Theor Appl Genet (2006) 112: 1104–1114
DOI 10.1007/s00122-006-0212-7

ORIGINAL PAPER

**Karl J. Schmid · Ottó Törjék · Rhonda Meyer
Heike Schmuths · Matthias H. Hoffmann
Thomas Altmann**

Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers

Received: 15 August 2005 / Accepted: 28 December 2005 / Published online: 2 February 2006

Abstract Population-based methods for the genetic mapping of adaptive traits and the analysis of natural selection require that the population structure and demographic history of a species are taken into account. We characterized geographic patterns of genetic variation in the model plant *Arabidopsis thaliana* by genotyping 115 genome-wide single nucleotide polymorphism (SNP) markers in 351 accessions from the whole species range using a matrix-assisted laser desorption/ionization time-of-flight assay, and by sequencing of nine unlinked short genomic regions in a subset of 64 accessions. The observed frequency distribution of SNPs is not consistent with a constant-size neutral model of sequence polymorphism due to an excess of rare polymorphisms. There is evidence for a significant population structure as indicated by differences in genetic diversity between geographic regions. Accessions from Central Asia have a low level of polymorphism and an increased level of genome-wide linkage disequilibrium (LD) relative to accessions from the Iberian Peninsula and Central Europe. Cluster analysis with the `structure` program grouped Eurasian accessions into $K=6$ clusters. Accessions from the Iberian Peninsula and from Central Asia constitute distinct populations, whereas Central and Eastern European accessions represent admixed populations in which genomes were reshuffled by historical recombination events. These patterns likely result from a rapid postglacial recolonization of Eurasia from glacial refugial populations. Our analyses suggest that mapping populations for association or LD mapping should be chosen from regional rather than a species-wide sample or identified genetically as sets of individuals with similar average genetic distances.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00122-006-0212-7 and is accessible for authorized users.

Communicated by O. Savolainen

Karl J. Schmid and Ottó Törjék contributed equally to this work.

K. J. Schmid
Max-Planck-Institute of Chemical Ecology, Jena, Germany

O. Törjék · R. Meyer · T. Altmann
University of Potsdam, Germany and Max-Planck-Institute of Molecular Plant Physiology, Golm, Germany

H. Schmuths
Institute for Plant Genetics and Crop Plant Research, Gatersleben, Germany

H. Schmuths
University of Nottingham, Loughborough, UK

M. H. Hoffmann
Institute of Geobotany and Botanical Gardens, Martin-Luther-University, Halle, Germany

K. J. Schmid (✉)
Department of Genetics and Evolution, Max-Planck-Institute of Chemical Ecology, Hans-Knöll-Str. 8, 07745 Jena, Germany
E-mail: schmid@ice.mpg.de

Introduction

Genetic variation within a species is influenced by evolutionary processes that affect the whole genome like demographic history and the breeding system, or that are variable across the genome like recombination rate, mutation rate, or selection. The pattern of genetic variation observed at any given locus has been generated by a combination of genome-wide and locus-specific factors. This has the consequence that the demographic history of a species needs to be accounted for in methods that utilize natural variation for mapping of genes involved in phenotypic variation [e.g., linkage disequilibrium (LD) mapping]. A significant association between genotypic and phenotypic variation may result from unrecognized population structure rather than from a causal relationship between genotypic and phenotypic variation at a given locus and lead to false positive associations (Pritchard and

Rosenberg 1999). In a similar fashion, a test of the hypothesis that genetic variation at a locus was influenced by selection requires disentangling the effects of different evolutionary processes on genetic variation.

The model plant *Arabidopsis thaliana* (L.) Heynh. exhibits a high level of genetic and phenotypic diversity (Mitchell-Olds 2001; Koornneef et al. 2004), and genomic resources are available to analyze this variation (Borevitz and Nordborg 2003). Genetic diversity was examined using a variety of genetic markers including allozymes (Abbott and Gomes 1989), RFLPs (Bergelson et al. 1998), AFLPs (Miyashita et al. 1999; Sharbel et al. 2000), microsatellites (Kuittinen et al. 1997; Vander Zwan et al. 2000) and sequence surveys of single genes (reviewed by Wright and Gaut 2005). In early studies, no correlation between genetic and geographic distance was found in samples representing the whole species range that includes Eurasia, North Africa, North America and East Asia (e.g., Miyashita et al. 1998; Bergelson et al. 1998). This was explained by a preference for disturbed habitats such as agricultural fields. A close association with human agriculture may have obscured historical patterns. In contrast, more recent studies uncovered the existence of a large-scale population structure (Sharbel et al. 2000; Nordborg et al. 2005). Sharbel et al. (2000) proposed a demographic model with glacial refugia on the Iberian Peninsula and Central Asia and a subsequent admixture of these populations in Central and Eastern Europe since the last glaciation.

Here we further characterize the population structure of *A. thaliana* by analyzing single nucleotide polymorphisms (SNPs). SNPs are abundant in the genome of *A. thaliana* (Jander et al. 2002; Schmid et al. 2003; Nordborg et al. 2005) and are useful markers for phylogeographic analyses (Brumfield et al. 2003), although ascertainment bias needs to be considered (Morin et al. 2004). We utilize a set of 'framework' SNP markers that are evenly distributed throughout the genome, separated by an average distance of 1.13 Mb (Törjék et al. 2003). To evaluate the effect of SNP ascertainment bias, we also sequenced a subset of the accessions at nine, short, randomly chosen loci. We analyze the current geographic population structure using a large number of accessions and compare levels of genetic variation between populations from different geographic regions in Eurasia to characterize demographic processes that may contribute to a genome-wide departure from a neutral model of sequence polymorphism (Schmid et al. 2005). The present survey includes nearly all natural accessions that were available from Arabidopsis stock centers in 2003, and an additional set of accessions that was recently collected in Central Asia, in one of the putative refugia. We find different levels of genetic diversity among geographic regions in the native species range and the existence of a large-scale population substructure.

