Natural Genetic Variation of *Arabidopsis thaliana* is Geographically Structured in the Iberian Peninsula

F. Xavier Picó*, Belén Méndez-Vigo†, José M. Martínez-Zapater† and Carlos Alonso-Blanco†

*: Estación Biológica de Doñana, CSIC, Sevilla-41013, Spain.
†: Departamento de Genética Molecular de Plantas, Centro Nacional de Biotecnología (CNB), Consejo Superior de Investigaciones Científicas (CSIC), Madrid-28049, Spain.
Running title: A. thaliana population structure in Iberia

Keywords: Arabidopsis, natural variation, genetic structure, isolation by distance, genome-wide analysis

Corresponding author:
Carlos Alonso-Blanco
Genética Molecular de Plantas
Centro Nacional de Biotecnología (CNB-CSIC)
Consejo Superior de Investigaciones Científicas (CSIC)
C/Darwin 3
Cantoblanco
Madrid 28049
Spain

e-mail: calonso@cnb.csic.es

Phone # (34) 915854688
Fax # (34) 915854506
Abstract

To understand the demographical history of *Arabidopsis thaliana* within its native geographic range, we have studied its genetic structure in the Iberian Peninsula region. We have analysed the amount and spatial distribution of *A. thaliana* genetic variation by genotyping 268 individuals sampled in 100 natural populations from the Iberian Peninsula. Analyses of 175 individuals from seven of these populations, with 20 chloroplast and nuclear microsatellite loci and 109 common single nucleotide polymorphisms, show significant population differentiation and isolation by distance. In addition, analyses of one genotype from 100 populations detected significant isolation by distance over the entire Peninsula, as well as among six Iberian subregions. Analyses of these 100 genotypes with different model-based clustering algorithms inferred four genetic clusters, which show a clear cut geographical differentiation pattern. On the other hand, clustering analysis of a world-wide sample showed a west-east Eurasian longitudinal spatial gradient of the commonest Iberian genetic cluster. These results indicate that *A. thaliana* genetic variation displays significant regional structure, and consistently support the hypothesis that Iberia has been a glacial refugium for *A. thaliana*. In addition, the Iberian geographical structure indicates a complex regional population dynamics, suggesting that this region contained multiple Pleistocene refugia with different contribution to the postglacial colonization of Europe.
Introduction

The annual wild weed species *Arabidopsis thaliana* is a model organism not only for molecular biology but also for ecological and evolutionary genetics, and hence, revealing the geographical structure of its genetic variation has become of paramount relevance (MITCHELL-OLDS and SCHMITT 2006). Quantification of genetic diversity within and among populations of *A. thaliana* and analysis of its spatial distribution pattern across the species geographic range is the basis to elucidate demographical (historical) and ecological influences (MITCHELL-OLDS and SCHMITT 2006). Furthermore, identification of genetic structure is relevant to map the causal genes responsible for the natural variation of adaptive traits by genome-wide association analysis. In these assays, knowledge of the genetic structure will reduce spurious correlations between genotype and phenotype due to historical relationships affecting the genetic background (CARDON and PALMER 2003; ZHAO et al. 2007).

*Arabidopsis thaliana* shows a world-wide geographic distribution, although its native range mainly spans Europe and Central Asia, while it is mostly naturalized elsewhere (reviewed in HOFFMANN 2002). It remains unknown if the main *A. thaliana* centre of origin is Central Asia or Europe/North Africa, both areas showing the highest diversity of related species (HOFFMANN 2002). Currently, there are wild genotypes (accessions) collected from more than 500 populations across its world distribution, which have been used to estimate the amount and patterns of genetic variation at a world-wide scale. These analyses have found significant population structure at global scale, as well as long range isolation by distance among different world regions (SHARBEL et al. 2000; NORDBORG et al. 2005; SCHMID et al. 2006;
OSTROWSKI et al. 2006; BECK et al. 2008). In addition, several laboratories have recently initiated the development of new A. thaliana collections for genetic variation studies at regional scale, in regions of the native distribution area such as Northern Europe (STENØIEN et al. 2005; BAKKER et al. 2006), France (LE CORRE 2005), Central Asia (SCHMID et al. 2006) and China (HE et al., 2007), as well as in regions of presumed recent introduction and expansion like Japan (TODOKORO et al. 1996) and North America (JØRGENSEN and MAURICIO 2004; BAKKER et al. 2006). Thus far, a significant regional correlation between genetic and geographical distances has been only observed in China, which supports a progressive natural dispersal under strong antrophogenic influence. In contrast, lack of such correlations within non-native regions has been interpreted as consequence of a recent colonization and expansion from mixed origin, or consequence of multiple colonizations (TODOKORO et al. 1996; JØRGENSEN and MAURICIO 2004; STENØIEN et al. 2005).

The Iberian Peninsula, located in the Western border of Eurasia and very close to Africa, has received some attention in several A. thaliana genetic structure studies due to its special interest for the biodiversity history of Europe (SYMONDS and LLOYD, 2003; NORDBORG et al. 2005; SCHMID et al. 2006). This region is a major part of the largest biodiversity hotspot in Europe, the Mediterranean Basin (MYERS et al. 2000) and it has been one of the most important Pleistocene glacial refugia for numerous plant and animal species of the European subcontinent (reviewed in HEWITT 2001; GOMEZ and LUNT, 2006). However, analyses of Iberian A. thaliana diversity have been mostly limited to small sets of genotypes collected in a single subregion of Spain.
(ROBBLELEN 1965; KUITTINEN et al. 2002; BECK et al. 2008). These studies have suggested that Iberia is genetically differentiated from other world regions (SYMONDS and LLOYD 2003; NORDBORG et al. 2005; SCHMID et al. 2006) and that it might be part of a Mediterranean glacial refugium for A. thaliana (SHARBEL et al. 2000).

To better understand the evolutionary history of A. thaliana in its native geographic range, we have systematically studied A. thaliana genetic structure in the Iberian Peninsula at different spatial levels. To achieve this goal, we have generated a collection of 268 individuals from 100 populations covering this region, and this has been used to determine the amount and spatial distribution of genetic variation. We have inferred the genetic structure by analysing genome-wide genotypes obtained with microsatellites (MSs) and single nucleotide polymorphisms (SNPs). We show that Iberian genetic diversity is geographically structured, which indicates population isolation in the past and provides evidence supporting the role of the Iberian Peninsula as a Pleistocene refugium for postglacial colonization of Europe. In addition, inference of four distinct Iberian genetic clusters spatially separated indicates a complex regional population dynamics that suggest the occurrence of multiple glacial refugia for A. thaliana in this region.

Materials and Methods

Plant material and sampling design

A hundred natural populations of Arabidopsis thaliana were surveyed in a region of around 800 x 700 km of the Iberian Peninsula (IP) and Menorca Island (Figure 1). These were spaced at an average distance of 326±180 km, with a
minimum and maximum of 1 and 898 km respectively. Populations were assigned to six geographical subregions defined according to the six major Iberian mountain systems, and using the largest rivers as main subregional borders (supplemental Table S1). Sampled populations cover most of *A. thaliana* distribution area in the IP and were located in a wide range of habitats (supplemental Figure S1 and supplemental Table S1). Population size was roughly estimated in the field and ranged from few individuals (less than 25) covering ~1 m$^2$ patch, to at least 1000 individuals in a 100 m tract (supplemental Table S1).

Seven of these populations, distanced between 86 and 564 km, were chosen for analysis of local population differentiation. These were extensively sampled following a transect where seeds from 20 to 32 plants were individually collected at a minimum distance of 0.5 m. Populations were named with 3 letters indicating the closest village or locality, followed by a different number code for each sampled individual.

Forty-three additional individuals from different local populations covering most of the rest of *A. thaliana* world distribution were also analysed. Five of them are new European individuals collected by the authors, while the rest were accessions obtained from public collections available at stock centres (supplemental Table S2).

**DNA isolation and marker genotyping**

DNA was isolated from 175 individuals of the seven extensively sampled IP populations and from one randomly chosen individual of the remaining 93 IP populations. From each sampled mother plant, a mix of leaf tissue from at least
six sister plants grown from the sampled seeds was used for DNA isolation in the seven large populations, which represented the DNA of the field mother plants. For the rest of IP and world individuals, leaf tissue was harvested from a single plant grown from sampled or stock centre multiplied seeds. DNA was isolated using a previously described protocol (BERNARTZKY AND TANKSLEY, 1986) without mercaptoethanol.

