zone unit.

and cpDNA datasets ($P = 0.01$, Fig. S3). In addition, weak but significant correlations were detected between chloroplast and nuclear sample allelic richness (A_{Sc} vs A_{Sn}, $R^2 = 0.1254$, $P = 0.0038$) and gene diversity (H_S vs H_E, $R^2 = 0.0992$, $P = 0.0106$).

Patterns of genetic diversity within and among units and subunits Considering the geographic location of the samples and the congruence between genetic data, samples were grouped into geographically distinct population units (Figs 2, 3). Considering $K = 2$ as the real number of clusters, three units were defined. The first unit had a north-western (NW) distribution and included samples showing high Q_{S_1} values ($Q_{S_1} > 0.75$) and chlorotypes A–E. The second unit had a south-eastern (SE) distribution and included samples showing high Q_{S_2} values ($Q_{S_2} > 0.75$) and chlorotypes G, H, J and L. The remaining samples showed high levels of admixture and chlorotypes from both main lineages. These samples were grouped into a third unit, called HZ for Hybrid Zones. Considering $K = 15$ as the real number of clusters led to the definition of 15 subunits within the three units (NW1–NW6, SE1–SE5 and HZ1–HZ4).

Significant differences in the levels of nuclear gene diversity (H_O, H_E) were detected among units (Table 1). Pairwise comparisons, however, revealed that significance was a result of the higher genetic diversity estimates in HZ compared with NW and SE samples (data not shown). Genetic differentiation among units was significant at both nDNA and cpDNA loci (Table 2a), and pairwise differentiation was highest between samples from NW and SE (Table 2b). Although differences in allele sizes at microsatellite loci were never informative, a phylogeographic signal was detected at cpDNA loci between NW and SE (Table 2). Within units, significant differentiation among subunits was detected from both datasets, but pairwise differentiation was lower than among units and no phylogeographic signal was apparent (Tables 1, 2). The level of differentiation among subunits differed significantly among the three units, with lower differentiation among subunits within HZ than within NW and SE (Table 1).

A significant positive correlation between linearized F_{STn} and $\log_e$(geographic distance) was observed for pairwise comparisons among samples within each unit (Fig. 4), suggesting within-unit limitation of gene flow by geographic distance (isolation by distance or IBD pattern; Slatkin, 1993). Similarly, NW–HZ and SE–HZ pairwise comparisons revealed a tendency towards positive correlation between linearized F_{STn} estimators and $\log_e$(geographic distance) (Fig. 4). By contrast, a negative correlation was

Table 2 (a) Hierarchical F statistics estimating the genetic differentiation among each level of genetic structure (unit, subunit and sample) in *Arabidopsis halleri* and (b) mean pairwise estimator of sample differentiation within and between units

(a)	Hierarchical level						
	$F_{U/T}$	P_U	$F_{sU/U}$	P_{sU}	$F_{S/sU}$	P_S	$F_{I/S}$
nDNA	0.07	< 0.001	0.20	< 0.001	0.10	< 0.001	0.10
cpDNA	0.12	< 0.001	0.60	< 0.001	0.28	< 0.001	

(b)	nDNA			cpDNA		
	NW	SE	HZ	NW	SE	HZ
NW	$\overline{F_{STn}} = 0.25$ $\overline{R_{ST}} = 0.17$ $P = 0.76$	$\overline{F_{STn}} = 0.37$ $\overline{R_{ST}} = 0.28$ $P = 0.53$	$\overline{F_{STn}} = 0.26$ $\overline{R_{ST}} = 0.28$ $P = 0.15$	$\overline{F_{STc}} = 0.62$ $\overline{N_{ST}} = 0.58$ $P = 0.58$	$\overline{\mathbf{F_{STc}}} = \mathbf{0.78}$ $\overline{\mathbf{N_{ST}}} = \mathbf{0.85}$ $\mathbf{P = 0.01}$	$\overline{F_{STc}} = 0.71$ $\overline{N_{ST}} = 0.73$ $P = 0.11$
SE		$\overline{F_{STn}} = 0.27$ $\overline{R_{ST}} = 0.34$ $P = 0.11$	$\overline{F_{STn}} = 0.28$ $\overline{R_{ST}} = 0.31$ $P = 0.26$		$\overline{F_{STc}} = 0.55$ $\overline{N_{ST}} = 0.58$ $P = 0.69$	$\overline{F_{STc}} = 0.58$ $\overline{N_{ST}} = 0.57$ $P = 0.51$
HZ			$\overline{\mathbf{F_{STn}}} = \mathbf{0.17}$ $\overline{\mathbf{R_{ST}}} = \mathbf{0.30}$ $\mathbf{P = 0.03}$			$\overline{F_{STc}} = 0.50$ $\overline{N_{ST}} = 0.50$ $P = 0.38$

(a) P_i are estimated P values for the effect of the ith hierarchical level. T, total; U, unit; sU, subunit; S, sample; I, individual.
(b) Differentiation was estimated using allele frequencies only (F_{ST}) or considering the contribution of allele sizes at microsatellite loci (R_{ST}) and molecular distances between chlorotypes (N_{ST}).
Bold indicates $P < 0.05$.
cpDNA, chloroplast DNA; nDNA, nuclear DNA.

detected for NW–SE sample pairs (Fig. 4), suggesting strong barriers to gene flow among these units (Slatkin, 1993; Hardy & Vekemans, 2001). The regression plot (Fig. 4) revealed that the correlation was a result of high genetic differentiation among Austrian (NW6) and Italian (SE2, SE3) samples which, although geographically close, are separated by the highest summits of the central Alps, as represented in Fig. 3.