Materials and methods

Plant material

Seeds of accessions analyzed were obtained from various sources. Accessions available through the *Arabidopsis* stock centers include Col-0 from G. Rédei (Univ. of Missouri-Columbia, USA); C24 from JP Hernalsteens (Vrije Universiteit Brussels, Belgium); Landsberg *erecta* from M. Koornneef (Wageningen University, Netherlands); Ag-0, An, Bch-1, Bur, Cal, Co, Cvi, Ei, Eil-0, Gr, Hi, Lip-0, Lm, Lu, Ob-0, Old-1, Per, Oy, Sue, Sg-1 and Te from S. Misera (Institut für Pflanzengenetik und Kulturpflanzenforschung, Gatersleben, Germany); a further 286 accessions from the Nottingham Arabidopsis Stock Center (NASC), the Arabidopsis Biological Resource Center (ABRC), Columbus, Ohio, and the SENDAI Arabidopsis seed stock center (SASSC), Sendai, Japan. An additional 41 accessions were included that originate from various locations in Russia and Uzbekistan (Schmuths et al. 2004). These accessions were collected in 2001 and 2002 by H. S. and M. H. Accessions were grouped into the following seven geographical regions: Scandinavia; British Isles, Central Europe (northern coast to the Alps); Iberian Peninsula; Southern Italy (south of the Po valley); Asia and Africa to reflect the physical barriers separating geographical regions during glaciation and include several refugial areas (Hewitt 1999).

SNP markers

Single nucleotide polymorphism markers used in this study are a subset of polymorphisms detected in a survey of 13 *A. thaliana* accessions (Schmid et al. 2003; Törjék et al. 2003). A framework set of 115 SNPs was identified from all available markers based on the following criteria (Törjék et al. 2003): (1) SNPs are polymorphic between the C24 and Col-0 accessions because they are also utilized for genotyping mapping populations derived from these two accessions; (2) a physical distance of ca. 1.15 Mb between adjacent SNPs (Fig. S1); (3) SNPs that were polymorphic between Ler and Col-0 in addition to C24/Col were preferred ($n = 62$).

DNA isolation and genotyping

Genomic DNA was extracted from 50 mg leaf tissue of two to three plants each using the NucleoSpin Multi-96 Plant kit (Macherey-Nagel, Düren, Germany) or the DNA easy extraction kit (QIAGEN, Hilden, Germany). All accessions were genotyped with the set of 115 SNP markers, established for matrix-assisted laser desorption/ionization time-of-flight (MALDI-ToF) analysis (performed by GAG-Bioscience GmbH, Bremen,

Germany). A subset of SNPs was genotyped in a smaller number of accessions using the SNaPshot™ method (Applied Biosystems). Both MALDI-ToF analysis and SNaPshot™ reactions were carried out as described (Törjék et al. 2003). Accessions with >20% missing data ($n=7$) and heterozygous genotypes (>5% of SNPs; $n=9$) were excluded from further analysis.

Error rate of SNP genotyping

The MALDI-ToF genotyping technology can have an error rate of up to 5% (Bray et al. 2001). We assessed the reliability of the MALDI-ToF genotypes using two approaches. First, five randomly chosen accessions were genotyped twice by MALDI-ToF with the whole set of SNP markers. Among the five duplicate sets ($5\times115=575$ comparisons), 521 genotypes (86.6%) could be obtained from both duplicates. Among these, only one (0.2%) differed between two duplicates. Second, we resequenced 29 SNPs from 16 accessions (464 genotypes) with the SNaPshot™ method. From the resequenced genotypes, 408 (87.9%) could be obtained from both replicates. Four genotypes (0.98%) differed between the MALDI-ToF and SNaPshot assays, and among those, three were determined to be heterozygous by either MALDI-ToF or SNaPshot assays. These controls indicate an overall error rate of less than 1%.

Analysis of SNP data

The frequency distribution of SNP markers was corrected for ascertainment bias and compared to a distribution based on the neutral equilibrium model as described by Nielsen et al. (2004). To reconstitute the unbiased allele frequency distribution, we used their base model, which assumes that the ascertainment sample of two accessions (Col-0 and C24) is part of the total sample of accessions. The maximum likelihood (ML) estimate of the corrected frequency distribution $P=(p_1,p_2,\ldots,p_{n-1})$ where p_i is the frequency of SNPs with the mutant allele frequency of i in a sample of n chromosomes, was calculated using Eq. 3 of Nielsen et al. (2004). The ML of estimated $\widehat{P}$ was compared with the likelihood obtained from the expected frequency distribution under a standard neutral model of constant population size (P_c), where p_i was calculated as $p_i = 1/i\sum_{i=1}^{n-1} 1/j$ with $0<i<n$. The likelihood ratio was calculated as $\log(L(\widehat{P})/L(P_c))$. To test whether the reconstituted allele frequency distribution differs significantly from a neutral distribution, the value of the likelihood ratio was compared to a distribution of simulated likelihood ratios. These were obtained by dividing ML estimates of simulated neutral frequency distributions generated using our ascertainment scheme with the likelihood of the neutral expected distribution, P_c.

Gene diversity was estimated as

$$H = [n/(n-1)]\left(1-\sum_{i=1}^{n} p_i^2\right)$$

where n is the number of alleles and p_i is the relative frequency of allele i of a single SNP (Nei 1987). Polymorphism across loci in a group of accessions was estimated as the mean of gene diversity among all SNP markers. Gene diversity estimates between groups of accessions from different geographic regions are not independent because of a shared underlying genealogy. For this reason, bootstrap analysis by resampling of accessions with 10,000 replicates was employed to calculate the 95% confidence interval of gene diversity estimates. The 2.5 and 97.5% percentiles of the bootstrap distribution were obtained using the bias-corrected method described by Dixon (2001, p. 279). A neighbor-joining (NJ) tree (Saitou and Nei 1987) was constructed with the `neighbor` program of the PHYLIP package (Felsenstein 1989) using a matrix of the proportion of uncorrected pairwise differences. SNPs with missing data in pairwise comparisons were excluded (pairwise deletion).

Correlations between matrices of genetic and geographic distances were calculated as the normalized Mantel statistic, r, (Mantel 1967) with 999 permutations, using R PACKAGE 4.0 (http://www.fas.umontreal.ca/BIOL/Casgrain/en/labo/R). Levels of pairwise gametic LD were estimated as r^2 (Hill and Robertson 1968). For comparisons of LD between populations, we calculated $|D'|$ (Lewontin 1964), because it is independent of allele frequencies, which may differ between populations. Tests for significant pairwise LD were conducted with the χ^2 statistic (Weir 1996) and sequential Bonferroni correction of P values (Sokal and Rohlf 1995).