Samples were genotyped at previously described microsatellite (BELL and ECKER 1994; PROVAN et al. 2000; LOUDET et al. 2002) and single nucleotide polymorphism loci (TÖRJÉCK et al. 2003; NORDBORG et al. 2005). Sixteen nuclear microsatellites (ncMSs) and four chloroplast microsatellites (cpMSs) were analysed (supplemental Table S3 and supplemental Figure S2). MS loci were amplified by PCR using a forward primer labelled with one of the Perkin-Elmer Applied Biosystems fluorochromes 6-FAM, NED, PET and VIC. PCR products of four differently labelled MSs were mixed in equal amounts and the fragments of five mixes (supplemental Table S3) were separated in an ABI PRISM 3700 DNA analyzer using GeneScan-500-LIZ (Applied Biosystems) as internal size standard. Electropherograms were visually inspected and manually scored using GeneScan 3.7 software (Applied Biosystems). Molecular sizes (bp) of DNA fragments were calculated and sizes were used to estimate the number of MS motif repeats, based on the available sequence of Columbia accession. MS alleles were recorded and analysed as the closest number of presumed motif repeats. MS loci showed an average frequency of missing data of 3.5%. However, nga111, nga172 and msat3.18 rendered 17.4, 16.4 and 8.0% of non-amplifying individuals, respectively, suggesting that these loci might contain null alleles or additional polymorphisms within the primer
sequences. The average frequency of missing data per individual was 1.7% for cpMSs and 3.2% for ncMSs. MS error rates of the molecular size estimates were calculated by amplifying and analysing duplicated samples of 20 genotypes, which provided an average genotyping error proportion of 0.046 per locus. All but one of the ncMSs are di- or trinucleotide repeat loci (supplemental Table S3) and showed an error rate of 0.019, while all cpMSs are mononucleotide repeats showing a 0.09 error rate. Most mistyped genotypes differed in one single base pair, whereas only a 0.007 error rate per locus was due to size estimates differing in more than one nucleotide.

Two different sets of nuclear SNP loci were analysed (supplemental Table S4 and supplemental Figure S2). Ninety-six Col/C24 SNP markers were selected because they are common polymorphisms in Central Europe (SCHMID et al. 2006). In addition, other 47 SNPs segregating in six Iberian accessions collected in several geographical subregions (Fei-0, Li-0, Pro-0, Se-0, Ts-1 and Ts-5) were randomly chosen from polymorphisms described by NORDBORG et al. (2005). SNPs were genotyped with SNPlex technique (Applied Biosystem), using three mixes of 47 or 48 loci (supplemental Figure 2S), through CEGEN genotyping service (http://www.cegen.org). Twenty-six Col/C24 SNP loci showing more than 25% missing data and eight other Col/C24 SNP loci that did not segregate among the individuals analysed in this work were discarded for further analyses. The final 62 Col/C24 SNP loci had a 6.0% average frequency of missing data. The 47 IP loci showed an average missing frequency of 4.6%. Ini-0 and Nac-0 individuals failed for the SNPlex mix containing the 47 IP SNP markers and were not included in some comparisons. SNPlex genotyping error rate was calculated by duplicated analysis of 25 individuals with the three
SNPlex mixes representing 117 SNP markers, which provided an average error rate per SNP locus smaller than 0.0004.

To integrate the genotypes generated in this work with previously published data, we first validated the SNP loci by genotyping several accessions included in other studies. Twenty-four accessions genotyped by SCHMID et al. (2006) and fourteen accessions analysed by NORDBORG et al. (2005) were genotyped for the Col/C24 SNPs and the IP SNP loci respectively. All except Sf-2 and LI-1 presented the same genotypes at all loci in both studies indicating that markers are the same but some accessions might be misclassified.

**Population data analysis**

Three sets of multilocus genotypes were generated in this work (supplemental Table S5): 1) genotypes of 100 IP individuals, one from each of the 100 populations; 2) genotypes of 175 IP individuals from the seven extensively sampled populations; 3) genotypes of 43 individuals from different populations from the rest of world. In addition, SNP genotypes of IP were compared with two world-wide SNP datasets previously generated (NORDBORG et al. 2005; SCHMID et al. 2006). To have comparable data from IP and world samples, a single random individual was selected from each local population and from each group of individuals with similar genotype, in previous world datasets. This selection did not affect greatly the analyses and comparisons presented in this work. Thus, we used two sets of 193 and 93 individuals from different world populations genotyped with 62 Col/C24 SNPs (SCHMID et al. 2006) and 47 IP SNPs (NORDBORG et al. 2005) respectively. Populations from outside the IP
were assigned to seven world geographical regions, according to BAKKER et al. (2006): Northern America, Western Europe, Eastern Europe (limit at longitude 60º W), Southern Europe and Northern Africa (here referred to as Mediterranean Basin), Northern Europe, Asia and Japan.

Genetic diversity was measured as percentage of polymorphic loci ($PL$), mean number of observed alleles per locus ($n_a$), mean allelic richness per locus ($R_S$), mean private allelic richness per locus ($R_P$), mean gene diversity ($H_S$) and number of multilocus haplotypes ($N_h$). These genetic parameters were estimated using the software programs FSTAT v. 2.9.3 (GOUDET 1995), POPGENE v. 1.32 (YEH et al. 1999) and HP-Rare v. 1.0 (KALINOWSKI 2005). Observed heterozygosities ($H_O$) and inbreeding coefficients ($F_{IS}$) were calculated for nuclear microsatellite loci using FSTAT. Outcrossing frequencies were estimated as $(1-F_{IS})/(1+F_{IS})$ according to ALLARD et al. (1968).

Linkage disequilibrium (LD) between pairs of loci was tested using the LD exact test implemented in FSTAT v. 2.9.3 (GOUDET 1995). The proportion of pairs of loci showing significant LD over the total number of possible pairs ($P_{ld}$) was estimated excluding polymorphic loci that only segregate as singleton alleles.

Genetic differentiation among populations or groups of individuals was estimated by hierarchical analysis of molecular variance (AMOVA, EXCOFFIER et al. 1992) using the program ARLEQUIN, v. 3.1 (EXCOFFIER et al. 2005). We calculated the $F_{ST}$ statistics analogous to the fixation index $F_{ST}$ (WEIR AND COCKERHAM 1984) and their significances from 1000 permutations. AMOVA tests were performed using multilocus genotypes on the following data sets: 1) seven local IP populations; 2) IP and world-wide individuals; 3) 100 IP
individuals classified into four genetic groups inferred with STRUCTURE; 4) 100 IP individuals grouped into six geographical subregions.

Genetic relationships among accessions were determined by neighbour-joining (NJ) analysis. Genetic distances between individuals were calculated as the proportion of allelic differences over the total number of alleles in the corresponding set of polymorphic loci, using the software program GGT v. 2.0 (VAN BERLOO 1999; http://www.plantbreeding.wur.nl/UK/software_ggt.html). NJ trees were constructed from 1000 bootstrap replicates using the software POPULATIONS v. 1.2.30 (http://bioinformatics.org/; O. Langella, unpublished) and drawn with MEGA v. 3.1 (KUMAR et al. 2004).

Phylogenetic relationships among chloroplast haplotypes (chlorotypes) were established and visualized as chlorotype frequency maps constructed with a median-joining network (BANDELT et al. 1999) using the program NETWORK v. 4.2 (http://www.fluxus-engineering.com). Given the large number of chlorotypes observed with four cpMS loci, the network presented was generated using only three markers (cp70189 was excluded; see supplemental Table S5). For clarity, chlorotypes detected in single individuals and unconnected to the network were removed from shown figures (11 out of 36 chlorotypes corresponding to 7.1% and 10.3% of the IP and world individuals, respectively).