Ecological niche modelling

MaxEnt uses the score of the area under the receiver operating characteristic curve (AUC) to evaluate the model performance (Phillips *et al.*, 2006). The AUC score was very high (0.945), suggesting a good fit between the model and the data (Fielding & Bell, 1997). EGVs with greatest predictive power were mostly temperature parameters and elevation (Table 3). The most suitable locations for *A. halleri* have intermediate values for the three most determinant EGVs (data not shown). This result is in agreement with a previous analysis showing that *A. halleri* prefers intermediate climatic conditions (Hoffmann, 2005).

The projection onto current climate layers resulted in a putative distribution covering most montane habitats of central Europe, from the Alps and the Carpathians to the Central European Highlands (Fig. S4). The potential distribution was rather continuous, whereas the species only occurs sparsely, in particular in European lowland areas. However, our results were consistent overall with the chorological data-based distribution provided by *Atlas Florae Europaeae* or Hoffmann's (2005) prediction. The projection onto palaeoclimate layers suggested that the species range was probably as large at the end of the LGM as it is currently. However, a shift in the distribution towards the western

and southern parts of Europe was predicted. In particular, the abundance of *A. halleri* at the end of the LGM was probably reduced in the Carpathian Mountains and greater in the Balkans. Finally, considering the extent of the ice sheets covering the Alps up to the end of the LGM (Ray & Adams, 2001), it is almost certain that, despite predictions, the species was not present in most parts of the Alps at that time.

Discussion

Phylogeographic structure of *A. halleri* in Europe

Microsatellite data supported a hierarchical pattern of population structure in *A. halleri*, with two allopatric and divergent genetic units (NW and SE) in north-western and south-eastern Europe, respectively, and substantial population structure within both units, probably as a result of an IBD pattern of gene flow. High genetic differentiation among geographically closest Austrian (NW) and Italian (SE) samples suggested that the highest summits of the Alps have remained a strong physical barrier to gene flow among units in western Europe. By contrast, hybridization was suggested by strong admixture signals along a large zone across Europe, from Slovenia to southern Poland (HZ samples). The observation of an IBD signal among NW or SE and HZ comparisons also supported the occurrence of genetic exchanges among units in this zone (Hardy & Vekemans, 2001) and higher genetic diversity within HZ samples (Comps *et al.*, 2001). Chlorotype data were highly congruent with microsatellite data and further revealed a significant phylogeographic signal among NW and SE samples. Thus, NW and SE genetic units appear to be two well-established phylogeographic units, and can be

Fig. 3 Synthetic representation of the current geographic distribution of detected units and subunits in *Arabidopsis halleri*. North-western (NW), south-eastern (SE) and hybrid zone (HZ) units are delimited by white, black and grey dashed lines, respectively.
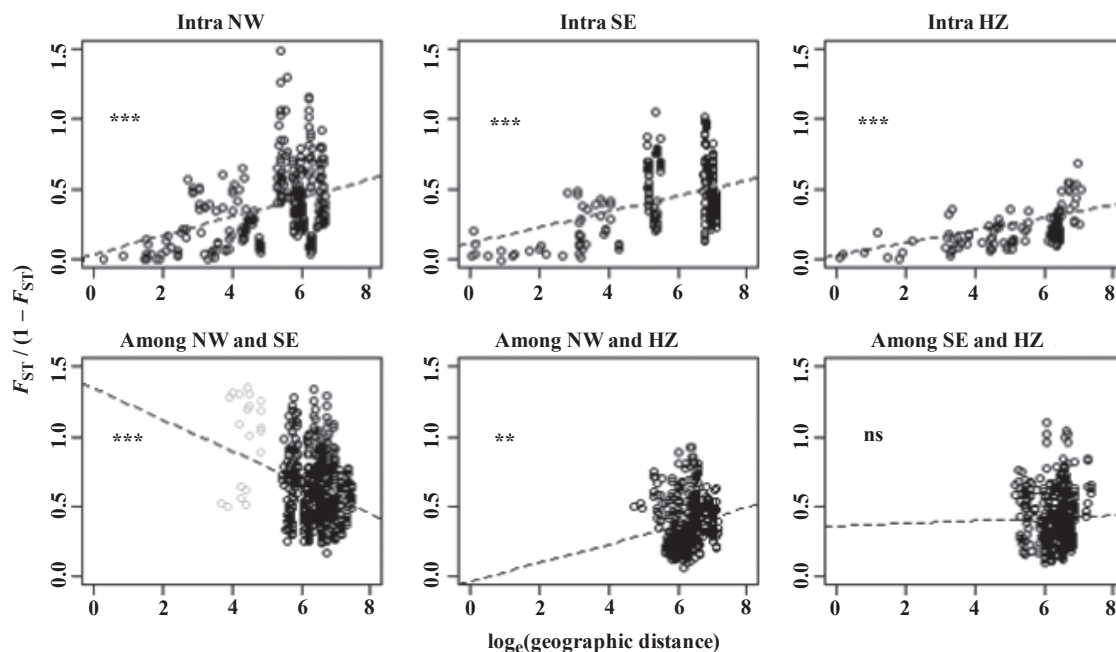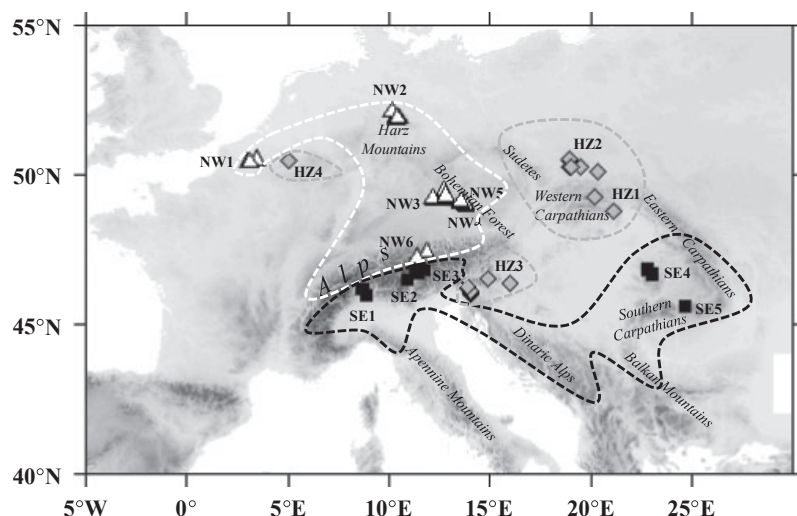
Fig. 4 $F_{ST}/(1 - F_{ST})$ computed for pairs of samples within or among units as a function of the geographic distance (logarithmic scale). Regression lines and significance of Mantel tests of correlation between matrices of geographic and genetic distances are also shown. NW, north-western unit; SE, south-eastern unit; HZ, hybrid zone unit; *, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$; ns, not significant.