SNP homogeneity

Geographic regions with a high degree of homogeneity (i.e., a high proportion of shared nonvariable SNPs) were identified by an ad hoc method. Accessions were mapped on a grid of (arbitrarily chosen) 50 rows and 214 columns that covers the Eurasian distribution range. SNPs were evaluated separately whether they were monomorphic or polymorphic within grids. If a cell and its neighboring cells were polymorphic for a certain SNP, it was given a value of 1, otherwise 0. In a last step, grid values of individual SNPs were added and used to manually identify the geographic regions with similar SNP homogeneities.

Clustering of accessions

Accessions were clustered with the `structure` program (version 2.0; Pritchard et al. 2000), which uses a Bayesian framework to infer the number of clusters, K, in a sample of genotypes. The likelihood of the data

given K is calculated and the value of K with the highest likelihood can be interpreted to correspond to the number of clusters in the sample. In addition to calculating the likelihood, `structure` also assigns every genome proportionally to each cluster and thus allows identification of groups of accessions that form a population. We used a model with admixture and allowed 10,000 runs for burn-in and an additional 100,000 runs to estimate parameters for populations. Because *A. thaliana* is a highly selfing species, genotypes were assumed to be haploid and heterozygous genotypes were treated as missing data. It should be noted that no correction for SNP ascertainment bias is currently implemented in `structure`. To visualize the assignment of accessions to different clusters, we used the `distruct` program (Rosenberg 2004) and the ArcView GIS (Environmental Systems Research Institute 1992).

Analysis of sequence data

Sequencing of short genomic regions, sequence assembly and annotation of coding regions was done as described using an automated sequence analysis pipeline (Schmid et al. 2003, 2005). The basic analysis of genetic variation in sequence alignments was performed with DnaSP 4.0 (Rozas and Rozas 1999). Sequence variation was examined as average pairwise nucleotide diversity, π (Nei 1987). All nucleotide polymorphisms were included, but insertion/deletion polymorphisms were excluded from the analysis. Bootstrap and other resampling analyses were done with scripts written in the Python programming language. If not indicated otherwise, statistical calculations were conducted with the R statistical package (http://www.r-project.org).

Data availability

Summary information about the accessions and SNP markers used in this study, and the genotype data in tabular form are available at http://www.mpimp-golm.mpg.de/arab-diversity. The SNP genotypes were also deposited in NCBI dbSNP under accession numbers: 49785508–49785622 and sequence alignments in NCBI GenBank under accessions numbers DU711026–DU711571.

Results

Summary of genotyping

We genotyped 351 accessions and retained 335 for further analysis after quality control. Among the 38,525 genotypes (115 SNPs × 335 accessions), 3,324 (8.6%) constitute missing data and 78 (0.2%) are heterozygous, which may result from contamination of seed stocks or DNA samples, the duplication of loci, or recent outcrossing in the natural environment. The proportions of missing data are not different among geographic regions (Table 1) and should not affect the inference of population structure. Only one SNP was triallelic (MASC01582), and all others biallelic. Genotypes were classified as being of Col-0 type or C24 type. The mean frequency of the Col-0 type genotypes (37.2%) was less than the frequency of the C24 type genotypes (53.5%). The two frequencies differ significantly (Wilcoxon's signed rank test, $V = 4{,}040.5$, $P = 0.04913$). This difference is likely a result of the ascertainment scheme, because we preferred SNPs that were polymorphic in Ler/Col-0 in addition to C24/Col-0, which has the effect that the chosen SNP has a population frequency of Col-0 allele of about 33%, which is close to the observed frequency. Most SNP markers are located in intergenic regions (64%) and a small fraction (8%) is located in coding regions at replacement sites. SNPs of the latter type have the lowest gene diversity among all SNP types, but the difference compared to other SNP types is not significant (one-way ANOVA: $F = 0.2271$, $df = 3$, $P = 0.87$); for this reason, all SNPs were combined in the following analyses.

Table 1 Genetic diversity, H, and proportion of missing data in different geographic regions

Region	N	H (95% CI)	Proportion of missing data (SE)
America	11	0.329 (0.133–0.376)	0.097 (0.051)
Africa	4	0.206 (0.0–0.206)	0.067 (0.029)
Iberian Peninsula	32	0.252 (0.175–0.302)	0.092 (0.005)
British Isles	11	0.232 (0.146–0.249)	0.087 (0.011)
Central Europe	188	0.245 (0.228–0.257)	0.083 (0.002)
Southern Italy	9	0.221 (0.138–0.221)	0.102 (0.014)
Scandinavia	10	0.221 (0.136–0.215)	0.091 (0.010)
Eastern Europe	22	0.185 (0.152–0.190)	0.102 (0.006)
Central Asia	40	0.095 (0.079–0.100)	0.080 (0.005)
Eastern Asia	1	–	0.043 (-)
Unknown origin	7	0.385 (0.133–0.462)	0.096 (0.056)
Total	335	0.245 (0.230–0.256)	0.086 (0.002)

95% Confidence intervals (CI) were determined by bootstrap analysis

Relationships among accessions

We identified genetically similar accessions by calculating a NJ tree that is based on the pairwise genetic distance (Fig. S2). As expected from the ascertainment procedure, the two accessions from which the SNPs were identified (Col-0 and C24) are maximally distant from each other. The tree reveals the previously observed star-like phylogeny with low bootstrap support of internal branches and long terminal branches. Only the newly collected Central Asian accessions deviate from a star phylogeny pattern. Some accessions are highly similar to C24 or Col-0. For example, the Co accession originating from Coimbra (Portugal) clusters closely with C24, indicating that the C24 accession, whose origin is unknown, is probably derived from Co and not from Col-0, as is sometimes suggested in the literature (e.g. Loridon et al. 1998). Individuals whose genotypes do not differ by more than two SNPs (the expected error rate of the MALDI-ToF assay) can be considered genetically identical. Using this criterion, 96 accessions were grouped into 35 sets of essentially identical accessions, of which several pairs consist of accessions from distant geographic origin (Table S1).