Genetic structure was inferred using the model-based clustering algorithms implemented in STRUCTURE v. 2.1 (PRITCHARD et al. 2000; FALUSH et al. 2003) and TESS v. 1.1 (FRANÇOIS et al. 2006). These Bayesian approaches were applied to the following sets of genotypes: 1) 100 and 193 individuals from different populations of the IP and rest of world
respectively; 2) 100 IP individuals from different populations. For STRUCTURE analyses, we used a similar setting to that described by NORDBORG et al. (2005). Basically, SNP multilocus genotypes were analysed with a haploid setting, using the linkage model with correlated allele frequencies, and running the algorithm with 50,000 MCMC iterations of burn-in length and 20,000 after-burning repetitions for parameter estimations. Genetic positions of loci were directly obtained from the consensus *A. thaliana* genetic map (http://www.arabidospsis.org) or by interpolation from the physically closest loci with known genetic position. To estimate the K number of ancestral genetic populations and the ancestry membership proportions of each individual in these clusters, the algorithm was run 10 times for each K value from 2 to 15. Differences between the data likelihood of successive K values were tested using the non-parametric Wilcoxon test for two related samples. The final K was estimated as the largest K value with significantly higher likelihood than that from K-1 runs (two-sided *P*<0.005). Similarity between runs was estimated using the symmetric similarity coefficient (NORDBORG et al. 2005) and the extent of membership in a single cluster was measured using the clusteredness coefficient (ROSENBERG et al. 2005). These parameters and the average matrix of cluster membership proportions of the 10 runs were computed using a Structure-sum R-script (EHRICH et al. 2006).

In contrast to STRUCTURE, TESS algorithm incorporates spatial population models assuming geographical continuity of allele frequencies by including the interaction parameter $\psi$, which defines the intensity of two neighbour individuals belonging to the same genetic cluster. In addition, TESS treats K as a variable to be estimated. Haploid multilocus genotypes were
analysed with TESS using the MCMC method, with the F-model and a $\psi$ value of 0 (which assumes a non-informative spatial prior) as well as with the admixture model and $\psi$ values between 0.5 and 0.7. For each model, the algorithm was run 200 times, each run with a total of 70,000 sweeps and 50,000 burn-in sweeps. K was estimated from the 10-20% runs with highest data likelihood. Similarity coefficients between runs and the average matrix of ancestry membership were calculated using CLUMPP v. 1.1 (JAKOBSSON and ROSENBERG 2007).

Estimated average matrices of membership proportions were graphically represented using DISTRACT software (ROSENBERG et al. 2002). Geographical distribution of ancestry matrices were represented by Kriging methods using the R-script available at http://www-timc.imag.fr/Olivier.Francois/admix_display.html (Olivier François, unpublished).

The relationship between genetic distance and Euclidean geographical distance among population pairs of the Iberian Peninsula was determined by Mantel correlation test (MANTEL 1967; SMOUSE et al. 1986) using ARLEQUIN (EXCOFFIER et al. 2005) and the isolation by distance (IBD) web service v. 3.13 (IBDWS, JENSEN et al. 2005). Genetic and geographical distances were log-transformed prior to analysis and the significance of correlations was calculated with 1000 randomizations. We estimated genetic distance between population pairs in two ways: when using local populations or groups of individuals we calculated the Slatkin’s linearized $F_{ST}$ expressed as $D = F_{ST}/(1-F_{ST})$ (SLATKIN 1995) with ARLEQUIN; when using populations represented by one randomly chosen individual we computed the proportion of pair-wise allelic differences as described above. Since Mah is an island population located more
than 300 km apart from the IP (Fig. 1), this genotype was dropped from these analyses. Isolation by distance analyses were carried out on the following sets of genotypes: 1) seven Iberian local populations 2) 99 IP individuals from different populations; 3) 99 IP individuals grouped into six geographical subregions, using the average geographical distance among all pairs of individuals from different subregions; 4) fifteen subgroups of the 99 IP individuals, corresponding to all different pair combinations of the six geographical subregions.

Results

General genetic diversity in the Iberian Peninsula

To estimate A. thaliana genetic diversity in the Iberian Peninsula (IP), we sampled 100 local populations in a region of 800 x 700 km (Figure 1 and supplemental Table S1). A randomly chosen individual from each population was genotyped with four sets of markers corresponding to 16 ncMSs, 4 cpMSs, 62 Col/C24 SNP and 47 IP SNP loci (Table 1 and Materials and Methods). As expected, genetic diversity estimates at MS loci were consistently larger than those at SNP loci. In addition, the two sets of microsatellites differed significantly, cpMSs showing considerably lower diversity values than ncMSs (Table 1). Twelve of the Col/C24 SNP loci (19.4%) were monomorphic while only one of the IP SNP markers (2.1%) was not polymorphic in the 100 individuals. IP SNP loci showed a slightly higher average minor allele frequency (MAF) (0.19±0.16) than Col/C24 SNPs (0.14±0.15) (supplemental Figure S3) and larger diversities were estimated with IP SNPs than with Col/C24 loci (Table 1). All 100 individuals showed different multilocus genotypes based on
any of the two sets of SNP loci or on the ncMSs, while cpMSs distinguished a total of 34 different chlorotypes. On average, we estimated that pairs of *A. thaliana* individuals collected in different natural populations differed in 55%, 81%, 26% and 27% of the polymorphic cpMSs, ncMSs, Col/C24 and IP SNP loci, respectively.

**Genetic diversity and differentiation of local populations**

To determine the genetic variation within Iberian local populations and their differentiation, we genotyped 175 individuals collected from seven large populations (20 to 32 per population) with the same four marker sets (Figure 1, Table 2). All microsatellites segregated among populations but only 46 loci from the 62 Col/C24 SNPs and 42 from the 47 IP SNP loci were polymorphic. A two to four-fold difference in genetic diversity was found among populations, depending on the parameter (supplemental Table S6). However, gene diversity estimates (\(H_s\)) of ncMSs and of the two sets of SNP loci were highly correlated (\(N = 7; r > 0.81; P < 0.026\)) indicating that both types of nuclear loci detected the same patterns of genetic variation. On average, pairs of *A. thaliana* individuals collected in the same local population differed in 38%, 58%, 16% and 13% of the polymorphic cpMSs, ncMSs, Col/C24 and IP SNP loci, respectively.

In total, 32 to 130 different genotypes were found depending on the marker set, the combined analysis of all nuclear loci resulting in the same 130 genotypes detected with ncMSs (supplemental Table S6). No identical multilocus genotype was found in different populations with any set of nuclear markers, while four chlorotypes were detected in several populations. The
combined analysis of both sets of SNP loci identified 88 different multilocus genotypes. From the joint analysis of all nuclear loci, we estimated that within the IP populations, on average, 50% of the individuals show haplotypes differing in at least 1 SNP locus and several ncMSs, 26% are genetically identical to other individuals, and 24% have nearly identical genotypes differing in 1 to 3 ncMS loci. In total, we found 28 pairs of individuals from the same population differing in a single ncMS locus. From this, we estimated a frequency of ~1% nearly identical individuals in an otherwise similar genetic background, per ncMS locus. This proportion was similar to the estimated ncMS error rate (see Materials and Methods). Therefore, we concluded that most nearly identical individuals detected with ncMSs in the same local population are not carrying de novo ncMS mutations, but probably, they are identical genotypes bearing MS genotyping errors.

Genome-wide analysis of linkage disequilibrium (LD) between pairs of nuclear loci in each population indicated that 5.6% to 62.3% of all pair-wise combinations of loci present significant LD (Table 2). On the other hand, local populations showed a mean observed heterozygosity per ncMS locus ($H_0$) of 0.02 and a mean outcrossing frequency of 2.5% (Table 2). However, substantial variation was found among populations since outcrossing frequencies ranged between 0.3% and 7.5 %. In total, 12, 7, 4 and 4 individuals appeared heterozygous for 1, 2, 3 and 5 ncMS loci respectively. Together, these individuals carried 58 heterozygous ncMS data points, 46 involving alleles already segregating in the corresponding population. Therefore, most heterozygous MS loci are probably generated by outcrossing.
Neighbour-joining analysis of the 175 multilocus genotypes showed that most individuals from the same population, but not all, group together (supplemental Figure S4). In addition, AMOVA estimates of genetic differentiation among the seven populations indicated that 33.6% of the genetic variation from all loci is present among populations. Similar average $F_{ST}$ values were estimated from the various sets of markers ($F_{ST} = 0.31-0.36; P < 0.0001$). Moreover, the seven populations differed genetically from each other since all $F_{ST}$ values between pairs of populations were significant ($F_{ST} = 0.12-0.66; P < 0.0001$). Therefore, Iberian populations are genetically differentiated, although more genetic variation is found within than among populations.