assumed to have resulted from vicariance among populations that have experienced a long-term barrier to gene flow (Avise *et al.*, 1987; Avise, 2000; Morrone, 2009). In this context, HZ samples can be assumed to represent a secondary contact zone among previously isolated genetic pools.

Origin of phylogeographic units

Our results suggest that the phylogeographic structure detected in *A. halleri* in Europe has resulted from the isolation of large population units on either side of the European Alps during the LGM. According to ecological niche modelling results, the species did not suffer large reductions of its range during the LGM. However, the species occurrence was then probably

reduced in the Carpathians, and must have been absent from most of the European Alps, largely covered by ice sheets (Ray & Adams, 2001). This latter assumption is corroborated by lower levels of genetic diversity in the Alpine (Austrian, Italian and Swiss) samples. In summary, the LGM species range must have mainly consisted of at least two large continuous populations, towards the Dinaric Alps in the Balkan Peninsula, and in the Central European Highlands. The LGM ice extent over the Alps (Ray & Adams, 2001; Schonswetter *et al.*, 2005) must have represented a strong physical barrier to gene flow between these populations, thus favouring vicariance.

Such a phylogeographic scenario differs from current paradigm scenarios for European taxa. As for numerous temperate species, we assumed only a few widespread phylogeographic units (Taberlet

Table 3 Bioclimatic variables used for the ecological niche modelling of *Arabidopsis halleri* and contribution of the variables to the model

Name	Description	Range (units) in area of interest			% contribution	Permutation importance
		Current climate (WorldClim)	Palaeoclimate layers (CCSM)	Palaeoclimate layers (MIROC)		
BIOCLIM 4	Temperature seasonality (SD × 100)	3586 to 9118 (°C × 100)	3442 to 9042 (°C × 100)	3442 to 9097 (°C × 100)	27	19
BIOCLIM 9	Mean temperature of the driest quarter	−107 to 252 (°C × 10)	−254 to 236 (°C × 10)	−220 to 231 (°C × 10)	16.3	16.7
BIOCLIM 11	Mean temperature of the coldest quarter	−143 to 113 (°C × 10)	−334 to 53 (°C × 10)	−234 to 105 (°C × 10)	9.9	1.1
BIOCLIM 10	Mean temperature of the warmest quarter	−25 to 257 (°C × 10)	−102 to 236 (°C × 10)	−46 to 231 (°C × 10)	9.1	1
BIOCLIM 5	Maximum temperature of warmest month	4 to 336 (°C × 10)	−17 to 328 (°C × 10)	−17 to 303 (°C × 10)	8.1	6.6
SRTM	Elevation	−28 to 3638 (m)	−28 to 3638 (m)	−28 to 3638 (m)	7.1	13.1
BIOCLIM 1	Mean annual temperature	−87 to 172 (°C × 10)	−187 to 157 (°C × 10)	−131 to 162 (°C × 10)	5.4	1.2
BIOCLIM 3	Isothermality (BIO2/BIO7) (× 100)	19 to 43	24 to 76	20 to 53	4.2	7.1
BIOCLIM 14	Precipitation of the driest period	2 to 172 (mm)	0 to 184 (mm)	1 to 204 (mm)	3.5	3.9
BIOCLIM 2	Mean diurnal range	42 to 124 (°C × 10)	50 to 412 (°C × 10)	40 to 132 (°C × 10)	3.5	3.9
BIOCLIM 18	Precipitation of the warmest quarter	41 to 593 (mm)	32 to 628 (mm)	36 to 700 (mm)	1.9	16.2
BIOCLIM 15	Precipitation seasonality (CV)	7 to 65 (mm)	7 to 67 (mm)	7 to 67 (mm)	1.3	3.4
BIOCLIM 12	Annual precipitation	312 to 2543 (mm)	299 to 2720 (mm)	217 to 3016 (mm)	0.9	1.7
BIOCLIM 6	Minimum temperature of coldest month	−174 to 85 (°C × 10)	−482 to 54 (°C × 10)	−300 to 67 (°C × 10)	0.8	2.8
BIOCLIM 17	Precipitation of the driest quarter	22 to 562 (mm)	13 to 601 (mm)	20 to 667 (mm)	0.4	1
BIOCLIM 7	Temperature annual range (BIO5 to BIO6)	157 to 344 (°C × 10)	151 to 627 (°C × 10)	147 to 372 (°C × 10)	0.3	0.9
BIOCLIM 13	Precipitation of the wettest period	34 to 281 (mm)	33 to 264 (mm)	24 to 347 (mm)	0.2	0.1
BIOCLIM 8	Mean temperature of the wettest quarter	−139 to 213 (°C × 10)	−216 to 169 (°C × 10)	−160 to 189 (°C × 10)	0.1	0.2
BIOCLIM 16	Precipitation of the wettest quarter	94 to 741 (mm)	90 to 744 (mm)	65 to 914 (mm)	0	0
BIOCLIM 19	Precipitation of the coldest quarter	63 to 710 (mm)	54 to 524 (mm)	44 to 876 (mm)	0	0