Frequency distribution of SNPs

To analyze genome-wide patterns of SNP variation, we first compared the frequency distribution of SNPs with the expectation of a neutral model of sequence polymorphism (constant population size, no population structure). Observed SNP frequencies were corrected for ascertainment bias using the base model of Nielsen et al. (2004), which assumes that SNPs are unlinked and drawn from a panmictic population of constant size. Since information about the ancestral state was not available, minor allele frequencies were used for the calculation. The comparison of the corrected observed distribution with the distribution expected under a neutral equilibrium model shows an excess of rare polymorphisms (Fig. 1). The likelihood ratio $\log(L(\widehat{P})/L(P_c))$ of observed and expected neutral frequency distributions was 98.49. To test for consistency with a neutral model, 100 simulated SNP frequency distributions were generated using coalescent theory and the same SNP ascertainment scheme. None of the simulated likelihood ratios (mean = 74.51; SD = 5.95) was larger than the observed one, leading us to conclude that the observed data are not consistent with a neutral model because of demographic factors or selection. Figure 1 shows that the deviation is due to an excess both of rare and high-frequency minor polymorphisms. The same result was obtained when we restricted the analysis to accessions from the Eurasian continent ($n = 233$).

Geographic population structure

We investigated whether a large-scale geographic population structure contributes to the deviation from a distribution expected under a neutral model by comparing allele frequencies in different geographic regions

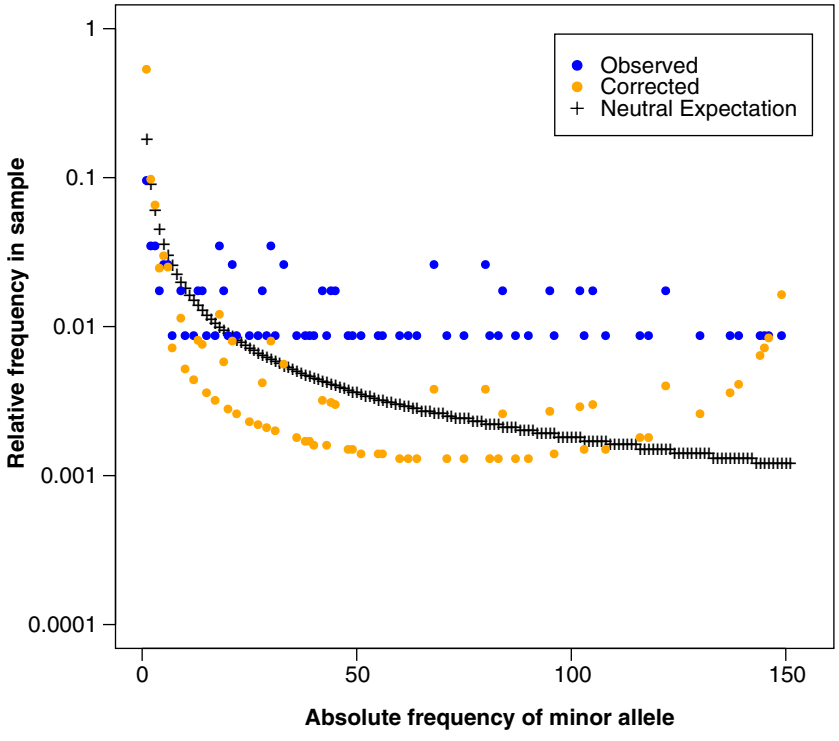

Fig. 1 Frequency distribution of the minor allele of 115 SNP markers genotyped in 304 accessions of *Arabidopsis thaliana*. Only one of each group of genetically identical accessions was included. Observed allele frequencies (*blue*) were corrected for ascertainment bias using the basic model described by Nielsen et al. (2004)

Table 2 Average pairwise linkage disequilibrium ($|D'|$) among SNP markers with a minor frequency of >0.1

Geographic region	Accessions	SNPs	Average $\|D'\|$ (95% CI)
Iberian Peninsula	32	59	0.375 (0.300–0.397)
Central Europe	188	66	0.157 (0.134–0.175)
Eastern Europe	22	48	0.375 (0.299–0.441)
Central Asia	40	25	0.357 (0.265–0.432)
Total	335	65	0.163 (0.142–0.180)

(Table 1). Within Eurasia, gene diversity was lowest among Central Asian accessions ($H=0.095$) and the highest among Iberian accessions ($H=0.252$). Bootstrap analysis showed that gene diversity values of Eastern European and Central Asian accessions are significantly lower than in other geographic regions of comparable size.

To verify that the observed differences in genetic diversity among geographic regions are not an artifact of the SNP ascertainment scheme, we conducted sequence surveys of nine randomly chosen, short genomic regions (average length 430 bp); such surveys produce unbiased estimates of genetic diversity. These regions were sequenced in most Iberian ($n=22$) and Central Asian accessions ($n=31$). Additional sequences from a set of 12 divergent accessions of mostly Central and Eastern European origin (“divergent set”) were obtained from Schmid et al. (2005). We identified 52 SNPs at the 9 loci (Table 3). Average levels of nucleotide diversity (π) differ between populations. They are the smallest in the Central Asian sample (0.00088), about two times larger in the Iberian sample (0.00166), and the highest in the diverse set (0.00306). We conducted pairwise comparisons of per-locus bootstrap estimates of π to test for differences between populations. Sequence diversity in the diverse set is higher here than those of Central Asian ($P=0.004$) and Iberian accessions ($P=0.009$). Diversity levels between the latter two sets are not significantly different ($P=0.17$).

A Mantel test using the SNP data indicates the presence of isolation by distance (i.e., a correlation between genetic and geographic distance) among Eurasian accessions ($n=308$, $r=-0.0645$, $P=0.035$). Significant positive correlations exist for all comparisons between Central Asia and other regions except Central Europe, and for the comparisons between the Iberian Peninsula with Central Europe and the British Isles, respectively (Table S2). Within geographic regions, there is isolation by distance among Iberian ($r^2=0.177$, $P<0.0001$) and Central Asian accessions ($r^2=0.323$, $P<0.0001$; Fig. 2). These results are essentially identical to that obtained by Sharbel et al. (2000) using unbiased AFLP markers. The most significant correlations are obtained when the Central Asian accessions are included, which likely result from the fact that they are genetically distinct (see below) and separated by a large distance from the other geographic regions. The observed relationship among Central Asian accessions is unexpected because genetic and geographic distances are positively correlated for accessions within a range of 1,000 km, but negatively correlated for larger distances (Fig. 2b). Thus, there are genetically similar accessions in geographically distant Central Asian locations, such as Siberia and Uzbekhistan.

