Molecular Ecology (2004) **13**, 3403–3413 doi: 10.1111/j.1365-294X.2004.02329.x

Neutral genetic variation among wild North American populations of the weedy plant *Arabidopsis thaliana* is not geographically structured

S. JØRGENSEN* and R. MAURICIO
Department of Genetics, Davison Life Sciences Complex, University of Georgia, Athens, GA 30602–7223, USA

Abstract

We investigated neutral genetic variation within and among 53 wild-collected populations of the weedy annual plant, *Arabidopsis thaliana*, in North America, using amplified fragment length polymorphism (AFLP) markers. *A. thaliana* is thought to have been introduced to North America from Eurasia by humans; such an introduction might be expected to leave a clear geographical signal in the genetic data. To detect such patterns, we sampled populations at several hierarchical geographical levels. We collected individuals from populations in two areas of the Southeast and one in the Midwest, as well as individuals from populations in the Pacific Northwest and Northeast. To estimate within-population variation, we sampled eight individuals from each of six populations in the Southeast and Midwest. Among all 95 individuals analysed, we detected 131 polymorphic AFLP fragments. We found no evidence for continental or regional diversification. Individuals sampled from Midwestern and Southeastern populations intermingled in a neighbour-joining tree, and Mantel tests conducted within the Midwestern and Southeastern regions as well as the full data set failed to detect any significant relationship between geographical and genetic distance. These results mirror those found for most global surveys of neutral genetic variability in *A. thaliana*. Surprisingly, we detected substantial amounts of neutral genetic variability within populations. The levels of genetic variation within populations, coupled with the nongeographical nature of divergence among populations, are consistent with contemporary gene flow and point to a complex and dynamic population history of *A. thaliana* in North America.

Keywords: AFLP, *Arabidopsis thaliana*, biogeography, diversity, genetic variability, population structure

Received 13 May 2004; revision received 5 July 2004; accepted 23 July 2004

Introduction

The mouse-ear cress, *Arabidopsis thaliana* (L.) Heynh., is a small, predominantly selfing, annual plant in the plant family Brassicaceae with a circumboreal distribution generally found in disturbed habitats (Lawrence 1976; Rollins 1993). A model organism in developmental and molecular genetics, limited knowledge of *A. thaliana* in its natural context has not prevented the species' adoption as a model system for ecological genetic studies (e.g. Mauricio & Rausher 1997).

An understanding of the amount and distribution of neutral genetic diversity in natural populations is fundamental to properly designing and interpreting the results of ecological genetic studies. This information is essential because the amount of neutral genetic variability found within and among populations will both influence the response to, and be influenced by, selective pressures and evolutionary processes such as gene flow and genetic drift (Hamrick & Godt 1996a; Futuyma 1998). For example, low levels of genetic variability could limit the response to natural selection (Fisher 1930). In fact, a number of studies have characterized the structure of neutral genetic variation

Correspondence: Rodney Mauricio. Fax: 706 542 3910; E-mail: mauricio@uga.edu
*Present address: Department of Geography, University of Hawai'i at Mānoa, Honolulu, Hawai'i 96822, USA

within *A. thaliana* at a global scale (King *et al*. 1993; Innan *et al*. 1997; Bergelson *et al*. 1998; Breyne *et al*. 1999; Miyashita *et al*. 1999; Erschadi *et al*. 2000; Sharbel *et al*. 2000; Hoffman & Schmuths 2001; Barth *et al*. 2002). One of the most surprising, yet consistent, results of these surveys is the general lack of any strong geographical patterning to the variation among sampled accessions.

Populations from North America have been poorly represented in global surveys of neutral genetic diversity and no comprehensive study of the relationships among populations on the continent exists. Thought to be native to eastern Europe and western Asia (Ratcliffe 1961; Lawrence 1976; Price *et al*. 1994; O'Kane & Al-Shehbaz 1997), the species' widespread presence in North America is often attributed to human activities such as the European exploration and colonization of the continent (e.g. Sharbel *et al*. 2000; Schmuths *et al*. 2004). Because the presence of *A. thaliana* in North America is believed to be relatively recent, the geographical and genetic relationships among populations may differ substantially from those sampled at the global scale and from the species' presumed native range. Populations in North America may additionally exhibit little neutral genetic variation due to founder effects or bottlenecks that might have occurred with colonization.

An additional limitation of earlier studies has been a reliance on material from seed stock centres where the provenance of collections is often obscure. In the *Arabidopsis* community, field-collected plants are often called 'ecotypes', borrowing the classic term of Turesson (1922, 1925) but not strictly his definition. For most of these 'ecotypes' a single individual collected from a single site has been deposited in either the *Arabidopsis* Biological Resource Center at The Ohio State University (Columbus, OH, USA) or the Nottingham Arabidopsis Stock Centre at the University of Nottingham (Loughborough, UK). These seeds are planted and grown to fruiting and all the resulting seeds are either pooled or seeds are separated by individual plant (single seed descent lines). Strong selection, although inadvertent, is particularly likely for such traits as early seed germination and flowering. It is only relatively recently that multiple individuals from a number of sites have been deposited in the stock centres.

In this study we used amplified fragment length polymorphism (AFLP) markers (Vos *et al*. 1995) to investigate the patterns of neutral genetic variability in wild-collected populations of *A. thaliana* in North America. We also used historical records to attempt to reconstruct the history of the introduction of *A. thaliana* to North America. Our primary objectives were to (i) determine whether there are geographical patterns to neutral genetic diversity among populations in North America, and (ii) provide a preliminary estimate of within-population neutral genetic variation in six populations from two different regions within North America.

Materials and methods

Plant material

All plant material used in this study was collected by the authors. A population was defined as plants growing within a discrete location, often a single agricultural field or isolated cemetery. The physical boundaries of each collection site were well delineated, and most sites were separated by at least 1 km. Seeds were collected from mature fruits of wild-growing individuals of *Arabidopsis thaliana* in 53 natural populations in North America (Table 1). Plants were collected along a linear transect with a minimum spacing of at least 2 m between sampled individuals to lessen the likelihood that we were collecting individuals from the same maternal plant. Populations were located primarily in the Midwestern United States, near Lake Michigan, and in the Southeastern United States in the Piedmont region of Georgia and North Carolina. Two populations from the Pacific Northwest and a single population from the Northeast, in central New York State, were also included.