CCSM, Community Climate System Model; MIROC, Model for Interdisciplinary Research on Climate.

et al., 1998; Hewitt, 1999; Schonswetter *et al.*, 2005; Weiss & Ferrand, 2007). However, the geographic distribution of units in temperate taxa suggested that refugial populations were located in the southern European peninsula. In some cases, nongenetic data supported the existence of smaller refugia in central Europe, but with no significant impact on large-scale population genetics ('cryptic refugia'; see Stewart & Lister, 2001; Provan & Bennett, 2008). By contrast, glacial survival at northern latitudes was commonly assumed for Alpine species (Schonswetter *et al.*, 2005; Holderegger & Thiel-Egenter, 2009). However, refugia were probably multiple, spatially restricted (microrefugia, *sensu* Rull, 2009) and close to the border or in central parts of the Alps. However, our results corroborated recent studies focused on the glacial history of herbaceous subalpine and montane species, including *A. lyrata* ssp. *petraea*, the closest relative of *A. halleri*, in Europe (Clauss & Mitchell-Olds, 2006; Mráz *et al.*, 2007; Ronikier *et al.*, 2008; Huck *et al.*, 2009; Ansell *et al.*, 2010). For such species, the existence of glacial macrorefugia in the lowland areas of central Europe, north of the Alps, was proposed (Ronikier *et al.*, 2008). Although the landscape was largely covered by permafrost, herbaceous subalpine and montane species probably occupied steppe–tundra habitats (Williams *et al.*, 1998; Schonswetter *et al.*, 2005).

Together, recent phylogeographic results on species that occur in both mountain and lowland habitats, in the Alps and the Carpathians, such as *A. halleri*, suggest that they could share an alternative paradigmatic phylogeographic pattern. This would confirm the role of among-species variation in ecological traits, in particular tolerance to climatic factors, on species response to climatic oscillations (Bhagwat & Willis, 2008). However, the exact

pattern remains unclear. Indeed, in addition to the lowlands or highlands of central Europe, *in situ* glacial survival in one or several areas within the Carpathians has also been hypothesized (Mráz *et al.*, 2007; Puşcaş, *et al.*, 2008; Ronikier *et al.*, 2008; Thiel-Egenter *et al.*, 2009). Although this was not supported by our data, the long-term survival of *A. halleri* in the Carpathians could explain the occurrence of specific chlorotypes in elevated frequency in Polish, Slovak and Romanian samples of *A. halleri* (chlorotypes F, I and L, respectively). It should be mentioned, however, that local haplotypes in certain regions may also result from interspecies hybridization, a phenomenon that seems to be common in the *Arabidopsis* genus (Schmickl *et al.*, 2010; Schmickl & Koch, 2011). Obviously, a more intensive sampling effort in the Carpathians would be necessary to discuss more precisely the phylogeographic history of *A. halleri* in this region.

Comparison of phylogeographic and ecogeographic units

In 2005, Pauwels *et al.* suggested that the recent colonization of metal-polluted soils had no influence on the *A. halleri* population genetic structure. However, this suggestion was based on limited information as the study only covered a restricted portion of the European species range, was based on cpDNA data only and did not detect any phylogeographic signal. In the present study, we revealed from cpDNA data that there is a phylogeographic break among European populations, and that phylogeographic units differ from edaphic ecotypes. Indeed, although the NW unit comprises both M and NM samples, the fact that M populations also occur in Italy (e.g. Shahzad *et al.*, 2010) strongly suggests that the SE unit also contains both ecotypes. As candidate

genes for the evolution of metal tolerance are exclusively nuclear (Roosens *et al.*, 2008a,b), one could assume that the convergent evolution of phylogeographically distant M populations towards higher tolerance levels (see Pauwels *et al.*, 2006) is associated with nuclear gene flow among units. However, the high level of congruence between nDNA and cpDNA population genetic patterns suggests that such gene flow did not occur. Together, our new results confirm that the colonization of polluted soils is more recent than the vicariance among phylogeographic units, occurred independently in both units and did not generate additional genetic breaks.