The a priori definition of geographic regions may not correspond with the genetic population structure and can confound estimates of genetic diversity and the analysis of historical processes. We, therefore, clustered accessions based on their genetic similarity and then asked whether genetically defined populations correlate with geographic regions. We first determined the degree of homogeneity among SNP markers within the Eurasian distribution range and defined regions having accessions of similar levels of homogeneity (see Materials and methods). Two spatial gradients with different levels of homogeneity were identified (Fig. 3). One gradient ranges from Central Asia (high degree of homogeneity corresponding to a low level of genetic diversity) to Central Europe (little homogeneity) and the other from Africa and the Iberian Peninsula (high homogeneity) to Central Europe.

To obtain a more fine-grained clustering, we grouped accessions into distinct clusters using the program `structure`, which implements a model-based clustering algorithm (Pritchard et al. 2000). We estimated the number of K clusters that is most consistent with the observed data. The average likelihood values of five runs for a given value of K increase gradually until $K=6$; at that point the likelihoods reach a maximum and then

Table 3 Summary of sequence survey of nine randomly chosen STS loci in different geographic regions

STS ID	Chr.	Position	Base pair	All accessions				Nucleotide diversity, π		
				n	S	θ_w	π	Central Asia ($n=31$)	Iberia ($n=22$)	Divergent set ($n=12$)
AtI20	1	17532497	422	62	3	1.51	1.87	0.53	2.11	2.28
At1est13	1	29475355	444	64	2	0.95	2.24	1.57	1.05	1.68
TIGR3187	2	17006898	470	56	2	0.92	0.15	0.00	0.19	0.35
At3est52	3	4822577	475	58	4	1.81	1.59	0.00	1.33	2.60
At3est47	3	6995953	364	57	2	0.59	0.27	0.00	0.29	0.77
AtIII19x5	3	8624955	365	62	6	3.49	2.87	1.06	2.10	7.26
AtIV20	4	14369614	599	60	2	0.72	0.11	0.00	0.00	0.55
AtV9	5	6413216	385	64	19	4.93	1.00	2.05	6.30	7.52
AtV23	5	16353490	323	62	12	7.88	3.23	2.69	1.56	4.49

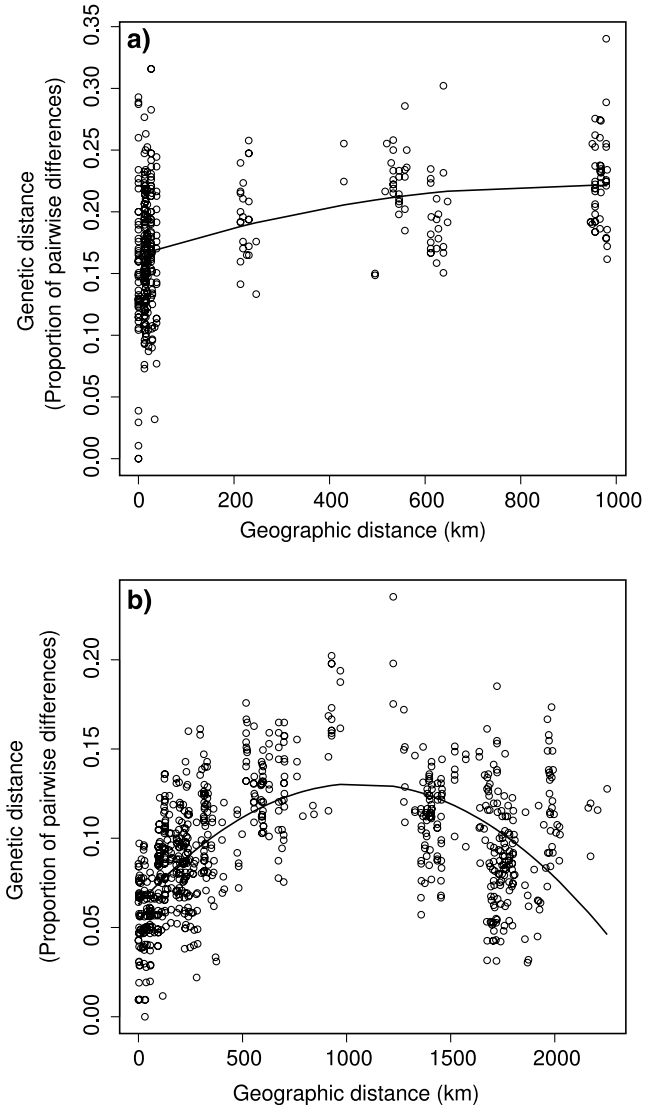

Fig. 2 Positive correlation between geographical and genetic distance indicates isolation by distance in *A. thaliana* (**a**) on the Iberian Peninsula and in Central Asia (**b**). The *line* corresponds to the fitted curve of a quadratic regression

drop for higher values of K (not shown). We also examined the geographic distribution of inferred clusters. With $K=3$, we observe one cluster consisting mainly of accessions from the Iberian Peninsula and the Mediterranian region, a second cluster with accessions from Central Europe, and a third cluster in Central Asia (Fig. 4a). This resembles a geographic structure similar to the one based on SNP homogeneity (Fig. 3). With $K=6$, several geographically restricted clusters are recognizable (Figs. 4a, S3), including a cluster of Iberian and Western European accessions, several different geographically overlapping clusters in Central Europe, a large cluster in Eastern Europe and Scandinavia, and a cluster in Central Asia. A similar result was obtained when the sequence survey data were clustered with `structure`. We obtained the best fit with $K=3$ (Fig. 4b). As observed with the SNP data, the Iberian and Central Asian accessions form distinct and homogeneous clusters. In summary, both the SNP and the sequence survey data indicate the presence of a large-scale population structure across an East–West gradient throughout the Eurasian continent. Accessions from the Iberian Peninsula and from Central Asia appear to be genetically distinct and homogenous clusters.

Analysis of LD

Although *A. thaliana* is a highly self-fertilizing species (Abbott and Gomes 1989), sufficient outcrossing has occurred in the history of the species to lead to a decay of LD within 250 kb (Nordborg et al. 2002, 2005). With the exception of five SNPs, the physical distance between SNPs is >1 Mb, and for this reason we expected them to be in linkage equilibrium. To verify this assumption and to test for different levels of LD in different populations, pairwise LD was estimated for all pairs of SNPs with a frequency of >0.1 of the minor allele. The mean r^2 for all 2,080 pairwise comparisons is 0.0183 and the median is 0.0080. Despite a low average r^2, 372 (18%) pairwise comparisons exhibit significant LD after Bonferroni correction. As expected from the large physical distance between markers, there is no correlation between r^2 and physical distance of markers, although five of the most significant six pairwise comparisons with $r^2>0.2$ include markers that are located adjacent to each other. The high proportion of significant pairwise correlations may be caused by a low-effective recombination rate and the presence of a population structure.