A single maternal individual was chosen from each of 47 populations for genetic analysis, as earlier-published studies had indicated that most of the neutral genetic variation in *A. thaliana* occurred among, rather than within, populations (Todokoro *et al*. 1995; Bergelson *et al*. 1998). However, to provide confirmation of this result, we chose to sample

Table 1 Locations of North American populations of *Arabidopsis thaliana* surveyed for this study, listed by geographical region

Population locations
Midwest
Berrien County, MI: MI1, MI2, MI3, MI4, MI5, MI6
La Porte County, IN: MI7, MI8, MI13
Jasper County, IN: MI10
St Joseph's County, IN: MI9, MI12, MI14
Starke County, IN: MI11
Southeast
Carteret County, NC: NC8, NC9, NC10
Durham County, NC: NC1, NC2, NC3, NC4, NC5, NC6, NC7, NC15, NC16, NC17
Person County, NC: NC12, NC13, NC14
Clarke County, GA: GA7
Madison County, GA: GA8, GA9, GA13, GA14
Oconee County, GA: GA1, GA2, GA3, GA6, GA15, GA16, GA17
Oglethorpe County, GA: GA10, GA11, GA12
Walton County, GA: GA4, GA5, GA18, GA19, GA20
Pacific Northwest
Vancouver, BC: PN1
King County, WA: PN2
Northeast
Tompkins County, NY: NY1

multiple individuals from a few randomly chosen populations. In these cases, eight maternal individuals were analysed from each of six additional populations, three each from the Midwest (MI1, MI4, MI13) and Southeast (GA5, GA11, GA18).

One seed from each field-collected maternal individual to be analysed was sown in an approximately 5 × 5 × 6-cm plastic pot filled with a soilless mix of peat moss, perlite, pine bark and vermiculite (Fafard no. 3B, Agawam, MA). The seeds were cold-stratified at 4 °C for 3 days then transferred to a single growth chamber with control for both daylength (14 h) and temperature (18 °C). Fresh leaf tissue was collected from young leaves and immediately snap-frozen in liquid nitrogen. We extracted genomic DNA using a standard phenol–chloroform protocol (Colosi & Schaal 1993; Bergelson *et al*. 1998).

AFLP analysis

Analysis of AFLP was performed with commercially available supplies from Applied Biosystems (ABI, Foster City, CA) for fluorescent marker detection. Restriction-ligation was conducted using the enzymes *Eco*RI and *Mse*I and enzyme-specific ligators supplied in ABI's Preselective Amplification kit (P/N 402004). Following overnight incubation at room temperature, the restriction-ligation reactions were diluted in Tris-EDTA buffer for preselective amplification. The preselective product was diluted to serve as the template for the subsequent rounds of selective amplifications. Three primer combinations were chosen for selective amplification (*Eco*RI ACT × *Mse*I CTT; *Eco*RI AGG × *Mse*I CAG; *Eco*RI AGC × *Mse*I CAG). For genotyping, 2 μL of the selective amplification polymerase chain reaction (PCR) product was combined with 7.4 μL of formamide and 0.6 μL of a ROX-500 internal size standard. The mixture was denatured at 95 °C for 4 min and snap-cooled before being run on an ABI Prism 3700 DNA Analyser. Raw data were collected using version 3.1 of ABI's GENESCAN® software, which automatically determines fragment size by interpolation to the internal size standard.

Each differently sized fragment identified was scored as either present or absent in each individual analysed. Thus, each AFLP marker was treated as dominant, the absence of a band being presumably recessive. We found no instances of codominance (i.e. two bands that appear as alternatives and are putatively allelic). In a predominantly selfing species such as *Arabidopsis thaliana*, heterozygotes should be rare (Bergelson *et al*. 1998) and the use of a dominant marker is unlikely to be a significant handicap (Vekemans *et al*. 2002). While the dominant expression of AFLP markers results in less per-locus information content compared to codominant markers (Vekemans *et al*. 2002), the large number of loci analysed tends to mitigate this problem (Parsons & Shaw 2001).

Data analysis

We tested for linkage disequilibrium (LD) among loci by comparing the observed variance (V_D) in the number of mismatches in the data set to the variance expected (V_E) under linkage equilibrium for each possible pair of haplotypes. We used the program LIAN (Haubold & Hudson 2000), which uses Monte Carlo simulations to compute the standardized index of association, as a measure of linkage disequilibrium, $I_A^S = (V_D/V_E - 1)/(r-1)$, where r is the number of loci.

Relationships among samples were estimated by a neighbour-joining analysis (Saitou & Nei 1987). Dice's (1945) coefficient of similarity was calculated between each pair of individuals: $S_{ij} = 2a/(2a + b + c)$, where a is the number of fragments shared between individuals, b is the number of fragments present in individual i only and c is the number of fragments present in individual j only. Similarity values near the maximum value of 1 indicate the samples share most fragments, while values near 0 indicate the samples have few fragments in common. The Dice coefficient was chosen because shared absences, which are more likely to be nonhomologous (Wolfe & Liston 1998), are excluded from the calculation. Similarities were calculated with version 8.0 of the program SPSS® (SPSS Inc. 1989–2002) and converted to distances by subtracting from 1. The neighbour-joining trees were constructed from 1000 bootstrap replicates using the program PAUP* version 4.1 (Swofford 2002).

We tested for a relationship between geographical distance and Dice's measure of genetic similarity using Mantel tests (Mantel 1967). A single individual was chosen randomly from each of the six intensively sampled populations to be included in these analyses. We tested for isolation by distance for the entire data set as well as for samples from the Midwestern and Southeastern regions separately. The Mantel analyses were conducted using the program NT-SYSPC (Rohlf 2002).

Neutral genetic diversity within the intensively sampled populations was assessed at the population, regional and species levels by estimating the percentage of polymorphic loci (*PL*) and the expected heterozygosity (H_E). For these calculations each fragment was considered a single, unique locus exhibiting Mendelian segregation of polymorphic fragments with a single dominant (amplified) and recessive (null) allele at each locus (Travis *et al*. 1996). A locus was considered polymorphic if the associated band was present in one or more, but not all, individuals of the population.