Unfortunately, our study did not include any voucher specimens, therefore reducing the strength of comparisons made between identified phylogeographic units and previously defined morphological subspecies. However, such questions open up very interesting new perspectives for future phylogeographic studies (Zink & Barrowclough, 2008). In our case, population genetic structure did not match the division of the species into the two European subspecies described by Al-Shehbaz & O'Kane (2002), and therefore our results call for a revision of subspecies delimitation, as suggested previously by Kolník & Marhold (2006) and Koch *et al.* (2008). The delimitation of two phylogeographic units also does not seem to support the definition of four European subspecies (Kolník & Marhold, 2006), but we suggest that the phylogeographic results and subspecies delimitation are not totally incongruent. Indeed, one subspecies, *A. h.* ssp. *ovirensis*, is only known from a single locality in eastern Austria (Kolník & Marhold, 2006; Koch *et al.*, 2008) and is probably missing from our sampling. The morphological type described as *A. h.* ssp. *tatrica* in Kolník & Marhold (2006) is endemic to the western Carpathians. It seems to include intermediates between ssp. *halleri* and ssp. *ovirensis* (Jones & Akeroyd, 1993 cited in Al-Shehbaz & O'Kane, 2002 and Kolník & Marhold, 2006). This observation fits remarkably well with the high levels of genetic admixture in samples from this region, suggesting that HZ samples may include *A. h.* ssp. *tatrica* specimens. In this case, the NW and SE units described could correspond to *A. h.* ssp. *halleri* and *A. h.* ssp. *dacica*, respectively. However, although, according to Kolník & Marhold (2006), the distribution of the taxa may be incompletely characterized, *A. h.* ssp. *dacica* occurs in the eastern and western Carpathians and the Balkan peninsula. In comparison, the distribution of the SE units clearly extends towards northern Italy and southern Switzerland (subunits SE1–SE3; Figs 2, 3).

Implications of phylogeographic structure for the genetic study of adaptive traits in *A. halleri*

Arabidopsis halleri is an emerging model involved in numerous studies using genomics, transcriptomics and quantitative trait loci mapping to unravel the genetic bases of metal tolerance in plants (e.g. Weber *et al.*, 2006; Willems *et al.*, 2007; Hanikenne *et al.*, 2008; Frérot *et al.*, 2010). A recent study suggested that metal tolerance evolved in association with speciation of *A. halleri* about 300 000 years ago (Roux *et al.*, 2011). As it is now clear that metal tolerance shows a quantitative polymorphism

differentiating M from NM accessions in *A. halleri* (Pauwels *et al.*, 2006; Meyer *et al.*, 2010), future mechanistic studies should undoubtedly also include the comparison between M and NM populations (e.g. Meyer *et al.*, 2009; Shahzad *et al.*, 2010). However, the noncongruence among ecotypes and phylogeographic units suggests that adaptive differentiation for tolerance traits between NM and M populations is more recent than vicariance among phylogeographic units. This agrees with the assumption that traits under selection allow the identification of a relatively recent barrier to gene flow (Zink & Barrowclough, 2008). This also fits with the fact that M populations occur on metalliferous sites of recent anthropogenic origin, and with the assumption that the colonization of metalliferous sites occurred independently in distant geographic areas (Pauwels *et al.*, 2005). However, we have further demonstrated that adaptive differentiation could have occurred several times independently in distinct phylogeographic units, that is, in vicariant gene pools that have suffered high genetic isolation over long periods of time. This significantly reduces the possibility of common ancestry of adaptive genes in phenotypically similar populations from different phylogeographic units (the trait originated in one unit and was secondarily transferred in others by hybridization; Rieseberg & Wendel, 1993; Vekemans, 2010). Alternatively, this strongly suggests that the evolution of populations towards similar ecological trait values in either phylogeographic unit reflects convergent evolution, potentially through distinct evolutionary pathways, so that a similar phenotype may not indicate comparable genetic composition at functional loci (Alonso-Blanco *et al.*, 2009). Thus, M or NM accessions from either phylogeographic unit may not be genetically exchangeable (Crandall *et al.*, 2000; Templeton, 2001). Finally, in contrast with many recent metal tolerance studies (e.g. Hanikenne *et al.*, 2008; Frérot *et al.*, 2010; Farinati *et al.*, 2011; Meyer *et al.*, 2011), which mainly used accessions from the northern edge of the species range as representative of the species (in Europe), we suggest that future studies should consider accessions from vicariant phylogeographic units as potentially representing independent selection events from differentiated gene pools.

References

Alonso-Blanco C, Aarts MG, Bentsink L, Keurentjes JJ, Reymond M, Vreugdenhil D, Koornneef M. 2009. What has natural variation taught us about plant development, physiology, and adaptation? *Plant Cell* **21**: 1877–1896.

Al-Shehbaz IA, O'Kane SLJ. 2002. Taxonomy and Phylogeny of *Arabidopsis* (Brassicaceae). In: Somerville CR, Meyerowitz EM, eds. *The Arabidopsis book*. Rockville, MD, USA: American Society of Plant Biologist, 1–22.

Ansell SW, Stenoien HK, Grundmann M, Schneider H, Hemp A, Bauer N, Russell SJ, Vogel JC. 2010. Population structure and historical biogeography of European *Arabidopsis lyrata*. *Heredity* **105**: 543–553.

Avise JC. 1989. Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* **43**: 1192–1208.

Avise JC. 2000. *Phylogeography: the history and formation of species*. Cambridge, MA, USA and London, UK: Harvard University Press.

Avise JC, Arnold J, Ball RMJ, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* **7**: 489–522.

Bhagwat SA, Willis KJ. 2008. Species persistence in northerly glacial refugia of Europe: a matter of chance or biogeographical traits? *Journal of Biogeography* **35**: 464–482.

Bierne N, Welch J, Loire E, Bonhomme F, David P. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology* **20**: 2044–2072.

Bomblies K, Weigel D. 2010. *Arabidopsis* and relatives as models for the study of genetic and genomic incompatibilities. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 1815–1823.