**Comparison between the Iberian Peninsula and the rest of world**

Genetic diversities of the IP and the rest of world distribution were compared by joint analyses of the 100 IP genotypes and 43 to 193 world-wide genotypes, depending on the marker type (Table 1 and Materials and Methods). Both groups of samples showed rather similar $H_s$ values for all sets of MS and SNP loci. However, Col/C24 SNP loci presented lower percentage of polymorphic loci and mean allelic richness in the IP than in the rest of world, while the opposite behaviour was observed for the IP SNP markers (Table 1). These results indicate that IP and the rest of world differ in their SNP allele frequencies and therefore, diversity estimates are differentially biased for the two sets of SNP loci. Since the SNP markers were selected from genotypes of Iberia and Central Europe (Materials and Methods) IP and the rest of world are biased towards higher diversity for the IP and the Col/C24 SNP loci, respectively.
Diversity estimates described for the two sets of SNP loci throughout the text reflect such marker ascertainment bias.

AMOVA analyses of genetic differentiation between the IP and the rest of world indicated that less than 5% of the variation distinguishes both groups ($F_{ST} = 0.019-0.041; P < 0.05$). This limited differentiation was mostly due to 1, 3, 12 and 11 of the cpMSs, ncMSs, Col/C24 and IP SNP loci, respectively, showing significantly different allele frequencies in the two samples ($P<0.01$).

On the other hand, the relationship among chlorotypes of the IP and the rest of world was compared by frequency network analysis of 143 individuals (Figure 2 and supplemental Table S5). In total, 36 chlorotypes were detected with three cpMSs, which were arranged in a complex phylogenetic network (see Materials and Methods). The six most frequent chlorotypes, showing frequencies higher than 4%, were found in both samples, while specific IP or world chlorotypes presented lower frequencies than 4%. However, IP and world samples differed in the frequency distribution of chlorotypes ($\chi^2 = 60.56; P<0.001$) and in the most common chlorotype (CH01 and CH18 for the IP and the rest of world respectively). In addition, eight IP specific chlorotypes appeared derived by single mutational steps from the commonest IP chlorotype, suggesting that this is the oldest Iberian chlorotype.

Finally, the overall structure of the IP genetic variation was inferred and compared to that of the rest of world distribution by using the clustering algorithm implemented in STRUCTURE. Four different genetic clusters were detected from the analysis of 293 multilocus genotypes obtained with the 62 polymorphic Col/C24 SNP loci (Figure 3 and supplemental Table S7). On average, individuals from outside the IP showed the largest membership
fractions in clusters 1 and 2, while clusters 3 and 4 presented the largest proportions in the Iberian genotypes. When individuals from outside the IP were classified in seven geographical regions, two spatio-longitudinal gradients of ancestry membership frequencies, with opposite directions, were detected in Eurasia. Cluster 2 showed the highest mean fraction in samples from Asia, its frequency decreasing in Eastern and Western Europe. In contrast, cluster 3 showed the highest frequency in the IP, and this decreases in Eastern Europe and Asia (Figure 3). In addition, cluster 1 appeared as the most frequent in Western Europe but the least common in IP, whereas cluster 4 shows a high frequency in the IP and the Mediterranean Basin but the lowest frequency in Western Europe. Individuals from North America and Asia showed at least two genetic clusters with average membership fractions lower than 10% indicating that samples from these regions contain less genetic variation than samples from Japan.

**Analyses of genetic structure in the Iberian Peninsula**

To establish the genetic structure of *A. thaliana* Iberian populations more precisely, we first carried out NJ analyses of the 100 individuals (supplemental Figure S5). These analyses detected several groups of five to ten genotypes, but bootstrap support was mostly low. Therefore, a model-based clustering approach was used to determine the structure of these populations (Figure 4 and supplemental Table S7). Four genetic clusters were inferred with STRUCTURE when analysing the multilocus genotypes obtained with all 95 polymorphic SNP loci. However, only three significant clusters were found when using 50 Col/C24 or 45 IP SNP loci (supplemental Table S7). Cluster
membership coefficients distinguished different but overlapping groups of individuals in the three analyses, illustrating the synergistic and complementary behaviour of both sets of SNP loci (data not shown). Basically, cluster C4 differentiated with the Col/C24 SNP loci in the world-wide analysis was split in two clusters when including the IP SNPs (for comparison, red colour IP individuals with high C4 membership fraction in the world-wide analysis of Figure 3 closely correspond to the red colour individuals in the K=2 analysis of Figure 4), while the remaining clusters C1 to C3 were assigned to two other clusters. Most individuals showed an estimated major membership proportion larger than 0.6 and therefore, they could be classified in four distinct genetic groups according to their largest ancestry membership fractions (Figure 4). This classification was in agreement with NJ analyses since individuals assigned to the same genetic group trend to cluster together in NJ trees (supplemental Figure S5). Nevertheless, it has been argued that in populations with continuous spatial distribution of genetic diversity, clusters detected by STRUCTURE might be influenced by uneven geographical distribution of samples (ROSENBERG et al. 2005; FRANÇOIS et al. 2006). Hence, clustering analyses were also performed using the algorithm implemented in TESS (see Materials and Methods). Four genetic clusters were also inferred in TESS analyses using different assumptions on the spatial distribution of genetic clusters (supplemental Figure S6). Most individuals showed the same major ancestry membership coefficient than that estimated with STRUCTURE, supporting the robustness of the inferred clusters. However, similarity coefficients among runs of the same TESS model were considerably lower than those estimated among STRUCTURE runs (supplemental Table S7).
Genetic diversities of the four STRUCTURE groups show a two-fold variation among groups, with groups 1 and 2 being consistently more diverse than groups 3 and 4 (supplemental Table S8). AMOVA analyses using all loci showed an average $F_{ST}$ differentiation of 0.11 over the four groups, $F_{ST}$ values being significant for all sets of markers (supplemental Table S9). The lowest differentiation was observed between groups 1 and 2, while the largest differentiations were estimated between group 3 and the remaining groups (supplemental Table S9).

We further analysed the four genetic groups for their chlorotype diversity (Figure 2B). Each genetic group carried numerous chlorotypes and no clearly distinct maternal origin of any group was observed. However, groups 1 and 2 bore at least four chlorotypes also present outside Iberia and located throughout the network. In contrast, groups 3 and 4 showed more restricted geographical and evolutionary chlorotype variation. Most individuals of group 3 carried different IP-specific chlorotypes, whereas group 4 contained the smallest number of IP-specific chlorotypes. These results suggest that group 3 is an Iberian specific group that has remained rather isolated from other world regions, whereas seeds of groups 1, 2 and 4 might have migrated between IP and other world regions.

**Analyses of geographical structure in the Iberian Peninsula**

To determine if *A. thaliana* genetic variation is spatially structured in Iberia, we first tested isolation by distance among the 100 genotypes collected from different local populations. Mantel tests showed that genetic distances are positively correlated with geographical distances, $r$ values ranging between 0.1
and 0.23 depending on the marker set ($P<0.004$). This correlation was maximum when using the genetic distances estimated from all loci ($r=0.28; P<0.001$). The IBD pattern of geographical structure was also analysed in the seven local populations extensively sampled (Figure 5A), Mantel test showing significant correlations for SNP loci ($r=0.55-0.60; P < 0.04$) and marginal significances for MS markers ($r=0.40-0.53; P = 0.06$).