Given the dominant nature of AFLP loci, the calculation of H_E required an estimation of inbreeding within populations. Failure to account for inbreeding will result in underestimating H_E (Clark 1997). When information about fixation is lacking, random mating (inbreeding coefficient

$f = 0$) is normally assumed (e.g. Travis *et al*. 1996). However, *A. thaliana* is thought to have a selfing rate of greater than 99% (Redei 1975; Abbott & Gomes 1989), corresponding to a high equilibrium fixation rate of f. Thus, the high selfing rate of *A. thaliana* also mitigates the problems with calculating expected heterozygosity with AFLP (Vekemans *et al*. 2002). However, because the actual selfing rate for these particular populations is unknown, we calculated all neutral genetic diversity measures using both the expected fixation rate under random mating ($f = 0$) and a high level of fixation expected given the rate of selfing ($f = 0.98$).

Total expected heterozygosity was calculated as the mean expected heterozygosity among loci. Because of the binary nature of AFLP data, H_E assumes a maximum value of 0.50, half the possible maximum for loci with an infinite number of equally frequent alleles. Neutral genetic structure among populations was determined by Nei's (1973) measures of genetic diversity, which include total genetic diversity (i.e. expected heterozygosity) (H_T), mean genetic diversity within species (H_S) and the proportion of genetic diversity occurring among species, $G_{ST} = (H_T - H_S)/H_T$. A procedure analogous to the calculation of H_E was used to estimate H_T and H_S. All genetic diversity parameters were estimated using version 1.32 of the program POPGENE (Yeh *et al*. 1999).

Divergence among the six intensively sampled populations was also assessed by an analysis of molecular variance (AMOVA, Excoffier *et al*. 1992). In this analysis, the total variance in the AFLP data set is partitioned into regional, among-population and within-population components. Paired genetic distances among the six populations were estimated using Φ_{ST}. As with its analogue, F_{ST} (Wright 1951), values of Φ_{ST} range from 0, indicating no differences between populations, to 1, which indicates no shared fragments between the compared populations. Unlike the calculation of F-statistics, including Nei's G_{ST}, the AMOVA analyses partitioned variation according to correlations among genotypes rather than variation in gene frequencies. This difference is because of the dominant expression of AFLP markers. The statistical significance of the variance components of the AMOVA and the paired comparisons were determined by nonparametric procedures using 999 random permutations. These analyses were performed with version 5.1 of the software package, GENALEX (Peakall & Smouse 2001).

Historical records of A. thaliana *in North America*

We obtained the collection records for specimens of *A. thaliana* from four herbaria with the earliest US collections: the Academy of Natural Sciences, Philadelphia (ANS), the Harvard University Herbaria (HUH), the New York Botanical Garden Herbarium (NYBG) and the Smithsonian Institution (SI). We also used online databases to search for records at other herbaria.

Results

We identified 131 polymorphic fragments from the 95 individuals sampled. There was little variation among primer-pairs in the number of fragments produced: ACT-CTT (41), AGG-CAG (42) and AGC-CAG (48). A large percentage (60%) of the polymorphic fragments occurred at frequencies less than 0.10, and none were observed at frequencies greater than 0.75. Approximately 15% of the fragments occurred at intermediate frequencies (0.30–0.70). When individuals from the intensively sampled populations were excluded only 109 fragments were present, all of which were polymorphic. Within the Southeast region, 99 of the 109 (91%) fragments were polymorphic and within the Midwest, 69 loci (63%) were polymorphic.

Relationships among individuals

Individuals did not group strictly according to geographical region in the neighbour joining analysis (Fig. 1). Individuals from populations sampled near Lake Michigan were located in a basal position on the tree and individuals tended to cluster by local geographical region at the tips of the tree. However, no broad, regional-level patterns were evident (Fig. 1). Clusters of individuals from Southeastern locales intermingled with clusters of individuals from Midwestern locations. Bootstrapping analysis indicated low support for many branches of the tree.

Variation within regions also did not appear to be geographically structured. When individuals from only the Southeast or only the Midwest were analysed, we still found that individuals from the intensively sampled populations did not produce discrete clusters. Mantel tests also failed to detect evidence of geographical structure at either the regional or continental scale. There was no evidence of isolation by distance for the full data set ($r = 0.134$, $t = 3.117$, $P = 0.99$) or within either the Midwestern ($r = -0.243$, $t = -1.480$, $P = 0.07$) or Southeastern regions ($r = 0.228$, $t = 4.246$, $P = 0.99$).

Significant linkage disequilibrium was detected within the full data set and in each of the two regions (Table 2). Because of significant disequilibrium among loci, relationships among individuals might become clearer from a principal components analysis (PCA), a technique based

Fig. 1 Neighbour-joining tree of individuals of *A. thaliana* determined from 131 polymorphic AFLP fragments, listed by abbreviations given in Table 1. For intensively sampled populations, sample number follows population identification. Included are those groups that occurred > 50% of the time (i.e. 50% majority-rule consensus); bootstrap values, determined from 999 permutations, are listed on each branch.

