Chloroplast DNA indicates a single origin of the allotetraploid *Arabidopsis suecica*

T. SÄLL,* M. JAKOBSSON,* C. LIND-HALLDÉN† & C. HALLDÉN‡

*Department of Cell and Organism Biology, Genetics, Lund University, Lund, Sweden
†Department of Mathematics and Natural Sciences, Kristianstad University, Kristianstad, Sweden
‡Department of Clinical Chemistry, Malmö University Hospital, Malmö, Sweden

Keywords:

Arabidopsis suecica;
Arabidopsis thaliana;
chloroplast;
DNA sequence polymorphism;
polyploid.

Abstract

DNA sequencing was performed on up to 12 chloroplast DNA regions [giving a total of 4288 base pairs (bp) in length] from the allopolyploid *Arabidopsis suecica* (48 accessions) and its two parental species, *A. thaliana* (25 accessions) and *A. arenosa* (seven accessions). *Arabidopsis suecica* was identical to *A. thaliana* at all 93 sites where *A. thaliana* and *A. arenosa* differed, thus showing that *A. thaliana* is the maternal parent of *A. suecica*. Under the assumption that *A. thaliana* and *A. arenosa* separated 5 million years ago, we estimated a substitution rate of 2.9×10^{-9} per site per year in noncoding single copy sequence. Within *A. thaliana* we found 12 substitution (single bp) and eight insertion/deletion (indel) polymorphisms, separating the 25 accessions into 15 haplotypes. Eight of the *A. thaliana* accessions from central Sweden formed one cluster, which was separated from a cluster consisting of central European and extreme southern Swedish accessions. This latter cluster also included the *A. suecica* accessions, which were all identical except for one 5 bp indel. We interpret this low level of variation as a strong indication that *A. suecica* effectively has a single origin, which we dated at 20 000 years ago or more.

Introduction

Molecular biology has revitalized many biological questions. One example is polyploidy where a large number of species were traditionally identified as diploid or polyploid on the basis of chromosome counts alone. By combining molecular markers with classical recombination mapping, it has become evident that a much larger number of species than previously thought have a polyploid origin or have ancestors that have gone through polyploidization during recent evolution. In a large proportion of these cases it has been found that so-called allopolyploidy has occurred, i.e. two or more species have combined their genomes to form a new species (see, e.g. Ramsey & Schemske, 1998, 2002). For example, *Brassica nigra*, which has been recognized as a diploid and which is a parent of two existing allopolyploids, *B. carinata* and *B. juncea* (U, 1935), was recently found to have an old allohexaploid genome structure (Lagercrantz & Lydiate, 1996; Lagercrantz, 1998). Similarly, *Zea mays*, which has also been regarded as a diploid, has been identified as an old allotetraploid (Gaut & Doebley, 1997). DNA sequencing, operating at a higher level of resolution than recombination mapping using molecular markers, provides an even more powerful way of identifying polyploidization events in the evolutionary history of species. The recent report of the genome sequence of the plant model organism *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000) shows that a major part of this genome exists as duplicated sequences. This indicates that *A. thaliana* probably has a tetraploid ancestry. However, whether this ancestor was allotetraploid or whether the duplicated sequences are instead the result of several independent segmental duplications is difficult to determine for certain, because of extensive local duplications and gene loss in combination with varying degrees of sequence divergence among different sequences.

Correspondence: Torbjörn Säll, Genetics, Department of Cell and Organism Biology, Lund University, Sölvegatan 29, S-223 62 Lund, Sweden.
Tel.: +46-46-222-7858; fax: +46-46-147-874;
e-mail: torbjorn.sall@gen.lu.se

One complication in the study of polyploids is the fact that they may be of multiple origin, i.e. the same parental species may have crossed several times to produce allopolyploid offspring (Soltis & Soltis, 1993, 1995). The general opinion until a few years ago (YA) was that most allopolyploids were of single origin (the species had arisen on one occasion only), but this view has recently changed. Today, most allopolyploids are considered to be of multiple origin and it is suggested that only a few species have single origins. Two proposed examples of allopolyploid species with single origins are the peanut *Arachis hypogaea* (Kochert *et al*., 1996) and the salt marsh grass *Spartina anglica* (Raybould *et al*., 1991). The number of origins is fundamental to understanding the pace of genome evolution in polyploids. If there is a single origin, all parts of the genome and all evolutionary processes related to them occur on the same time scale. If there are multiple origins, then certain parts of the genome may show a long history of isolation between the parental species whereas others are more recent, which naturally will complicate the analysis of events. Thus if one intends to study a polyploid, it is of fundamental importance to first determine its number of origins.

An allopolyploid species which has recently attracted considerable interest is *A. suecica* (2n = 26). The reason for this is that one of its parental species is *A. thaliana*, which is the leading model organism in plant biology. One of the important features of *A. thaliana* (2n = 10) is its small genome size of approximately 125 Mb (The Arabidopsis Genome Initiative, 2000). *Arabidopsis suecica*'s other parent, *A.* (*Cardaminopsis*) *arenosa* (2n = 4x = 32 and 2n = 16), also has a small genome, approximately 150 Mb (unpublished results). Thus it is safe to conclude that *A. suecica* has one of the smallest genomes among polyploid plants. *Arabidopsis suecica* is mainly found in central Sweden and southern Finland, but isolated populations are found elsewhere in the two countries (Hultén, 1971; unpublished results). *Arabidopsis thaliana*, on the other hand, is rare in central Sweden but common in southern Sweden and southern Finland, whereas the tetraploid *A. arenosa* is quite common in central Sweden but rare in Finland. The diploid form of *A. arenosa* is confined to eastern Europe, primarily Slovakia (Mesicek, 1970). However, whether *A. suecica* originated from the diploid or tetraploid form of *A. arenosa* is not clear. Hultgård (1987) and Suominen (1994) proposed that *A. suecica* originated behind the retreating ice of the last ice age, approximately 10 000 YA.