*A. thaliana* spatial structure in the IP was further evaluated by classifying the 100 populations in six geographical subregions (Figure 1 and supplemental Table S1). Genetic diversities were rather similar in the six IP subregions (supplemental Table S10), but AMOVA analyses indicate significant differentiation among all subregions for three sets of loci (supplemental Table S9). As shown in Figure 5B, a strong significant correlation was found between geographical and genetic distances among the six subregions ($r=0.64, P=0.014$). This significant regional isolation by distance was in agreement with the clustering obtained by NJ analysis, where individuals from the same subregion trend to cluster together (supplemental Figure S5). To identify geographical barriers that might limit genetic flow, we also performed Mantel tests with the individual genotypes from each of the 15 pair combinations of the six geographical subregions. All pair comparisons of subregions showed significant positive correlation between geographical and genetic distances. The lowest values corresponded to comparisons involving pairs of subregions I to IV ($r = 0.21-0.26; P<0.005$). In contrast, comparisons of subregions V and VI with the rest of subregions showed considerably larger correlations ($r = 0.30-0.62; P<0.001$), indicating that the north-east subregions have been more isolated from the rest of Iberia.
Finally, geographical structure of A. thaliana genetic variation was also inferred from the four genetic clusters previously established by model-based approaches. As shown in Figure 5C, genetic clusters derived with STRUCTURE were not evenly distributed across Iberia, but they appear mostly restricted to particular subregions. Overall, clusters 1, 2, 3 and 4 are located mainly in the north-west, north-east, centre/south-east and south-west areas, respectively. Genetic clusters inferred from a TESS model assuming non-informative spatial prior showed nearly the same geographical distribution than STRUCTURE clusters (supplemental Figure S6A and S6B). Results from both algorithms differed mainly in the frequencies of clusters 1 and 2, TESS analysis leading to higher and lower frequencies of these groups respectively. In addition, genetic clusters inferred with TESS using an interaction parameter value of 0.7 also show considerable spatial overlapping with STRUCTURE clusters (supplemental Figure S6C). However, the latter TESS model estimated a more restricted geographical distribution of clusters 2 and 3, and a broader distribution of cluster 4.

Discussion

Arabidopsis thaliana is widely distributed as a native species in the Iberian Peninsula. In this region, A. thaliana appears not only in agricultural fields and other relatively anthropogenic habitats but also in a wide range of naturally-disturbed habitats, from mesic and xeric grasslands to Atlantic and Mediterranean forests (supplemental Figure S1). In this work, we have developed an Iberian collection of 268 individuals sampled from 100
populations, which correspond to 181 distinct genotypes as estimated from presumed neutral cpMSs, ncMSs and nuclear SNP loci.

**Estimating genetic diversity, differentiation and structure with microsatellites and single nucleotide polymorphisms**

As expected from the different molecular nature and mutation rate of MS and SNP loci, genetic diversities were higher when based on microsatellite than on SNP loci. Consistently, lower genetic differentiation values were often found with MS than SNP loci. In addition, ncMS loci did not enable inference of genetic clusters when analysed with Bayesian model-based clustering algorithms. This was not exclusively due to the low number of ncMS loci considered because the combination of ncMS and SNP markers showed reduced clustering power than SNP loci (data not shown). Probably, this is also consequence of the high variability of the ncMS analysed in *A. thaliana*, which is predicted to generate a large proportion of molecular convergence (homoplasy) in small populations evolving with high mutation rates (ESTOUP *et al.* 2002). Thus, SNP loci were more useful than microsatellites for *A. thaliana* genetic structure analyses. However, population genetic parameters estimated from SNP loci were also biased because SNP markers were ascertained from small sets of genotypes with uneven geographical distribution. Selection of SNPs from small panels of genotypes biases the sets of SNP loci towards polymorphisms with intermediate allele frequencies (reviewed in BRUMFIELD *et al.* 2003). Accordingly, we mainly analysed common sequence polymorphism since only 27-44% of the SNP loci show MAF<5% (supplemental Figure S3), compared with 55% of synonymous SNPs presenting MAF<5% in random
samples (NORDBORG et al. 2005). Such distortion increases overall diversity estimates and decreases differentiation of recent branches of genealogical trees (BRUMFIELD et al. 2003). However, due to the detected geographical structure, selection of SNP loci from genotypes of particular regions, like Iberia or Central Europe, mainly biases population parameter estimates in those regions. The joint analysis of two sets of SNP loci selected from different world regions reduced SNP ascertainment bias and increased the capability to infer genetic clusters.

It must be emphasized that as discussed for human populations (ROSENBERG et al. 2005), A. thaliana genetic clusters detected by STRUCTURE algorithm are probably due to small IP geographical discontinuities of allele frequencies, and such clusters only represent a small fraction of the total genetic variation (average $F_{ST}$ differentiation among IP clusters is 0.11). It has been shown that the clustering capacity of STRUCTURE is affected by several factors, like the number of loci (ROSENBERG et al. 2005). The two sets of SNP loci used in this work were selected to contain nearly 50 loci, since this is the minimum number estimated for consistent inference of genetic clusters (ROSENBERG et al. 2005). Furthermore, it has been argued that uneven spatial sampling in the experimental design might also affect the STRUCTURE clustering patterns (ROSENBERG et al. 2005). As proposed by FRANÇOIS et al. (2006), the robustness of STRUCTURE clusters detected in this work has been tested with TESS algorithm, which incorporates spatial models for geographical continuity of allele frequencies. A large concordance was observed among the results obtained with both algorithms.
Therefore, uneven geographical distribution of samples and the specific bias of MS and SNP loci are not expected to affect any major conclusion of this work.

**Genetic diversity within local populations of *A. thaliana***

The seven Iberian local populations analysed in this work contain substantial amount of genetic variation, since all of them include several chlorotypes and around 50% of the individuals of each population show distinct nuclear multilocus genotypes. Average $F_{ST}$ values indicate that 66% of the genetic variation is segregating within these populations. Different proportions of within population diversity have been found in other world regions, values ranging from 77-36% in North America (BERGELSON *et al.* 1998; JØRGENSEN and MAURICIO 2004), 57% in Europe (BAKKER *et al.* 2006), 54% in China (HE *et al.* 2007), 41% in France (LE CORRE 2005), 12% in Norway (STENØIEN *et al.* 2005) and 0% in Japan (TODOKORO *et al.* 1997). In agreement with BECK *et al.* (2008), the large fraction of genetic variation segregating within the Iberian populations at MS and SNP loci might suggest a south-north latitudinal gradient of local population diversity in Europe. However, caution must be taken when comparing genetic differentiations from previous studies because these have been estimated with different types of markers. Moreover, the genetic diversity of local populations might depend on population size, age and habitat. The Iberian populations analysed in this work are large populations of thousands of individuals, permanent during several years, and growing in mostly natural habitats. Such populations might contain larger amount of genetic variation than smaller and recent populations collected in more homogeneous habitats. Accordingly, large coordinated studies avoiding differences due to type and
number of molecular markers, kind of local population and sample size, are necessary to compare the distribution of genetic diversity within and among local populations in different world regions.

Iberian populations differ considerably in their genetic diversity. In agreement with previous studies (STENØIEN et al. 2005; BAKKER et al. 2006) several results indicate that migration, outcrossing, and de novo mutation, differentially contribute to the variation within populations. First, chloroplast and nuclear haplotype analyses show that most populations contain not only related but also genetically unrelated individuals (supplemental Figure S4) suggesting that seed migration is an important factor contributing to within population diversity. Second, considerable variation is found among populations for the proportion of pairs of loci showing significant LD, which indicates a different contribution of cross-fertilization and recombination, or of demographical factors affecting LD. In agreement, large variation among populations was also found for outcrossing rate estimates. Iberian populations presented an average outcrossing frequency of 2.5%, which is slightly larger than previous MS-based estimates (LE CORRE et al., 2005; STENØIEN et al. 2005; BAKKER et al. 2006). This is probably due to the larger diversity of Iberian populations and not to mistyping errors because the three main sources of MS scoring errors (stutter bands, large-allele dropout and null alleles) lead to underestimations of heterozygosity (reviewed in DEWOUDY et al., 2006). Moreover, in contrast to previous MS studies, we estimated heterozygosity of field plants instead of individuals raised from field seed, which suggests that heterozygous individuals derived from outcrossing might have higher fitness under natural conditions. Finally, the presence of some individuals heterozygous for ncMS alleles that
otherwise are undetected in the same local population suggests that part of this heterozygosity might be generated by *de novo* ncMS mutations. However, true detection of individuals carrying new or homoplastic MS alleles generated by *de novo* mutation requires exhaustive genome-wide genotyping with extremely low error rate, which cannot be achieved by the standard methods of MS analysis used in this work.