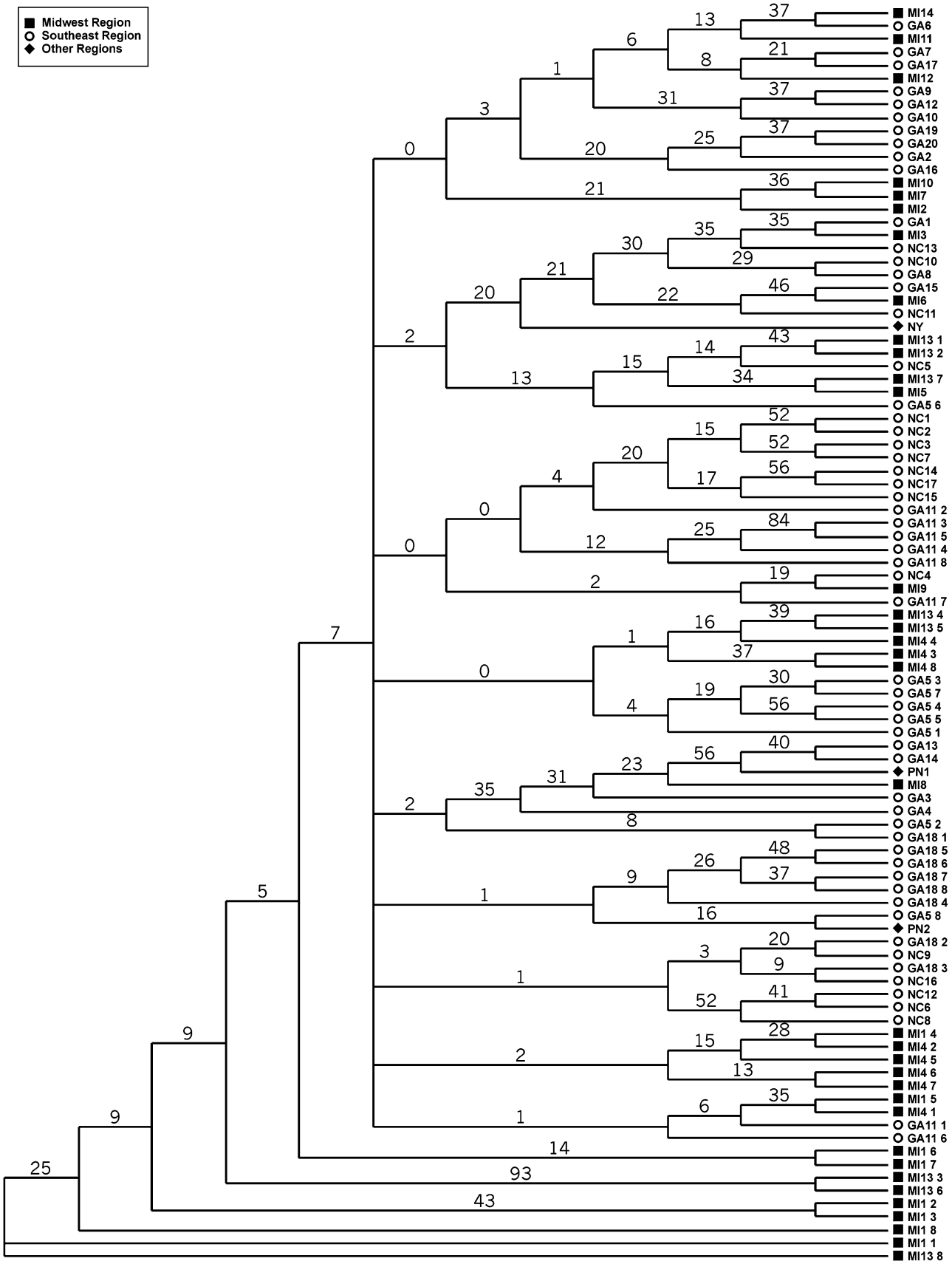
Midwest Region
Southeast Region
Other Regions
MI14
GA6
MI11
GA7
GA17
MI12
GA9
GA12
GA10
GA19
GA20
GA2
GA16
MI10
MI7
MI2
GA1
MI3
NC13
NC10
GA8
GA15
MI6
NC11
NY
MI13 1
MI13 2
NC5
MI13 7
MI5
GA5 6
NC1
NC2
NC3
NC7
NC14
NC17
NC15
GA11 2
GA11 3
GA11 5
GA11 4
GA11 8
NC4
MI9
GA11 7
MI13 4
MI13 5
MI4 4
MI4 3
MI4 8
GA5 3
GA5 7
GA5 4
GA5 5
GA5 1
GA13
GA14
PN1
MI8
GA3
GA4
GA5 2
GA18 1
GA18 5
GA18 6
GA18 7
GA18 8
GA18 4
GA5 8
PN2
GA18 2
NC9
GA18 3
NC16
NC12
NC6
NC8
MI1 4
MI4 2
MI4 5
MI4 6
MI4 7
MI1 5
MI4 1
GA11 1
GA11 6
MI1 6
MI1 7
MI13 3
MI13 6
MI1 2
MI1 3
MI1 8
MI1 1
MI13 8

Table 2 Linkage disequilibrium in *A. thaliana* calculated from 131 polymorphic AFLP loci. Only those samples for which no missing data were used, and regional estimates were calculated from individuals in the intensively sampled populations. Significance was determined from 100 Monte Carlo simulations

Geographic region	n	V_D	V_E	I_A^S	P
All samples	84	54.679	17.418	0.017	≤ 0.01
Midwest	22	39.102	15.013	0.012	≤ 0.01
Southeast	24	38.346	14.559	0.013	≤ 0.01

on correlations or nonindependent associations among variables. We performed such a PCA analysis but we failed to detect any discrete clusters, and no geographical patterning was discernable.

Genetic diversity within and among populations

One hundred and ten (84%) loci were polymorphic among plants in the six intensively sampled populations (Table 3). The mean percentage of polymorphic loci within populations was 40.2%, and varied little among the populations or regions. Among populations from the Midwestern region, *PL* varied from 35.9 to 42.8%, with a pooled regional value of 69.5%. Values ranged from 35.1 to 45.0% among the populations sampled from the Southeastern region, which had a pooled *PL* of 67.9%. Species-level expected heterozygosity (assuming $f = 0.98$) was 0.179, and was also relatively uniform among the two regions. Expected heterozygosity within Midwestern populations varied from 0.120 for population MI4 to 0.150 for population MI13. In the Southeast, values ranged from 0.122 for population GA11 to 0.139 for population GA18.

At the species level, the neutral genetic structure detected among all six populations was 0.279 (Table 4). When the populations were pooled within regions, substantially less structure was detected ($G_{ST} = 0.069$). These values indicate that only about 25% of the differences among populations are due to regional differences. Instead, a majority of the variation among populations is due to variation within regions. Similar levels of neutral genetic structure were detected among populations from the Midwestern ($G_{ST} = 0.218$) and Southeastern regions ($G_{ST} = 0.234$).

The AMOVA analysis indicated that approximately 77% of the variation in the data set was from genotypic variation within populations (Table 5). Only 3% of the variation could be attributed to regional differences, and the final 20% was due to differences among populations within regions. Despite differences in the proportion of the total variance, values for all three hierarchical levels were significant. Pairwise determinations of Φ_{ST} varied from 0.072 to 0.303. In all cases, the differentiation between populations was significant (Table 6). The mean value for comparisons among the Midwestern populations was 0.199; for the Southeastern populations, the mean value was 0.216. Comparisons between populations from different regions produced values from 0.072 to 0.303 with a mean of 0.230.