Brej T, Fabiszewski J. 2006. Plants accumulating heavy metals in the Sudety Mts. *Acta Societatis Botanicorum Poloniae* **75**: 61–68.

Chen C, Durand E, Forbes F, Francois O. 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7**: 747–756.

Clauss MJ, Mitchell-Olds T. 2006. Population genetic structure of *Arabidopsis lyrata* in Europe. *Molecular Ecology* **15**: 2753–2766.

Comps B, Gomory D, Letouzey J, Thiebaut B, Petit RJ. 2001. Diverging trends between heterozygosity and allelic richness during postglacial colonization in the European beech. *Genetics* **157**: 389–397.

Crandall KA, Bininda-Emonds OR, Mace GM, Wayne RK. 2000. Considering evolutionary processes in conservation biology. *Trends in Ecology and Evolution* **15**: 290–295.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**: 2611–2620.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**: 47–50.

Farinati S, DalCorso G, Panigati M, Furini A. 2011. Interaction between selected bacterial strains and *Arabidopsis halleri* modulates shoot proteome and cadmium and zinc accumulation. *Journal of Experimental Botany* **62**: 3433–3447.

Fielding AH, Bell JF. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**: 38–49.

Francois O, Ancelet S, Guillot G. 2006. Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**: 805–816.

Francois O, Blum MG, Jakobsson M, Rosenberg NA. 2008. Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genetics* **4**: e1000075.

Frérot H, Faucon M-P, Willems G, Godé C, Courseaux A, Darracq A, Verbruggen N, Saumitou-Laprade P. 2010. Genetic architecture of zinc

hyperaccumulation in *Arabidopsis halleri*: the essential role of QTL × environment interactions. *New Phytologist* **187**: 355–367.

Goudet J. 1995. FSTAT version 1.2: a computer program to calculate *F*-statistics. *Journal of Heredity* **86**: 485–486.

Goudet J. 2005. HIERFSTAT, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes* **5**: 184–186.

Guillot G, Santos F. 2009. A computer program to simulate multilocus genotype data with spatially autocorrelated allele frequencies. *Molecular Ecology Resources* **9**: 1112–1120.

Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A, Motte P, Kroymann J, Weigel D, Kramer U. 2008. Evolution of metal hyperaccumulation required *cis*-regulatory changes and triplication of *HMA4*. *Nature* **453**: 391–395.

Hardy OJ, Vekemans X. 2001. Patterns of allozyme variation in diploid and tetraploid *Centaurea jacea* at different spatial scales. *Evolution* **55**: 943–954.

Hardy OJ, Vekemans X. 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology* **2**: 618–620.

Hewitt GM. 1999. Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society* **68**: 87–112.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**: 1965–1978.

Hoffmann MH. 2005. Evolution of the realized climatic niche in the genus *Arabidopsis* (Brassicaceae). *Evolution* **59**: 1425–1436.

Holderegger R, Thiel-Egenter C. 2009. A discussion of different types of glacial refugia used in mountain biogeography and phylogeography. *Journal of Biogeography* **36**: 476–480.

Huck S, Büdel B, Kadereit JW, Printzen C. 2009. Range-wide phylogeography of the European temperate-montane herbaceous plant *Meum athamanticum* Jacq.: evidence for periglacial persistence. *Journal of Biogeography* **36**: 1588–1599.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**: 1801–1806.

Jalas J, Suominen J, eds. 1994. *Atlas Florae Europaeae. Distribution of Vascular Plants in Europe. Vol 10. Cruciferae (Sisymbrium to Aubrieta)*. Helsinki, Finland: The Committee for Mapping the Flora of Europe & Societas Biologica Fennica Vanamo.

Koch MA, Matschinger M. 2007. Evolution and genetic differentiation among relatives of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **104**: 6272–6277.

Koch M, Wernisch M, Schmickl R. 2008. *Arabidopsis thaliana*'s wild relatives: an updated overview on systematics, taxonomy and evolution. *Taxon* **57**: 933–943.

Kolník M, Marhold K. 2006. Distribution, chromosome numbers and nomenclature conspect of *Arabidopsis halleri* (Brassicaceae) in the Carpathians. *Biologia* **61**: 41–50.

Kruckeberg AR. 1986. An essay: The stimulus of unusual geologies for plant speciation. *Systematic Botany* **11**: 455–463.

Liu C, Berry PM, Dawson TP, Pearson RG. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* **28**: 385–393.

Macnair M, Gardner M. 1998. The evolution of edaphic endemics. In: Howard DJ, Berlocher SH, eds. *Endless forms, species and speciation*. New York, NY, USA: Oxford University Press, 157–171.

Mallet J. 2001. Subspecies, semispecies, superspecies. In: Levin SA, ed. *Encyclopedia of biodiversity*. San Diego, CA, USA: Academic Press, 523–526.

Meyer C-L, Kostecka AA, Saumitou-Laprade P, Créach A, Castric V, Pauwels M, Frérot H. 2010. Variability of zinc tolerance among and within populations of the pseudometallophyte species *Arabidopsis halleri* and possible role of directional selection. *New Phytologist* **185**: 130–142.

Meyer C-L, Peisker D, Courbot M, Craciun AR, Cazale AC, Desgain D, Schat H, Clemens S, Verbruggen N. 2011. Isolation and characterization of *Arabidopsis halleri* and *Thlaspi caerulescens* phytochelatin synthases. *Planta* **234**: 83–95.