To evaluate the effect of population structure on LD, we compared genome-wide LD between geographic regions (Iberian Peninsula, Central Europe, Eastern Europe and Central Asia). First, we computed the average $|D'|$ for all accessions from a geographic region (Table 2). Genome-wide LD is lowest among Central European accessions, intermediate among Central Asian and Eastern European and highest among Iberian accessions. Because numbers of accessions and SNPs differed among geographic regions, we specifically tested whether the low level of LD among Central European accessions is an artifact of these differences. Average $|D'|$ values for Central European accessions were computed using the same number of randomly selected accessions and SNP markers as in the other regions as given in Table 2 (i.e., 40 accessions and 25 SNPs in comparison with Central Asian accessions). We then counted how often among 1,000 repetitions the average $|D'|$ value of Central European accessions was larger than in the other region. The genome-wide LD among Central European accessions is significantly lower than among from Central Asian ($P=0.015$, one-tailed) and Iberian accessions ($P=0.023$), but not among Eastern European accessions ($P=0.095$). These analyses suggest that

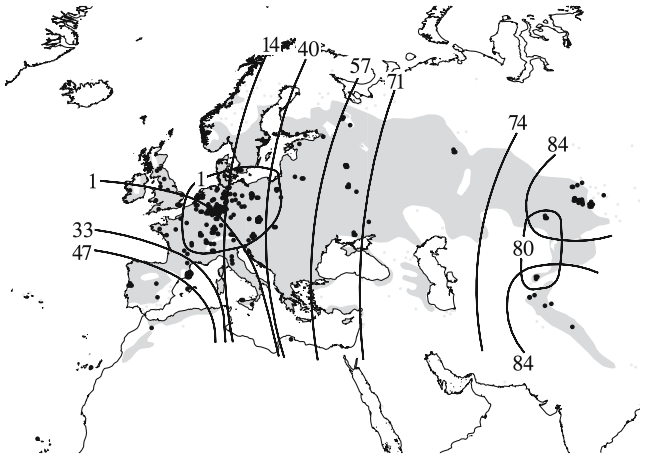

Fig. 3 Distribution of homogeneity of 115 SNPs markers among *A. thaliana* accessions across the Eurasian distribution range. The composite map shows the locations of the accessions included in the analyses and the isolines show the absolute number of SNPs that are invariant (i.e., not polymorphic) in a geographic region

demographic processes contribute to differences in pairwise LD between geographic regions.

Discussion

Possible effects of sampling and ascertainment bias

We first consider the effects of sampling and marker ascertainment bias on observed patterns of genetic variation. The accessions included in our sample were collected independently by different researchers and do not represent a well-designed hierarchical sampling scheme. Central Europe has been extensively sampled, the Iberian Peninsula and Central Asia to a lesser degree; the western part of Russia and potential glacial refugia such as Italy and the Balkans, which are part of the natural species range (Hoffmann 2002), are not well represented in current collections (Fig. 3). The analyses of both gene diversity and genome-wide LD demonstrate that a sampling bias due to a larger number of accessions from Central Europe relative to other geographic regions of a similar size does not confound estimates of genetic variation. The observed differences between geographic regions are not an effect of different sample sizes.

The SNP markers used in this study were selected from a panel of two accessions (Col-0 and C24). An ascertainment panel of two accessions leads to an overrepresentation of high frequency SNPs (Eberle and Kruglyak 2000) and to a loss of resolution in the detection of recent historical events, because the resulting genealogies tend to have reduced terminal branch lengths (Brumfield et al. 2003). This bias affects estimates of population parameters such as gene diversity, H. Several methods for ascertainment bias correction are available (e.g., Kuhner et al. 2000; Nielsen et al. 2004), but they are currently implemented in only few computer programs and are not able to account for population structure. Future SNP markers to be used in genotyping of *A. thaliana* should be derived from a large panel consisting of accessions from the whole species range (Akey et al. 2003). We accounted for ascertainment bias in the analysis of the frequency distribution of SNPs, and obtained independent evidence by sequence surveys for estimates of regional diversity and the inference of geographic population structure. The major conclusions drawn from our analyses should not be confounded by ascertainment bias because consistent results were obtained with both types of data.

The effect of demographic history on genetic variation

Accessions of *A. thaliana* show a genome-wide excess of rare polymorphisms and harbor a set of loci with a very high level of polymorphism (Nordborg et al. 2005; Schmid et al. 2005). In this study, a similar result was obtained by the analysis of the frequency spectrum of SNPs due to an excess of polymorphisms with either a low or high frequency of the minor allele (Fig. 1). This confirms that a standard null model is not appropriate for tests of a neutral model of sequence polymorphism in this species, at least for samples that include accessions from the whole species range. A complex demographic history, such as a combination of population growth and structure, and not selection at multiple loci appear to be largely responsible for this deviation from a neutral model (Nordborg et al. 2005; Schmid et al. 2005). Important demographic factors likely include the presence of a large-scale population structure, historical changes in population size, and migration. Our data support such an interpretation because of different levels of genetic diversity and LD among geographic regions, (Tables 1, 2) and the clustering of accessions along an East–West gradient in Eurasia (Figs. 4, S3). Accessions were grouped into $K=6$ clusters based on the SNP data and into $K=3$ clusters based on the sequence data. By sequencing 876 genomic fragments in 96 accessions, Nordborg et al. (2005) grouped the data into $K=8$ clusters and the geographic distribution of these clusters is similar to the one observed in this study. A recent simulation study of hierarchical population models showed that the `structure` program tends to identify the uppermost level of a population hierarchy and that within these clusters there can be sublevels of structuring (Evanno et al. 2005). In inbreeding species such as *A. thaliana*, `structure` likely overestimates the total number of subpopulations (Falush et al. 2003). For these reasons, clusters inferred by `structure` should not be viewed as separate panmictic populations, but as groups of genetically similar accessions.

The absence of a population structure in early studies of genetic variation in *A. thaliana* was interpreted to be the result of human disturbance. Although we also find some evidence for human-induced long-distance

migration, the previously observed lack of a geographic pattern may result from the small numbers of accessions and markers used and does not hold up if a larger number of markers and accessions are analyzed.