Table 3 The percentage of polymorphic loci (*PL*) and expected heterozygosity (H_E) detected within six populations of *A. thaliana*. Heterozygosity was calculated assuming both random mating ($f = 0$) and a selfing rate of 0.99 ($f = 0.98$). Standard errors are given in parentheses

	H_E		
Population	*PL*	$f = 0$	$f = 0.98$
Midwest			
MI1	39.7	0.096 (0.051)	0.122 (0.059)
MI4	35.9	0.105 (0.059)	0.120 (0.062)
MI13	42.8	0.111 (0.054)	0.150 (0.067)
Pooled Midwest	69.5	0.125 (0.031)	0.167 (0.034)
Southeast			
GA5	45.0	0.091 (0.047)	0.122 (0.053)
GA11	35.1	0.103 (0.059)	0.122 (0.064)
GA18	42.8	0.119 (0.060)	0.139 (0.063)
Pooled Southeast	67.9	0.131 (0.032)	0.166 (0.036)
Species	84.0	0.135 (0.022)	0.179 (0.023)

Table 4 Nei's (1973) measures of neutral genetic diversity and structure calculated for six intensively sampled populations of *A. thaliana*. Values were calculated assuming both random mating ($f = 0$) and a selfing rate of 0.99 ($f = 0.98$)

	$f = 0$			$f = 0.98$		
Analysis	H_T	H_S	G_{ST}	H_T	H_S	G_{ST}
Species level						
As six populations	0.135	0.104	0.229	0.179	0.129	0.279
As two regions	0.135	0.128	0.050	0.179	0.167	0.069
Regional level						
Midwest	0.126	0.104	0.172	0.167	0.130	0.218
Southeast	0.131	0.104	0.203	0.167	0.127	0.234

History of A. thaliana *in North America*

The earliest dated specimens of *A. thaliana* were collected in Pennsylvania in 1843 (ANS), although there exist potentially earlier, undated collections (Mauricio & Rausher 1997). Other specimens in the ANS collection include samples from Pennsylvania (1850), Kentucky (1850), Texas (1860) and New York (1864). Other localities with samples from the 1860s include samples from Massachusetts (HUH, 1861), Delaware (NYBG, 1863) and Michigan (NYBG, 1868). Samples of *A. thaliana* collected in the late 19th century include

Table 5 Hierarchical analysis of molecular variance (AMOVA) for six populations of *A. thaliana*. The significance value was determined from a permutation test

Source of variation	d.f.	Sums of squares	Variance component	% Total variation	*P*-value
Six populations					
Among populations	5	155.35	2.792	22.39	0.001
Within populations	40	387.11	9.678	77.61	0.001
Regional level					
Between regions	1	38.70	0.414	3.28	0.020
Among populations	4	116.65	2.544	20.13	0.001
Within populations	40	387.11	9.678	76.59	0.001

Table 6 Pairwise differentiation among six populations of *A. thaliana*. Φ_{ST} values are below the diagonal. Probability values are based on 999 permutations

	1	2	3	4	5
1 MI1					
2 MI4	0.298**				
3 MI13	0.113**	0.186**			
4 GA5	0.240***	0.185**	0.072*		
5 GA11	0.287***	0.246***	0.303***	0.277***	
6 GA18	0.245***	0.241***	0.250***	0.180**	0.191***

*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

Washington, DC (NYBG, 1877), New Jersey (NYBG, 1878), North Carolina (HUH, 1874), West Virginia (SI, 1878), Arkansas (SI, 1887), Tennessee (SI, 1889), Connecticut (SI, 1896) and Wisconsin (SI, 1896). After 1896 collections in the Smithsonian Institution Herbarium become more numerous, with multiple collections per year. Records from the Jepson Herbarium indicate the presence of *A. thaliana* in California in 1892, and the University of British Columbia Herbarium has its first record of *A. thaliana* in British Columbia by 1915.

Discussion

Global surveys of neutral genetic diversity in *A. thaliana* indicate few broad-scale geographical patterns of diversification (King *et al*. 1993; Innan *et al*. 1997; Bergelson *et al*. 1998; Breyne *et al*. 1999; Miyashita *et al*. 1999; Erschadi *et al*. 2000). Our study focused exclusively on North America, where this species is thought to be introduced. Anthropogenic dispersal from Europe to eastern North America, with subsequent expansion to the Midwest and West, might lead us to expect a stronger geographical signature to the pattern of neutral genetic diversity than found globally. None the less, we found no evidence to support any strong geographical structure to our genetic data from wild North American populations of *A. thaliana*.

That is not to say that there is no evidence in the literature for an association between geographical and genetic structure, particularly with respect to European vs. Asian accessions (Sharbel *et al*. 2000). However, even when a geographical signal has been detected, it has not been strong (Sharbel *et al*. 2000; Barth *et al*. 2002; Schmuths *et al*. 2004). For example, Sharbel *et al*. (2000) detected significant, but weak, isolation by distance among populations sampled from the presumed native range of *A. thaliana* in Eurasia. Barth *et al*. (2002), who examined individuals from Africa, Eurasia and North America with cleaved amplified polymorphic sequences (CAPS) and intersimple sequence repeats (ISSR) data, also found evidence for weak geographical structure: neither of their marker data sets revealed a significant relationship between geographical and genetic distance, but a multivariate analysis of their CAPS data did indicate some clustering of groups according to geographical origin (Barth *et al*. 2002).

When included in global surveys, North American accessions tend to be embedded within and interspersed among populations from Eurasia (Innan *et al*. 1997; Bergelson *et al*. 1998; Miyashita *et al*. 1999; Barth *et al*. 2002). This pattern has generally been interpreted as consistent with the presumed non-native status of *A. thaliana* in North America. However, the circumboreal distribution of *A. thaliana* is a pattern common to many species in the Brassicaceae (Rollins 1993), and the observed relationships of North American populations to Eurasian populations could also have resulted from a prehistorical, nonanthropogenic introduction to North American. Whether the populations in the eastern and western portions of North America are the result of a single or multiple introductions to the continent is also unclear. Samples for the Pacific Northwest were only weakly represented in our study (two individuals). However, their relationships to samples from eastern North America were not consistent: instead of clustering together, they appeared in two different clusters with individuals from Georgia. This pattern could indicate that populations in the Pacific Northwest are derived from eastern populations, that multiple introductions occurred in both regions with some populations being founded by the same source, or that contemporary gene flow occurs between populations in eastern and western North America.