Meyer C-L, Vitalis R, Saumitou-Laprade P, Castric V. 2009. Genomic pattern of adaptive divergence in *Arabidopsis halleri*, a model species for tolerance to heavy metal. *Molecular Ecology* **18**: 2050–2062.

Montarry J, Andrivon D, Glais I, Corbiere R, Mialdea G, Delmotte F. 2010. Microsatellite markers reveal two admixed genetic groups and an ongoing displacement within the French population of the invasive plant pathogen *Phytophthora infestans*. *Molecular Ecology* **19**: 1965–1977.

Morrone JJ. 2009. *Evolutionary biogeography. An integrative approach with case studies*. New York, NY, USA: Columbia University Press.

Mráz P, Gaudeul M, Rioux D, Gielly L, Choler P, Taberlet P. 2007. Genetic structure of *Hypochaeris uniflora* (Asteraceae) suggests vicariance in the Carpathians and rapid post-glacial colonization of the Alps from an eastern alpine refugium. *Journal of Biogeography* **34**: 2100–2114.

Pauwels M, Frerot H, Bonnin I, Saumitou-Laprade P. 2006. A broad-scale analysis of population differentiation for Zn tolerance in an emerging model species for tolerance study: *Arabidopsis halleri* (Brassicaceae). *Journal of Evolutionary Biology* **19**: 1838–1850.

Pauwels M, Roosens N, Frerot H, Saumitou-Laprade P. 2008a. When population genetics serves genomics: putting adaptation back in a spatial and historical context. *Current Opinion in Plant Biology* **11**: 129–134.

Pauwels M, Saumitou-Laprade P, Holl A-C, Petit D, Bonnin I. 2005. Multiple origin of metallicolous populations of the pseudometallophyte *Arabidopsis halleri* (Brassicaceae) in central Europe: the cpDNA testimony. *Molecular Ecology* **14**: 4403–4414.

Pauwels M, Willems G, Roosens N, Frérot H, Saumitou-Laprade P. 2008b. Merging methods in molecular and ecological genetics to study the adaptation of plants to anthropogenic metal-polluted sites: implications for phytoremediation. *Molecular Ecology* **17**: 108–119.

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modelling of species geographic distributions. *Ecological Modelling* **190**: 231–259.

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

Provan J, Bennett KD. 2008. Phylogeographic insights into cryptic glacial refugia. *Trends in Ecology and Evolution* **23**: 564–571.

Puşcaş M, Choler P, Tribsch A, Gielly L, Rioux D, Gaudeul M, Taberlet P. 2008. Post-glacial history of the dominant alpine sedge *Carex curvula* in the European alpine system inferred from nuclear and chloroplast markers. *Molecular Ecology* **17**: 2417–2429.

Rajakaruna N. 2004. The edaphic factor in the origin of plant species. *International Geology Review* **46**: 471–478.

Ray N, Adams JM. 2001. A GIS-based vegetation map of the world at the last glacial maximum (25,000–15,000 bp). *Internet Archaeology*. **11**. URL http://intarch.ac.uk/journal/issue11/rayadams_toc.html

Raymond M, Rousset F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**: 248–249.

Rieseberg LH, Wendel JF. 1993. Introgression and its consequences in plants. In: Harrison R, ed. *Hybrid zones and the evolutionary process*. Oxford, UK: Oxford University Press, 70–109.

Ronikier M, Costa A, Aguilar JF, Feliner GN, Küpfer P, Mirek Z. 2008. Phylogeography of *Pulsatilla vernalis* (L.) Mill. (Ranunculaceae): chloroplast DNA reveals two evolutionary lineages across Central Europe and Scandinavia. *Journal of Biogeography* **35**: 1650–1664.

Roosens NH, Willems G, Godé C, Courseaux A, Saumitou-Laprade P. 2008a. The use of comparative genome analysis and syntenic relationships allows extrapolating the position of Zn tolerance QTL regions from *Arabidopsis halleri* into *Arabidopsis thaliana*. *Plant and Soil* **306**: 105–116.

Roosens NH, Willems G, Saumitou-Laprade P. 2008b. Using *Arabidopsis* to explore zinc tolerance and hyperaccumulation. *Trends in Plant Science* **13**: 208–215.

Rosenberg NA. 2004. DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**: 137–138.

Roux C, Castric V, Pauwels M, Wright SI, Saumitou-Laprade P, Vekemans X. 2011. Does speciation between *Arabidopsis halleri* and *Arabidopsis lyrata* coincide with major changes in a molecular target of adaptation. *PLoS ONE* **6**: e26872.

van Rossum F, Bonnin I, Fenart S, Pauwels M, Petit D, Saumitou-Laprade P. 2004. Spatial genetic structure within a metallicolous population of *Arabidopsis halleri*, a clonal, self-incompatible and heavy-metal-tolerant species. *Molecular Ecology* **13**: 2959–2967.

Rull V. 2009. Microrefugia. *Journal of Biogeography* **36**: 481–484.

SAS Institute. 2002. *SAS User's Guide: Statistics. Version 9.1.* Cary, NC, USA: SAS Institute.

Schmickl R, Jørgensen MH, Brysting AK, Koch MA. 2010. The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evolutionary Biology* **10**: 98.

Schmickl R, Koch MA. 2011. *Arabidopsis* hybrid speciation processes. *Proceedings of the National Academy of Sciences, USA* **108**: 14192–14197.

Schonswetter P, Stehlik I, Holderegger R, Tribsch A. 2005. Molecular evidence for glacial refugia of mountain plants in the European Alps. *Molecular Ecology* **14**: 3547–3555.