Glacial refugia and postglacial recolonization

Most of the present Eurasian distribution range was an unsuitable habitat for *A. thaliana* during the last glacial maximum (LGM), about 18,000 years ago (Frenzel et al. 1992). Populations were likely separated into glacial refugia on the Iberian Peninsula, Southern Italy, the Balkans, and portions of Central Asia (Comes and Kadereit 1998). The occupation of these refugia may have contributed to the genetic differentiation between refugia by genetic drift or local adaptation. Patterns of SNP variation in the present study appear to be consistent with glacial refugia and a postglacial range shift. Accessions from Iberia and Central Asia belong to distinct and genetically diverged clusters. In contrast, accessions from Central and Eastern Europe constitute recently admixed populations with a high level of secondary genetic diversity. Both the geographic patterns of SNP homogeneity and the inferred population structure suggest that the admixture zone includes the northern part of Central Europe and Eastern Europe that roughly correspond to the location of postglacial "suture zones" observed in other species (Hewitt 1999). The geographic distribution of inferred clusters can then be interpreted as having resulted from a recolonization of Central Europe from refugia on the Iberian Peninsula and in Central Asia as suggested by Sharbel et al. (2000) (Figs. 4, S3). However, observed patterns of genome-wide LD are not consistent with such an explanation, because LD is higher in the putative refugia and lowest in Central Europe. Under the above model, the opposite is expected, namely a high LD in admixed populations and reduced LD in refugia. Genetic variation among Central Asian accessions is 2–2.5 times lower when compared to European and Iberian accessions and there are genetically similar accessions in distant (> 1,000 km) locations of Central Asia (Fig. 2b) which suggests a recent and rapid colonization by a small number of migrants. A similar pattern was obtained with a set of accessions from northern Sweden and Finland, at the northern edge of the distribution range (Nordborg et al. 2005). Because Central European populations are more polymorphic, there may have been more opportunities for a decay of LD than in the less diverse Central Asian populations.

Implications for genetic mapping

Naturally occurring genetic variation is a useful resource for the genetic mapping of complex phenotypic traits (Alonso-Blanco and Koornneef 2000). Population-based mapping approaches by means of LD mapping or association studies are increasingly

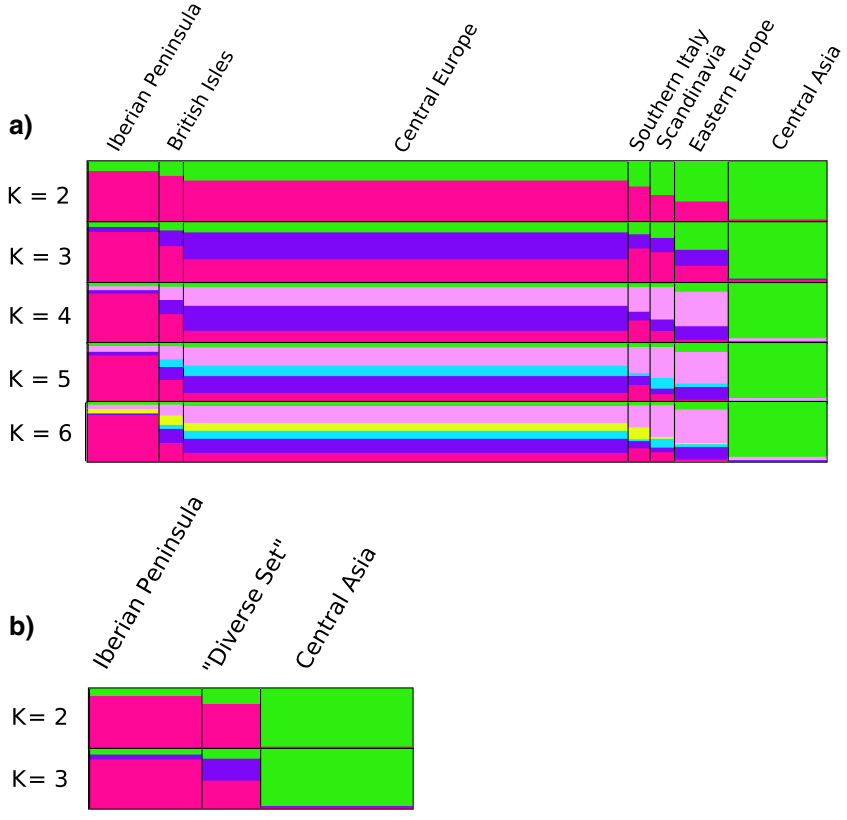

Fig. 4 Population structure as inferred with the `structure` program (version 2.0) with numbers of clusters ($K = 2,3,...$). Each individual is assigned proportionally to one of the clusters. The proportions are indicated by the relative lengths of different colors, representing K clusters. To highlight differences between geographic regions, the assignment to clusters is shown as an average proportion for the every geographic region. **a** Analysis of 115 SNPs in 335 accessions and **b** 9 genome sequence tags (GSTs) from a total of 66 accessions

being used to investigate the genetic basis of traits such as flowering time variation (Olsen et al. 2004; Stinchcombe et al. 2004; Caicedo et al. 2004; Shindo et al. 2005; Lempe et al. 2005). Association studies need to control for population stratification to avoid spurious associations between markers and phenotype. Therefore, it is necessary to determine whether a population structure is present and how a suitable mapping population of unrelated individuals with a similar genetic distance can be identified. In *A. thaliana*, individuals from the same local population can be genetically different indicating that they originated from multiple source populations, whereas geographically distant accessions can be highly similar (Table S1). For this reason, it may be worth considering to derive populations for association studies from genetically homogenous geographic regions such as Central Asia, or define them genetically rather than geographically by using genome-wide markers and clustering programs like `structure`. Such an approach was taken by Caicedo et al. (2004), who identified a set of 95 accessions using the AFLP data of Sharbel et al. (2000), to investigate epistasis among flowering time genes. These accessions are mostly of Central European origin and do not show any substructure suggesting they represent an unstratified population suitable for association studies. However, a significant clustering among Central European accessions was observed by Nordborg et al. (2005) and in this study (Figs. 4, S3), which indicates that a reliable inference of genetic relationships and population structure in *A. thaliana* requires large sets of genetic markers to avoid spurious associations between traits and markers.