Regional patterns of neutral genetic diversity in North America

The neighbour-joining analysis indicated little regional or subregional diversification in eastern North America, mirroring patterns found globally. Although individuals from the intensively sampled populations collected from Berrien County, Michigan (MI1, MI4) and La Porte County, Indiana (MI13) occupied the most basal positions on the neighbour-joining tree, in general populations from the Midwest and Southeast were intermingled. There is some evidence for local substructuring of *A. thaliana*, because the smallest groups in the neighbour-joining tree tended to include only individuals from either the Midwestern or Southeastern region. However, these conclusions are tempered by low bootstrap values throughout the tree. Despite sampling more recently established populations from a smaller geographical region, the statistical support for our neighbour-joining analysis was only marginally better than that of Sharbel *et al*. (2000). Only a handful of bootstrap values in our analysis exceeded 50%: in about half the cases, they were observed for small clusters of individuals sampled from the same location. In most the remaining cases, bootstrap values in excess of 50% were associated with individuals sampled from different populations from within the Southeastern region. A single group contained individuals from two populations in Georgia as well as an individual sampled from the Pacific Northwest.

Patterns of neutral genetic variability in *A. thaliana* from published reports have been interpreted to indicate a rapid spread followed by little gene flow among populations (Bergelson *et al*. 1998; Miyashita *et al*. 1999; Erschadi *et al*. 2000; Barth *et al*. 2002). For example, Miyashita *et al*. (1999) analysed 38 globally distributed ecotypes and their neighbour-joining analysis produced a tree with a star-like topology, with each ecotype having a long branch. They interpreted these patterns to indicate a rapid diversification in the past with little migration between locations. Given the consistency of findings at multiple spatial scales in multiple locations, including North America, the recent past may mean only several hundred years ago, consistent with a role for humans in the spread and movement of individuals among populations of *A. thaliana* (Bergelson *et al*. 1998; Sharbel *et al*. 2000).

The historical herbaria records seem to support an anthropogenic dispersal of *A. thaliana* in North America consistent with recent colonization: plants tend to appear in herbaria collections earlier in the East and later in the Midwest and West. However, these data should be viewed with some caution. First, herbarium records from the US colonial period (pre1800s) were often not dated specifically. The dates of the birth and death of the collector, or the dates of the collection expedition, are sometimes the only information we have of the date of collection of the specimen. In addition, many early botanists were paid by European museums to collect new plants from the New World, and there could well have been a bias among both collectors and museums against collecting weedy plants known from Europe. Even if we accept the records at face value, there are enough contradictions to raise questions about strictly anthropogenic dispersion. For example, *Arabidopsis thaliana* appears in herbaria records in Texas in 1860, but in the eastern states of Tennessee and Connecticut in 1889 and 1896, respectively.

Although a rapid pulse of diversification followed by little among-population gene flow has been the favoured hypothesis for explaining patterns of neutral genetic variability in *A. thaliana*, other explanations do exist. One possibility is that these populations originated very recently from a mixed origin, and have not yet lost variants by genetic drift. Our finding of a lack of regional or subregional differentiation could be accounted for by this hypothesis. Contemporary seed dispersal of what are essentially well-established, nonrecombining haplotypes coupled with extinction/recolonization could also generate the patterns seen in *A. thaliana* (Bergelson *et al*. 1998). We have collected some data on the persistence of wild populations of *A. thaliana* in the Southeast that suggests extensive extinction of local populations (RM, unpublished data).

Non-anthropogenic mechanisms may also facilitate gene flow among populations of *A. thaliana*. The reportedly high degree of selfing observed in natural populations (Abbott & Gomes 1989) suggests that seed dispersal is most likely to be the source of gene flow among populations. *A. thaliana*, like many annual weeds characteristic of disturbed habitats, should be successful at colonizing new locations (Griffith *et al*. 2004).

The primary seed dispersal mechanism of *A. thaliana* is thought to be gravity. The species possesses a dehiscent fruit that, when mature, breaks open, spilling the seeds. However, because the seeds are of extremely small size, long-distance dispersal may be an important, but underappreciated, mechanism mediating gene flow in *A. thaliana* (Cain *et al*. 2000). In particular, wind and convective updrafts could be a source of long-distance dispersal (Carlquist 1980; Tackenberg *et al*. 2003). Certainly, the presence of unique multilocus haplotypes in *A. thaliana* found distributed across the globe argues for either long-distance dispersal or rapid, worldwide, population expansion (Bergelson *et al*. 1998).

Uplift of *A. thaliana* seeds to the upper atmosphere could occur by strong thermals or in association with the uplift occurring along surface frontal boundaries, which usually coincides with the location of the polar jet stream in the upper atmosphere (Aguado & Burt 2003). Both violent storm activity (e.g. tornadoes, thunderstorms with strong updrafts) and undulations in the polar jet stream occur in the spring, precisely at the time that fruits of *A. thaliana* are

maturing and seeds are being released. If the seeds were lifted into the upper atmosphere by convective updrafts and reached the polar jet stream they could potentially be transported 100s of kilometres. In fact, wind connectivity, more so than geographical proximity, explains the bryophyte and pteridophyte floristic affinities of Southern Hemisphere extra-tropical land masses (Muñoz *et al*. 2004). Given the almost continuous ring of land occurring at high latitudes in the Northern Hemisphere, wind connectivity may be an even more important phenomenon than for southern landmasses. Thus, the jet stream could be a mechanism of north-to-south, south-to-north and west-to-east seed dispersal.

Genetic diversity within populations

The most striking and unexpected result of the present study was the presence of significant within population neutral genetic variation. Individuals from the six intensively sampled populations failed to cluster together in the neighbour-joining analysis. For example, three individuals from population MI4 cluster together with individuals from population MI13, while the other five individuals appear in a different cluster with individuals from population MI1. These patterns are similar to those of Bergelson *et al*. (1998), who found distinctly different haplotypes in two populations sampled from the same agricultural field in Ascot, UK and Erschadi *et al*. (2000) who also found distinct differences among populations sampled from the same location in Cologne, Germany.

Similar haplotypes were often distributed widely in North America. For example, the individuals from the Pacific Northwest populations appeared in two different small groups, each with individuals from Georgia. Clusters frequently included a mixture of individuals from both the Midwestern and Southeastern regions. Again, the lack of geographical patterns is consistent with that found at the global scale (Bergelson *et al*. 1998; Erschadi *et al*. 2000).