Shahzad Z, Gosti F, Frérot H, Lacombe E, Roosens N, Saumitou-Laprade P, Berthomieu P. 2010. The five *AHMTP1* zinc transporters undergo different evolutionary fates towards adaptive evolution to zinc tolerance in *Arabidopsis halleri*. *PLoS Genetics* **6**: e1000911.

Slatkin M. 1993. Isolation by distance in equilibrium and nonequilibrium populations. *Evolution* **47**: 264–279.

Stewart J, Lister A. 2001. Cryptic northern refugia and the origins of the modern biota. *Trends in Ecology and Evolution* **16**: 608–613.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F. 1998. Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**: 453–464.

Templeton AR. 2001. Using phylogeographic analyses of gene trees to test species status and processes. *Molecular Ecology* **10**: 779–791.

Thiel-Egenter C, Holderegger R, Brodbeck S, Gugerli F. 2009. Concordant genetic breaks, identified by combining clustering and tessellation methods, in two co-distributed alpine plant species. *Molecular Ecology* **18**: 4495–4507.

Thorpe RS, Malhotra A, Black H, Daltry JC, Wuster W. 1995. Relating geographic pattern to phylogenetic process. *Philosophical Transactions of the Royal Society B: Biological Sciences* **349**: 61–68.

Turesson G. 1922. The species and the variety as ecological units. *Hereditas* **3**: 100–113.

Vallender R, Robertson RJ, Friesen VL, Lovette IJ. 2007. Complex hybridization dynamics between golden-winged and blue-winged warblers (*Vermivora chrysoptera* and *Vermivora pinus*) revealed by AFLP, microsatellite, intron and mtDNA markers. *Molecular Ecology* **16**: 2017–2029.

Vekemans X. 2010. What's good for you may be good for me: evidence for adaptive introgression of multiple traits in wild sunflower. *New Phytologist* **187**: 6–9.

Verbruggen N, Hermans C, Schat H. 2009. Molecular mechanisms of metal hyperaccumulation in plants. *New Phytologist* **181**: 759–776.

Waltari E, Hijmans RJ, Peterson AT, Nyari AS, Perkins SL, Guralnick RP. 2007. Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS ONE* **2**: e563.

Weber M, Trampczynska A, Clemens S. 2006. Comparative transcriptome analysis of toxic metal responses in *Arabidopsis thaliana* and the Cd^{2+}-hypertolerant facultative metallophyte *Arabidopsis halleri*. *Plant, Cell & Environment* **29**: 950–963.

Weiss S, Ferrand N, eds. 2007. *Phylogeography of southern European refugia: evolutionary perspectives on the origins and conservation of European biodiversity*. Berlin, Germany: Springer.

Willems G, Drager DB, Courbot M, Godé C, Verbruggen N, Saumitou-Laprade P. 2007. The genetic basis of zinc tolerance in the metallophyte *Arabidopsis halleri* ssp. *halleri* (Brassicaceae): an analysis of quantitative trait loci. *Genetics* **176**: 659–674.

Williams M, Dunkerley D, De Deckker P, Kershaw P, Chappell J. 1998. *Quaternary environments*. London, UK: Arnold.

Wilson EO, Brown WL. 1953. The subspecies concept and its taxonomic application. *Systematic Zoology* **2**: 97–111.

Winker K. 2010. Subspecies represent geographically partitioned variation, a gold mine of evolutionary biology, and a challenge for conservation. *Ornithological Monographs* **67**: 6–23.

Zink RM, Barrowclough GF. 2008. Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology* **17**: 2107–2121.

Supporting Information

Additional supporting information may be found in the online version of this article. [Correction added after online publication 17 Jan 2012: the Supporting Information file has been renamed to replace 'Notes S1' with 'Methods S1'.]

Fig. S1 Bayesian clustering results of nuclear DNA (nDNA) markers obtained using either Structure or TESS.

Fig. S2 Plot of chlorotypes on the first two axes of canonical analysis.

Fig. S3 Distribution of among-sample pairwise genetic distance at nuclear DNA (nDNA) loci as a function of the pairwise genetic distance at chloroplast DNA (cpDNA) loci.

Fig. S4 Present and past geographic range of *Arabidopsis halleri* in Europe.

Table S1 Locations of sampled populations of *Arabidopsis halleri* and sample sizes for chloroplast DNA (cpDNA) and nuclear microsatellite analysis

Table S2 Distribution of chlorotypes in samples, subgroups and groups

Table S3 Genetic diversity indices at nuclear DNA (nDNA) and chloroplast DNA (cpDNA) loci for each sample

Methods S1 Analysis of variance testing the effect of chlorotype composition on genetic composition at nuclear loci.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

About *New Phytologist*

- *New Phytologist* is an electronic (online-only) journal owned by the New Phytologist Trust, a **not-for-profit organization** dedicated to the promotion of plant science, facilitating projects from symposia to free access for our Tansley reviews.

- Regular papers, Letters, Research reviews, Rapid reports and both Modelling/Theory and Methods papers are encouraged. We are committed to rapid processing, from online submission through to publication 'as ready' via *Early View* – our average time to decision is <25 days. There are **no page or colour charges** and a PDF version will be provided for each article.

- The journal is available online at Wiley Online Library. Visit **www.newphytologist.com** to search the articles and register for table of contents email alerts.

- If you have any questions, do get in touch with Central Office (np-centraloffice@lancaster.ac.uk) or, if it is more convenient, our USA Office (np-usaoffice@ornl.gov)

- For submission instructions, subscription and all the latest information visit **www.newphytologist.com**