Inference of *A. thaliana* diversity centres and genetic refugia from the world-wide geographical structure

As discussed above, the broad diversity of habitats occupied by *A. thaliana* in the Iberian Peninsula, together with the large genetic variation observed within Iberian populations suggest a longer demographical history of *A. thaliana* in Iberia than in other world regions. Furthermore, several results suggest that the Iberian Peninsula was a diversity centre and a European glacial refugium for *A. thaliana*. First, Iberia contains similar amount of neutral genetic diversity per locus than the rest of world distribution area. Second, Iberia is genetically differentiated from the rest of world as indicated by allele frequencies at single SNP and MS loci, as well as by multilocus chlorotype frequencies. Third, the global-scale geographical structure of genetic variation inferred by model-based clustering analysis shows a pattern of genetic differentiation of Iberia and the rest of Eurasia that further supports this hypothesis (Figure 3). Two main conclusions are drawn from the joint analysis of genotypes from IP and the rest of world. On one hand, the Iberian Peninsula appears populated by several distinct genetic lineages and not by a single homogenous genetic group of populations, as previously described from the analysis of a limited number of
accessions (SYMONDS and LLOYD, 2003; SCHMID et al. 2006). On the other hand, two longitudinal gradients of cluster frequencies are detected in Eurasia suggesting a double post-glacial colonization of Central Europe, in agreement with spatial gradients and IBD patterns previously found (SHARBEL et al. 2000; OSTROWSKY et al. 2006; SCHMID et al. 2006). The presence of a west-east frequency decrease of the major IP cluster (C3) and an east-west frequency decrease of the main Asian genetic cluster (C2), point to a postglacial colonization of Central Europe from Iberia and Asia. However, given the high Western Europe frequency of another cluster inferred in this analysis (C1 in Figure 3), it is very likely that other Eurasian refugium(a) contributed to its colonization. Furthermore, the low genetic diversity found in Asian accessions with SNP markers (present work and SCHMID, et al., 2006) does not support that this region was the primary centre of diversity and origin of A. thaliana. In agreement, it has been recently suggested that A. thaliana arose in the Caucasus region (BECK et al. 2008). However, the limited number and distribution of genotypes studied from Asia and Southern Europe, together with their restricted genotyping, precludes solid conclusions on the origin and expansion of A. thaliana in Eurasia.

**Causes and consequences of the geographical structure of A. thaliana in the Iberian Peninsula**

Several results demonstrate that A. thaliana neutral genetic variation is spatially structured in the Iberian Peninsula and indicate a limited genetic flow among geographical subregions. First, significant isolation by distance was detected in the entire IP and among IP subregions, based on pair-wise comparisons of the
Global $F_{ST}$ values indicate an average differentiation among subregions of ~8%, which is comparable to the subregional differentiation estimated from the local populations extensively analysed (0.7-16.6%, data not shown). Thus, local (short distance) dispersal of seeds seems an important process contributing to found new populations in Iberia. Second, the geographical distribution of the genetic clusters inferred by model-based approaches shows a clear cut spatial differentiation pattern. Consistently, the four genetic groups inferred with different population models appear mainly distributed in the four Iberian quadrants (Figure 5C and supplemental Figure S6).

The current geographical structure in four largely parapatric groups indicates isolation of genetic lineages and (meta)populations in the past. One possible scenario to generate such isolation can be speculated from the last Pleistocene glaciations, 20,000-40,000 years ago (PÉREZ ALBERTI et al. 2004). As reviewed by GOMEZ and LUNT (2006), the fragmented nature of suitable Iberian habitats favoured the occurrence of multiple glacial refugia isolated from each other. Phylogeographic studies of many European flora and fauna species have shown strong genetic subdivision in the Iberian Peninsula, the spatially separated distribution of genetic lineages being interpreted as remnant landmarks of those Pleistocene refugia. Comparative analyses demonstrate that phylogeographic Iberian patterns of different species broadly overlap among them, as well as with areas of high endemism. These studies indicate the internal complexity of Iberia as a glacial refugium and have led to the proposal of several Iberian glacial refugia that are shared by multiple species (GOMEZ and LUNT 2006). Interestingly, the geographical distribution
of *A. thaliana* genetic clusters inferred in this work partly overlaps with some of the refugia described for other species. Therefore, we hypothesize that Iberia was not a single *A. thaliana* refugium during Pleistocene glaciations but it provided several southern and northern refugia.

On the other hand, the geographical structure of *A. thaliana* suggests that there have been physical and/or environmental barriers limiting genetic flow throughout Iberia during the postglacial period. Mantel correlation analyses with pairs of geographical subregions suggest that the Ebro river basin (limiting regions V and VI) is a major Iberian geographical barrier, subregion VI showing a 0.16 average $F_{ST}$ differentiation from the rest of Iberia. In addition, the Central System mountains (stretching east-west across the IP centre and separating subregions II and III) might have also limited genetic flow, since the lowest genetic differentiations were estimated between subregions I and II, and subregions III and IV. Furthermore, current spatial structure might have been partly originated by other isolating factors or unknown demographical processes acting more recently in Holocene. In addition, given the latitudinal climatic diversity of Iberia, it cannot be discarded that this structure is also sustained by natural selection of genetic lineages adapted to different subregional climatic environments during glacial and/or postglacial isolation.

Finally, the Iberian geographical structure provides new insights on the contribution of the IP refugia to current European *A. thaliana* diversity. We hypothesize two different postglacial colonization waves deriving from Iberia and involving the IP genetic groups 1/2 and 3/4. Genetic groups 1 and 2 cover the northern half of the IP (Figure 5C) and are mainly related to cluster 3 of the world-wide analysis, which shows a west-east spatial gradient in Eurasia.
(Figure 3). Hence, groups 1 and 2 from Northern IP seem to have contributed mainly to the post-glacial colonization of Western and Northern Europe. On the other hand, cluster 4 was identified in the world-wide analysis at relatively high frequency only in the Mediterranean Basin (Figure 3), while it shows lower frequency in the rest of the Eurasian continent. This cluster is differentiated in the IP genetic clusters 3 and 4 (Figure 4), which are found mainly in Southern Iberia. Interestingly, most individuals of IP group 3 carry Iberian specific chlorotypes suggesting that this cluster is specific of Iberia. Furthermore, IP group 3 shows the largest $F_{ST}$ differentiations among the four IP groups, pointing to this group as the most ancient Iberian genetic lineage. These results, together with the presence of chlorotypes shared with other world regions in IP cluster 4 (Figure 2B), suggest that IP group 4 might have been differentiated from IP specific cluster 3 and expanded later to the Mediterranean region. However, the opposite direction of differentiation and migration cannot be discarded, since it is unknown how the expansion of $A. thaliana$ in the Mediterranean Basin occurred before and after the last glaciation. Further studies of Southern Europe peninsulas and the North Africa region will elucidate the demographical history of $A. thaliana$ in the Mediterranean Basin, where other glacial refugia have been located for many European species (HEWITT, 2001).

Acknowledgements
Authors wish to thank Dorothée Ehrich for assistance with Structure-sum R-script and François Olivier for help with representation of ancestry matrixes by Kriging methods. We also thank to Manuel Piñeiro (Ees, Glo, Mdd and Nac), César Gómez-Campo (Sne), Pere Fraga (Mah), Eduardo Sánchez (Mol and
Smt), José A. Aguilar (Fue), Herminia Zaldívar (Mep), Pablo Catarecha (Tor) and Ivan del Olmo (Cha) for kindly providing the samples indicated between brackets. The authors thank Mercedes Ramiro for excellent technical assistance. This work was funded by grant BIO2004-00533 from the Ministerio de Educación y Ciencia to C.A.-B. F.X.P. was funded by grant GENOMICS/2003/12 from the European Science Foundation travel program. B.M.-V. was funded by a salary fellowship from the FICYT (Principado de Asturias, Spain).

Text citations

ALLARD, R., S. JAIN and P. WORKMAN, 1968 The genetics of inbreeding species. Advances in Genetics 14: 55-131.