The substantial amount of genetic variability observed within the six intensively sampled populations in North America was largely unexpected given the reported high selfing rate of *A. thaliana* (Abbott & Gomes 1989) and the results of several other studies in which only low levels of variation within populations of *A. thaliana* were detected. These studies include those that did (Breyne *et al*. 1999) and did not (Konieczny & Ausubel 1993; Hanfstingl *et al*. 1994; Todokoro *et al*. 1995; Bergelson *et al*. 1998) use AFLP markers. Typically, selfing species have low levels of neutral genetic diversity within populations but substantial differentiation among populations (Allard *et al*. 1968; Hamrick & Godt 1996b).

The overall divergence among the six intensively sampled populations was, at $G_{ST} = 0.28$, less than that seen in other selfing species with gravity-dispersed seeds, which average $G_{ST} \sim 0.5$ (Hamrick & Godt 1996b). In fact, Bergelson *et al*. (1998), determined that approximately 64% of the genetic variation occurred among rather than within populations of 11 globally distributed populations of *A. thaliana*. Because of the different sampling methodologies (i.e. global vs. North America only), genetic (RFLP of protein coding regions vs. anonymous AFLP loci) and mathematical techniques used to describe genetic variation, direct comparisons with our study are difficult. To facilitate comparison, we reanalysed data from solely the North American populations of Bergelson *et al*. (1998) to produce estimates of genetic structure (i.e. G_{ST}) analogous to our calculations. Values of G_{ST} varied substantially: 0.20, 0.69 and 0.83 for the loci *Adh*, *Dhs1* and *Gpa1*, respectively, with a mean of 0.57. The large variation among loci in G_{ST} could indicate that genetic variation at the loci sampled by Bergelson *et al*. (1998) is not neutral (see Hanfstingl *et al*. 1994). Furthermore, because nucleotide diversity at the loci sampled by Bergelson *et al*. (1998) was unusually low, their estimates of neutral genetic structure may be inflated.

Like the G_{ST} analyses, our AMOVA analysis indicated that a majority of the total AFLP variation detected within and among the six intensively sampled populations of *A. thaliana* from eastern North America was due to variation within populations. Although only one-fifth of the total AFLP variation was due to differences among populations, the divergence among all pairs of the six intensively sampled populations was significant. Interestingly, the least differentiation was detected between the nongeographically related population pair MI13–GA5 ($\Phi_{ST} = 0.07$, vs. a range of 0.11–0.30 between the other population pairs).

Our result of high within-population diversity raises the possibility that outcrossing rates may be higher in natural populations in North America than estimated previously. In the field in North Carolina, several species of flies (Syrphidae), small bees (Apidae) and skippers (Hesperiidae) actively visit and probe flowers and transport pollen of *A. thaliana* (Mauricio & Rausher 1997) and we have observed active pollinator visitation in Georgia. Given the surprising levels of within population genetic diversity, we are currently conducting more detailed within-population genetic analyses.

Linkage disequilibrium was expected given the presumed breeding system of *A. thaliana*, the small size of its genome and the possibility that the species has been introduced recently into North America. In fact, we did find significant linkage disequilibrium, although it was not correlated with the geographical relationships among populations. Nordborg *et al*. (2002) have suggested that linkage disequilibrium may be substantial within local populations due to founder events. Our data do not allow us to conclude that the LD we detected was due to any particular evolutionary force, including founder events or natural selection.

In summary, the introduction of *Arabidopsis thaliana* to North America has not left a clear genetic signature. For example, had the introduction of *A. thaliana* involved a single recent introduction, with subsequent spread, we would have detected a relationship between geography and neutral genetic variation. However, we found no evidence of isolation by distance among populations and the levels of genetic diversity within and among populations in the Midwestern and Southeastern regions were virtually uniform. The substantial levels of genetic diversity within populations coupled with the nongeographical nature of divergence among populations are consistent with contemporary gene flow with either a recent or pre-Columbian introduction from either single or multiple sources. A third possibility is a decidedly nonequilibrium scenario with recent (even continuing) introduction to North America from multiple sources with insufficient time for drift to have caused differentiation among populations. Distinguishing among these hypotheses will require more extensive and detailed analyses of neutral genetic variation over space and time.

Acknowledgements

We thank A. Tull and M. Boyd in the Franklin College Plant Biology Greenhouses for their expert plant care, S. Held and E. Rodriquez for technical assistance and R. Baucom, D. Charlesworth, E. Gonzales, J. Ross-Ibarra, J. Stinchcombe and an anonymous reviewer for helpful comments on earlier versions of the manuscript. J. Barber, D. Boufford, P. Holmgren, E. Schuyler and V. Funk provided information on collections of *Arabidopsis thaliana* in their respective institutions. This material is based upon work supported by the National Science Foundation under grant no. 0129191.

References

Abbott RJ, Gomes MF (1989) Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. *Heredity*, **62**, 411–418.

Aguado E, Burt JE (2003) *Understanding Weather and Climate*, 3rd edn. Prentice Hall, Upper Saddle River, NJ.

Allard RW, Jain SK, Workman PL (1968) The genetics of inbreeding species. *Advances in Genetics*, **14**, 55–131.

Barth S, Melchinger AE, Lübberstedt TH (2002) Genetic diversity in *Arabidopsis thaliana* L. Heynh. investigated by cleaved amplified polymorphic sequence (CAPS) and inter-sample sequence repeat (ISSR) markers. *Molecular Ecology*, **11**, 495–505.

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

Breyne P, Rombaut D, Van Gysel A, Van Montagu M, Gerats T (1999) AFLP analysis of genetic diversity within and between *Arabidopsis thaliana* ecotypes. *Molecular and General Genetics*, **261**, 627–634.

Cain ML, Milligan BG, Strand AE (2000) Long-distance seed dispersal in plant populations. *American Journal of Botany*, **87**, 1217–1227.

Carlquist S (1980) *Hawaii: a Natural History*, 2nd edn. Pacific Tropical Botanical Garden, Honolulu.

Clark AG (1997) Estimating nucleotide divergence with RAPD data. In: *Fingerprinting Methods Based on Arbitrarily Primed PCR* (eds Micheli MR, Bova R), pp. 219–226. Springer-Verlag, Berlin.

Colosi JC, Schaal BA (1993) Tissue grinding with ball bearings and a vortex mixer. *Nucleic Acids Research*, **21**, 1051–1052.