Acknowledgements This work was funded by the German Ministry of Science (BMBF) as part of the GABI project (#0312275A) to T. A. and by the Emmy-Noether program of the Deutsche Forschungsgemeinschaft (Schm 1354-2/2) to K. J. S. We are grateful to Henriette Ringys-Beckstein, Maik Zehnsdorf and Melanie Lück for excellent technical assistance. We also thank K. Bachmann, M. Clauss, B. Haubold, M. Koornneef, A. Lawton-Rauh, T. Mitchell-Olds, S. Ramos-Onsins and E. Wheeler for discussions and comments on an earlier version of the manuscript.

References

Abbott RJ, Gomes MF (1989) Population genetic structure and outcrossing rate of *Arabidopsis thaliana*. Heredity 42:411–418

Akey J, Zhang K, Xiong M, Jin L (2003) The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. Mol Biol Evol 20:232–242

Alonso-Blanco C, Koornneef M (2000) Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. Trends Plant Sci 5:22–29

Bergelson J, Stahl E, Dudeck S, Kreitman M (1998) Genetic variation between and within populations of *Arabidopsis thaliana*. Genetics 148:1311–1323

Borevitz J, Nordborg M (2003) The impact of genomics on the study of natural variation in *Arabidopsis*. Plant Phys 132:718–725

Bray M, Boerwinkle E, Doris P (2001) High-throughput multiplex SNP genotyping with MALDI-Tof mass spectrometry: practice, problems and promise. Hum Mutat 17:296–304

Brumfield R, Beerli P, Nickerson D, Edwards S (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249–256

Caicedo A, Stinchcombe J, Olsen K, Schmitt J, Purugganan M (2004) Epistatic interaction between *Arabidopsis FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci USA 101:15670–15675

Comes H, Kadereit J (1998) The effect of quaternary climatic changes on plant distribution and evolution. Trends Plant Sci 3:432–438

Dixon P (2001) The Bootstrap and the Jackknife. In: Scheiner S, Gurevich J (eds) Design and analysis of ecological experiments. Oxford University Press, Oxford, pp 267–288

Eberle M, Kruglyak L (2000) An analysis of strategies for discovery of single-nucleotide polymorphisms. Genet Epidemiol 19:S29–S35

Environmental Systems Research Institute R Inc (1992) Arc/Info. Environmental Systems Research Institute, Red lands

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software `structure`: a simulation study. Mol Ecol 14:2611–2620

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

Felsenstein J (1989) PHYLIP—Phylogeny Inference Package, Version 3.2. Cladistics 5:164–166

Frenzel B, Pécsi M, Velichko A (1992) Atlas of paleoclimates and paleoenvironments of the northern hemisphere. Gustav Fischer, Stuttgart

Hewitt G (1999) Post-glacial re-colonization of European biota. Biol J Linn Soc 68:87–112

Hill W, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–331

Hoffmann M (2002) Biogeography of *Arabidopsis thaliana* (L.) *Heynh. (Brassicaceae)* J Biogeogr 29:125–134

Jander G, Norris S, Rounsley S, Bush D, Levin I, Last R (2002) *Arabidopsis* map-based cloning in the post-genome era. Plant Phys 129:440–450

Koornneef M, Alonso-Blanco C, Vreugdenhil D (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. Ann Rev Plant Biol 55:141–172

Kuhner M, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics 156:439–447

Kuittinen H, Mattilan A, Savoulainen O (1997) Genetic variation at marker loci and in quantitative traits in natural populations of *Arabidopsis thaliana*. Heredity 79:144–152

Lempe J, Balasubramanian S, Sureshkumar S, Singh A, Schmid M, Weigel D (2005) Diversity of flowering responses in wild *Arabidopsis thaliana* strains. PloS Genet 1:e6

Lewontin R (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67

Loridon K, Cournoyer B, Goubely C, Depeiges A, Picard G (1998) Length polymorphism and allele structure of trinucleotide microsatellites in natural accessions of *Arabidopsis thaliana*. Theor Appl Genet 97:591–604

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Mitchell-Olds T (2001) *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. Trends Ecol Evol 16:693–700

Miyashita NT, Kawabe A, Innan H (1999) DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified random fragment length polymorphism analysis. Genetics 152:1723–1731

Miyashita NT, Kawabe A, Innan H, Terauchi R (1998) Intra- and interspecific DNA variation and codon bias of the alcohol dehydrogenase (*Adh*) locus in *Arabis* and *Arabidopsis* species. Mol Biol Evol 15:1420–1429

Morin P, Luikart G, Wayne R, the SNP workshop group (2004) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19:208–216

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nielsen R, Hubisz M, Clark A (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 32:3435–3445

Nordborg M, Borevitz J, Bergelson J, Berry C, Chory J, Hagenblad J, Kreitman M, Maloof J, Noyes T, Oefner P, Stahl E, Weigel D (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:190–193

Nordborg M, Hu T, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg N, Shah C, Wall J, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005) The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:e196

Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD (2004) Linkage disequilibrium mapping of *Arabidopsis CRY2* flowering time alleles. Genetics 167:1361–1369

Pritchard J, Rosenberg N (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

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

Rosenberg N (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Schmid K, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005) A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 169:1601–1615

Schmid K, Rosleff-Sörensen T, Stracke R, Törjek O, Altmann T, Mitchell-Olds T, Weisshaar B (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res 13:1250–1257

Schmuths H, Hoffmann M, Bachmann K (2004) Geographic distribution and recombination of genomic fragments on the short arm of chromosome 2 of *Arabidopsis thaliana*. Plant Biol 6:128–139

Sharbel T, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol Ecol 9:2109–2118

Shindo C, Aranzana M, Lister C, Baxter C, Nicholls C, Nordborg M, Dean C (2005) Role of *FRIGIDA* and *LOWERING LOCUS C* in determining variation in flowering time of *Arabidopsis*. Plant Phys 138:1163–1173

Sokal RR, Rohlf FJ (1995) Biometry. Sinauer Associates, Sunderland

Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, Halldorsdottir SS, Purugganan MD, Schmitt J (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. Proc Natl Acad Sci USA 101:4712–4717

Törjék O, Meyer R, Müssig C, Schmid K, Weisshaar B, Mitchell-Olds T, Altmann T (2003) Establishment of a high-efficiency SNP-base framework marker set for *Arabidopsis*. Plant J 36:122–140

Vander Zwan C, Brodie S, Campanella J (2000) The intraspecific phylogenetics of *Arabidopsis thaliana* in worldwide populations. Syst Bot 25:47–59

Weir B (1996) Genetic data analysis. II. Sinauer Associates, Sunderland

Wright S, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506–519