BAKKER, E. G., E. A. STAHL, C. TOOMAJIAN, M. NORDBORG, M. KREITMAN et al., 2006 Distribution of genetic variation within and among local populations of Arabidopsis thaliana over its species range. Mol Ecol 15: 1405-1418.

BANDELT, H.J., FOSTER, P. and A. RÖHL, 1999 Median-joining network for inferring intraspecific phylogenies. Mol. Biol. Evol. 16.

BECK, J. B., H. SCHMUTHS and B. A. SCHAAAL, 2008 Native range genetic variation in Arabidopsis thaliana is strongly geographically structured and reflects Pleistocene glacial dynamics. Mol Ecol 17: 902-915.

BELL, C. J., and J. R. ECKER, 1994 Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics 19: 137-144.
BERGELSON, J., E. STAHL, S. DUDEK and M. KREITMAN, 1998 Genetic variation within and among populations of Arabidopsis thaliana. Genetics 148: 1311-1323.

BERNARTZKY, R., and S. TANKSLEY, 1986 Genetics of acting-related sequences in tomato. Theor. Appl. Genetics 72: 314-324.

BONIN, A., D. EHRICH and S. MANEL, 2007 Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Mol Ecol 16: 3737-3758.

BRUMFIELD, R. T., BEERLI, P., NICKERSON, D.A. and S.V. EDWARDS, 2003 The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol. Evol. 18: 249-256.

CARDON, L. R., and L. J. PALMER, 2003 Population stratification and spurious allelic association. Lancet 361: 598-604.

DEWOODY, J., J. NASON and V. HIPKINS, 2006 Mitigating scoring errors in microsatellite data from wild populations. Mol. Ecol. Notes 6: 951-957.

EHRICH, D., 2006 AFLPDAT: a collection of r functions for convenient handling of AFLP data. Mol. Ecol. Notes 6: 603-604.

ESTOUP, A., P. JARNE and J. M. CORNUET, 2002 Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol 11: 1591-1604.

EXCOFFIER, L., G. LAVAL and S. SCHENEIDER, 2005 Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47-50.

EXCOFFIER, L., P. SMOUSE and J. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes:
application to human mitochondrial DNA restriction data Genetics 131: 479–491.

FALUSH, D., M. STEPHENS and J. K. PRITCHARD, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

FRANCOIS, O., S. ANCELET and G. GUILLOT, 2006 Bayesian clustering using hidden Markov random fields in spatial population genetics. Genetics 174: 805-816.

GOMEZ, A., and D. LUNT, 2006 Refugia within refugia: patterns of phylogeografic concordance in the Iberian Peninsula, pp. 155-188 in Phylogeography of Southern European Refugia, edited by S. WEISS and N. FERRAND, The Netherlands.

GOUDET, J., 1995 FSTAT: A computer program to calculate F- statistics. Journal of Heredity: 485-486.

HE, F., D. KANG, Y. REN, L. J. QU, Y. ZHEN et al., 2007 Genetic diversity of the natural populations of Arabidopsis thaliana in China. Heredity 99: 423-431.

HEWITT, G.M., 2001 Speciation, hybrid zones and phylogeography – or seeing genes in space and time, Mol. Ecol. 10: 537-549.

HOFFMANN, M. H., 2002 Biogeography of Arabidopsis thaliana (L.) Heynh. (Brassicaceae). Journal of Biogeography 29: 125-134.

JAKOBSSON, M., and N. A. ROSENBERG, 2007 CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801-1806.
JENSEN, J. L., A. J. BOHONAK and S. T. KELLEY, 2005 Isolation by distance, web service. BMC Genet 6: 13.

JØRGENSEN, S., and R. MAURICIO, 2004 Neutral genetic variation among wild North American populations of the weedy plant Arabidopsis thaliana is not geographically structured. Mol Ecol 13: 3403-3413.

KALINOWSKI, S., 2005 HP-Rare: a computer program for performing rarefaction on measures of allelic diversity. Mol. Ecol. Notes 5: 187-189.

KUITTINEN, H., SALGUERO, D. and M. AGUADÉ (2002) Parallel patterns of sequence variation within and between populations at three loci of Arabidopsis thaliana. Mol. Biol. Evol. 19: 2030-2034.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5: 150-163.

LE CORRE, V., 2005 Variation at two flowering time genes within and among populations of Arabidopsis thaliana: comparison with markers and traits. Mol Ecol 14: 4181-4192.

LOUDET, O., S. CHAILLOU, C. CAMILLERI, D. BOUCHEZ and F. DANIEL-VEDELE, 2002 Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. Theor Appl Genet 104: 1173-1184.

MANTEL, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer Research 27.

MITCHELL-OLDS, T., and J. SCHMITT, 2006 Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. Nature 441: 947-952.
MYERS, N., R.A. MITTERMEIER, C.G. MITTERMEIER, G.A.B. DA FONSECA and J. KENT, 2000 Biodiversity hotspots for conservation priorities. Nature 403: 853-858

NORBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN et al., 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3: e196.

OSTROWSKI, M. F., J. DAVID, S. SANTONI, H. MCKHANN, X. REBOUD et al., 2006 Evidence for a large-scale population structure among accessions of Arabidopsis thaliana: possible causes and consequences for the distribution of linkage disequilibrium. Mol Ecol 15: 1507-1517.

PÉREZ ALBERTI, A., M. VALCÁRCEL DÍAZ and R. B. CHAO, 2004 Pleistocene glaciation in Spain in Quaternary glaciations: extent and chronology, (Part I : Europe) edited by J. EHELERS and P. L. GIBBARD. Elsevier, The Netherlands.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

PROVAN, J., 2000 Novel chloroplast microsatellites reveal cytoplasmic variation in Arabidopsis thaliana. Mol Ecol 9: 2183-2185.

ROBBLELEN, G., 1965 The LAIBACH standard collection of natural races. Arabidopsis Information Service 2.

ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD et al., 2002 Genetic structure of human populations. Science 298: 2381-2385.
ROSENBERG, N. A., S. MAHAJAN, S. RAMACHANDRAN, C. ZHAO, J. K.
PRITCHARD et al., 2005 Clines, clusters, and the effect of study design on
the inference of human population structure. PLoS Genet 1: e70.

SCHMID, K. J., O. TORJEK, R. MEYER, H. SCHMUTHS, M. H. HOFFMANN et
al., 2006 Evidence for a large-scale population structure of Arabidopsis
thaliana from genome-wide single nucleotide polymorphism markers.
Theor Appl Genet 112: 1104-1114.

SHARBEL, T. F., B. HAUBOLD and T. MITCHELL-OLDS, 2000 Genetic
isolation by distance in Arabidopsis thaliana: biogeography and postglacial
colonization of Europe. Mol Ecol 9: 2109-2118.

SLATKIN, M., 1995 A measure of population subdivision based on
microsatellite allele frequencies. Genetics 139: 457-462.

SMOUSE, P., J. LONG and R. SOKAL, 1986 Multiple regression and
correlation extensions of the Mantel test of matrix correspondence.
Systematic Zoology 35: 627-632.

STENØIEN, H. K., C. B. FENSTER, A. TONTERI and O. SAVOLAINEN, 2005
Genetic variability in natural populations of Arabidopsis thaliana in northern
Europe. Mol Ecol 14: 137-148.

SYMONDS, V. V., and A. M. LLOYD, 2003 An analysis of microsatellite loci in
Arabidopsis thaliana: mutational dynamics and application. Genetics 165:
1475-1488.

TODOKORO, S., R. TERAUCHI and S. KAWANO, 1996 Microsatellite
polymorphisms in natural populations of Arabidopsis thaliana in Japan. Jpn
J Genet 70: 543-554.
TÖRJÉCK, O., D. BERGER, R. C. MEYER, C. MUSSIG, K. J. SCHMID et al., 2003 Establishment of a high-efficiency SNP-based framework marker set for Arabidopsis. Plant J 36: 122-140.

VAN BERLOO, R., 1999 GGT: software for the display of graphical genotypes. Journal of Heredity 90: 328-329.

WEIR, B., and C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution: 1358-1370.

YEH, F., R. YANG and T. BOYLE, 1999 POPGENE, version 1.32. Microsoft Windows-Based Freeware for Population Genetic Analysis. University of Alberta, Edmonton.