Dice LR (1945) Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302.

Erschadi S, Haberer G, Schöniger M, Torres-Ruiz RA (2000) Estimating genetic diversity of *Arabidopsis thaliana* ecotypes with amplified fragment length polymorphisms (AFLP). *Theoretical and Applied Genetics*, **100**, 633–640.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Fisher RA (1930) *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

Futuyma DJ (1998) *Evolutionary Biology*, 3rd edn. Sinauer, Sunderland, MA.

Griffith C, Kim E, Donohue K (2004) Life-history variation and adaptation in the historically mobile plant *Arabidopsis thaliana* (Brassicaceae) in North America. *American Journal of Botany*, **91**, 837–849.

Hamrick JL, Godt MJW (1996a) Conservation genetics of endemic plant species. In: *Conservation Genetics: Case Histories from Nature* (eds Avise JC, Hamrick JL), pp. 281–304. Chapman & Hall, New York.

Hamrick JL, Godt MJW (1996b) Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, **351**, 1291–1298.

Hanfstingl U, Berry A, Kellogg EA *et al.* (1994) Haplotype divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: role for both balancing and directional selection? *Genetics*, **138**, 811–828.

Haubold B, Hudson RR (2000) LIAN 3.0: detecting linkage disequilibrium in multilocus data. *Bioinformatics*, **16**, 847–849.

Hoffmann MH, Schmuths H (2001) *Arabidopsis thaliana* as a tool for biosystematics: studies in molecular phylogeography. *International Organization of Plant Biosystematists Newsletter*, **33**, 17–21.

Innan H, Terauchi R, Miyashita NT (1997) Microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana*. *Genetics*, **146**, 1441–1452.

King G, Nienhaus J, Hussey C (1993) Genetic similarity among ecotypes of *Arabidopsis thaliana* estimated by analysis of restriction fragment length polymorphisms. *Theoretical and Applied Genetics*, **86**, 1028–1032.

Konieczny A, Ausubel F (1993) A procedure for quick mapping of *Arabidopsis* mutants using ecotype specific markers. *Plant Journal*, **4**, 403–410.

Lawrence MJ (1976) Variations in natural populations of *Arabidopsis thaliana* (L.). Heynh. In: *The Biology and Chemistry of the Cruciferae* (eds Vaughan JG, MacLeod AJ, Jones BMG), pp. 167–190. Academic Press, London.

Mantel NA (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Mauricio R, Rausher MD (1997) Experimental manipulation of putative selection agents provides evidence for the role of natural enemies in the evolution of plant defense. *Evolution*, **51**, 1435–1444.

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

Muñoz J, Felicísmo ÁM, Cabezas F, Burgaz AR, Martínez I (2004) Wind as a long-distance dispersal vehicle in the Southern Hemisphere. *Science*, **304**, 1144–1147.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA*, **70**, 3321–3323.

Nordborg M, Borevitz JO, Bergelson J *et al.* (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, **30**, 190–193.

O'Kane SL Jr, Al-Shehbaz IA (1997) A synopsis of *Arabidopsis* (Brassicaceae). *Novon*, **7**, 323–327.

Parsons YM, Shaw KL (2001) Species boundaries and genetic diversity among Hawaiian crickets of the genus *Laupala* identified using amplified fragment length polymorphism. *Molecular Ecology*, **10**, 1765–1772.

Peakall R, Smouse PE (2001) *Genalex V5: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research*. Australian National University, Canberra, Australia. http://www.anu.edu.auu/BoZo/GenAlEx/

Price RA, Palmer JD, Al-Shehbaz IA (1994) Systematic relationships of *Arabidopsis*: a molecular and morphological perspective. In: *Arabidopsis* (eds Meyerowitz EM, Somerville CR), pp. 7–19. Cold Spring Harbor Laboratory Press, New York.

Ratcliffe D (1961) Adaptation to habitat in a group of annual plants. *Journal of Ecology*, **49**, 187–203.

Redei GP (1975) *Arabidopsis* as a genetic tool. *Annual Review of Genetics*, **9**, 111–127.

Rohlf FJ (2002) NTSYSPC: Numerical Taxonomy System, ver. 2.1. Exeter Publishing, Ltd. Setauket, New York.

Rollins RC (1993) *The Cruciferae of Continental North America*. Stanford University Press, Stanford, CA.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

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

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

SSPS Inc. (1989–2002) *SPSS for Windows*, Release 8.0, 1998. SPSS Inc., Chicago.

Swofford D (2002) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4.0 beta 10. Sinauer, Sunderland, MA.

Tackenberg O, Poschlod P, Kahmen S (2003) Dandelion seed dispersal: the horizontal wind speed does not matter for long-distance dispersal — it is updraft! *Plant Biology*, **5**, 451–454.

Todokoro A, Terauchi R, Kawano S (1995) Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. *Japanese Journal of Genetics*, **70**, 543–554.

Travis SE, Maschinshi J, Keim P (1996) An analysis of genetic variation in *Astragalus cremnophylax* var. *cremnophylax*, a critically endangered plant, using AFLP markers. *Molecular Ecology*, **5**, 735–745.

Turesson G (1922) The genotypical response of the plant species to its habitat. *Hereditas*, **3**, 266–281.

Turesson G (1925) The plant species in relation to habitat and climate. *Hereditas*, **6**, 147–236.

Vekemans X, Beauwens T, Lemaire M, Roldán-Ruiz I (2002) Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and size. *Molecular Ecology*, **11**, 139–151.

Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP: a new concept for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Wolfe AD, Liston A (1998) Contributions of the polymerase chain reaction to plant systematics and evolutionary biology. In: *Molecular Systematics of Plants II* (eds Soltis DE, Soltis PS, Doyle JJ), pp. 43–86. Kluwer, New York.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Yeh FC, Yang R, Boyle T (1999) POPGENE, version 1.32. Microsoft Window-Based Freeware for Population Genetic Analysis. University of Alberta, Edmonton. Available at: http://www.ualberta.ca/~fyeh/index.htm.

Stacy Jørgensen is a biogeographer with a main research interest in spatial evolutionary patterns and processes. Rodney Mauricio is a plant evolutionary ecologist whose own research focuses on the evolution of adaptive traits, using wild populations of *Arabidopsis thaliana* as a model system. He has studied natural populations of *A. thaliana* since 1989.