ZHAO, K., M. J. ARANZANA, S. KIM, C. LISTER, C. SHINDO et al., 2007 An Arabidopsis example of association mapping in structured samples. PLoS Genet 3: e4.
Tables

**Table 1.** Genetic diversity of *A. thaliana* in the Iberian Peninsula (*IP*) and the rest of world (*RW*). Table shows: sample size (*N*); percentage of polymorphic loci (*PL*); number of observed alleles (*n*); allelic richness (*R*); private allelic richness (*RP*); gene diversity (*H*); and the number of multilocus haplotypes (*NH*).

*n*, *R*, *RP* and *H* are mean values ± SD estimated from 4 cpMSs, 16 ncMSs, 62 Col/C24 and 47 IP SNP loci.

| Marker set       | Region | N   | PL       | n | R  | RP      | H  | NH  |
|------------------|--------|-----|----------|---|----|---------|----|-----|
| cpMSs            |        |     |          |   |    |         |    |     |
|                   | *IP*   | 100 | 100      | 6.50 ± 2.52 | 5.74 ± 1.86 | 1.35 ± 1.78 | 0.55 ± 0.18 | 34  |
|                   | *RW*   | 43  | 100      | 5.75 ± 2.22 | 5.75 ± 2.22 | 1.35 ± 1.54 | 0.57 ± 0.20 | 23  |
| ncMSs            |        |     |          |   |    |         |    |     |
|                   | *IP*   | 100 | 100      | 18.69 ± 8.38 | 15.91 ± 7.20 | 5.72 ± 3.89 | 0.82 ± 0.18 | 100 |
|                   | *RW*   | 43  | 100      | 14.44 ± 5.30 | 14.35 ± 5.29 | 4.15 ± 1.42 | 0.83 ± 0.23 | 43  |
| Col/C24 SNPs     |        |     |          |   |    |         |    |     |
|                   | *IP*   | 100 | 80.6     | 1.81 ± 0.40 | 1.80 ± 0.39 | 0.01 ± 0.04 | 0.19 ± 0.18 | 100 |
|                   | *RW*   | 193 | 100      | 2.00 ± 0.00 | 1.96 ± 0.11 | 0.18 ± 0.35 | 0.22 ± 0.15 | 193 |
| IP SNPs          |        |     |          |   |    |         |    |     |
|                   | *IP*   | 98  | 97.9     | 1.98 ± 0.15 | 1.98 ± 0.15 | 0.09 ± 0.29 | 0.27 ± 0.18 | 98  |
|                   | *RW*   | 93  | 91.3     | 1.91 ± 0.28 | 1.91 ± 0.28 | 0.02 ± 0.15 | 0.26 ± 0.18 | 92  |
Table 2. Genetic diversity of seven *A. thaliana* local populations from the Iberian Peninsula. Table shows: sample size (*N*); number of multilocus genotypes (*NG*); number of multilocus haplotypes (*NH*); mean gene diversity (*HS*); percentage of pairs of nuclear loci showing significant LD (*Pld*); mean observed heterozygosity (*HO*), inbreeding coefficient (*FIS*) and outcrossing rates (*OR*) estimated from 16 ncMS loci.

|     | Agu  | Cdc  | Leo  | Mar  | Pra  | Qui  | San  | Average |
|-----|------|------|------|------|------|------|------|---------|
| *N* | 21   | 32   | 20   | 30   | 24   | 27   | 21   | 25.0 ± 4.8 |
| *NG*| 14   | 30   | 18   | 25   | 19   | 17   | 19   | 20.3 ± 5.4 |
| *NH*| 8    | 24   | 8    | 16   | 11   | 5    | 16   | 12.6 ± 6.5 |
| *HS*| 0.13 ± 0.18 | 0.25 ± 0.26 | 0.26 ± 0.25 | 0.19 ± 0.24 | 0.33 ± 0.27 | 0.13 ± 0.22 | 0.28 ± 0.27 | 0.22 ±0.27 |
| *Pld*| 12.72 | 5.58 | 22.85 | 14.43 | 31.36 | 62.34 | 8.71 | 22.6 ± 19.6 |
| *HO*| 0.05 ± 0.06 | 0.02 ± 0.02 | 0.03 ± 0.06 | 0.03 ± 0.04 | 0.01 ± 0.02 | 0.01 ± 0.02 | 0.00 ± 0.01 | 0.02 ± 0.02 |
| *FIS*| 0.87 ± 0.13 | 0.97 ± 0.04 | 0.96 ± 0.08 | 0.94 ± 0.06 | 0.99 ± 0.02 | 0.96 ± 0.05 | 0.99 ± 0.02 | 0.96 ± 0.04 |
| *OR*| 0.07 ± 0.07 | 0.02 ± 0.02 | 0.02 ± 0.04 | 0.03 ± 0.03 | 0.01 ± 0.01 | 0.02 ± 0.03 | 0.00 ± 0.01 | 0.03 ± 0.02 |
Figure legends

Figure 1. Geographical location of *A. thaliana* populations of the Iberian Peninsula surveyed in this work. Numbers I to VI and dotted lines indicate the six IP geographical subregions considered in spatial analyses. Red spots depict the location of the seven large populations used for local population differentiation analysis.

Figure 2. Chlorotype networks of *A. thaliana* individuals from the Iberian Peninsula and the rest of world. A) Network of chlorotypes present in the IP and in the rest of world. Pie sizes are proportional to the total chlorotype frequency, while gray and black sectors correspond to relative frequencies in IP and in the rest of world respectively. B) Network of Iberian chlorotypes. Pie sizes are proportional to IP frequency and the four coloured sectors correspond to relative frequencies in the IP genetic groups inferred with STRUCTURE. Yellow, blue, green and red depict IP genetic groups 1, 2, 3 and 4, respectively. Each branch corresponds to one mutational step between chlorotypes. Non-observed mutational steps between chlorotypes are indicated by perpendicular dashes. See Materials and Methods and supplemental Table S5 for details.

Figure 3. World-wide population structure of *A. thaliana*. Upper panel shows each individual as a vertical line divided in segments representing the estimated membership proportions in the four ancestral genetic clusters inferred with STRUCTURE from 62 Col/C24 SNP loci. Individuals are classified in eight geographical regions indicated on the top. Individuals within each region are arranged according to estimated cluster membership proportions. The lower
The table shows the number of individuals and the mean membership fractions in the four genetic clusters (C1 to C4) for each region. Standard deviations of all mean membership proportions were smaller than 0.001 (not shown).

**Figure 4.** Population structure of *A. thaliana* in the Iberian Peninsula. Genetic relationships among 100 IP individuals from different populations were estimated with STRUCTURE using 95 polymorphic SNP loci. Each individual is depicted as a vertical rectangle divided in segments representing the estimated membership proportions in the ancestral genetic clusters (K) fitted in the model. Individuals are arranged according to estimated cluster membership proportions. Arrows at the bottom indicate individuals from the seven populations used for local population differentiation analyses.

**Figure 5.** Geographical structure of *A. thaliana* in the Iberian Peninsula. A) Correlation between geographical and genetic distances in seven local populations extensively sampled. B) Correlation between geographical and genetic distances among six IP geographical subregions. In A and B, genetic distances are estimated from all 115 polymorphic MS and SNP loci. C) Geographical location and genetic composition of 100 individuals from different populations. Each population is shown as a pie chart representing the membership proportions in the four genetic clusters inferred with STRUCTURE (colours of clusters as in Figure 4; cluster 1, yellow; cluster 2, blue; cluster 3, green; cluster 4, red). * indicate the seven populations used for local population differentiation analyses.
A

Genetic distance $[F_{ST}/(1-F_{ST})]

Geographical distance (km)

$F_{ST}=0.34$

$r=0.63$

$R^2=0.39$

B

Genetic distance $[F_{ST}/(1-F_{ST})]

Geographical distance (km)

$F_{ST}=0.08$

$r=0.64$

$R^2=0.41$

C

Map of Spain with different colored circles indicating genetic diversity.

Legend:
- Yellow
- Blue
- Green
- Red

Map scale: 50 km at 0, 250 km at corner.