Hylander (1957) was the first to suggest the parentage of *A. suecica* from data on morphology and chromosome counts. Recently, different molecular techniques have been applied to investigate Hylander's proposal. Kamm *et al.* (1995) sequenced nuclear AT-rich tandem repeats and O'Kane *et al.* (1996) sequenced nuclear rDNA internal transcribed spacer sequences (ITS) from single *A. suecica* accessions. Both of these investigations confirmed Hylander's proposal. Price *et al.* (1994) used restriction analysis of chloroplast DNA (cpDNA) and Mummenhof & Hurka (1994) used isoelectric focusing of *Rubisco* to investigate whether *A. thaliana* or *A. arenosa* was the mother species of *A. suecica*. Both studies found *A. thaliana* to be the mother, but neither ruled out the possibility of multiple origins as they studied only single accessions of *A. suecica.* More recently, artificially synthesized *A. suecica* have been formed in the laboratory from tetraploid *A. thaliana* (mother) and tetraploid *A. arenosa* (father) plants, whereas the reciprocal cross-failed (Comai *et al*., 2000).

We have chosen to study cpDNA sequence variation in *A. suecica* and its parental species in order to assess the number of maternal origins of this allotetraploid species. Chloroplast DNA has been widely used in phylogenetic studies of plants. Compared with nuclear sequences, the advantages of cpDNA are that recombination is rare or nonexistent, the genome is maternally inherited and the rate of structural and sequence evolution is slow (Palmer, 1987; Palmer *et al*., 1988). Studies at the species level have the same advantages except that the slow rate of sequence evolution will result in small amounts of sequence variation. The determination of a 154-kb sequence of the *A. thaliana* chloroplast genome (Sato *et al*., 1999) allowed us to sequence 12 loci covering more than 4 kb of mostly noncoding sequence. We analysed 48 accessions of *A. suecica* and a total of 32 accessions of its parental species (see Materials & methods). The 48 *A. suecica* samples represent the whole known breeding area of the species. The primary goal of our study was to determine the number of origins of *A. suecica*. In cases of multiple origins, reciprocal combinations are possible; all of the *A. suecica* accessions were therefore investigated for two of the above loci. The second goal of our study was to determine which accessions of *A. thaliana* have the most similar chloroplast genomes to that of *A. suecica*. A third goal was to infer, if possible, the time of origin or origins of *A. suecica* or, in other words, to estimate the maximum number of generations separating different accessions of *A. suecica*.

Materials and methods

Plant material and experimental design

A total of 80 accessions of *A. arenosa*, *A. thaliana* and *A. suecica* were analysed in this study (Table 1). The seven *A. arenosa* accessions were exclusively from Sweden, and the 48 *A. suecica* accessions were from Sweden (38) and Finland (10). All but one of the 25 *A. thaliana* accessions were from Europe, with most coming from Scandinavia (13 from Sweden, three from Finland, two from Denmark and one from Norway). Fresh young leaves were harvested from greenhouse-grown plants for immediate DNA extraction using a Plant DNeasy kit from Qiagen. The DNA concentrations and the integrity of the extractions were determined using

Table 1 Description of the seven *Arabidopsis arenosa*, 25 *A. thaliana* and 48 *A. suecica* plants used in this study. The chloroplast loci sequenced for each accession are also shown.

Name	Species	Location	Loci sequenced
A:140	*A. arenosa*	Nyåker (SW)	All
A:170	*A. arenosa*	Gottne (SW)	1, 2, 8 and 11
A:210	*A. arenosa*	Kvarnå (SW)	1, 2, 8 and 11
A:350	*A. arenosa*	Edsäter (SW)	1, 2, 8 and 11
A:380	*A. arenosa*	Sillre (SW)	1, 2, 8 and 11
A:520	*A. arenosa*	Noppikoski (SW)	1, 2, 8 and 11
A:660	*A. arenosa*	Sunne (SW)	1, 2, 8 and 11
T:1	*A. thaliana*	Vänersborg (SW)	All
T:50	*A. thaliana*	Kristianstad (SW)	All
T:160	*A. thaliana*	Västervik (SW)	All
T:81	*A. thaliana*	Karhumäki (FI)*	All
T:93	*A. thaliana*	Tvärminne (FI)*	All
T:104	*A. thaliana*	Nurmes (FI)*	All
Oy-0	*A. thaliana*	Oystese (NO)†	All
Ct-1	*A. thaliana*	Catania (IT)†	All
T:10	*A. thaliana*	Lilla Edet (SW)	All
T:20	*A. thaliana*	Tollarp (SW)	All
T:40	*A. thaliana*	Hässleholm (SW)	All
T:70	*A. thaliana*	Lund (SW)	All
T:700	*A. thaliana*	Anten (SW)	All
T:340	*A. thaliana*	Höör (SW)	All
T:350	*A. thaliana*	Klevshult (SW)	All
T:360	*A. thaliana*	Mantorp (SW)	All
T:370	*A. thaliana*	Kungs-Husby (SW)	All
T:380	*A. thaliana*	Stavsnäs (SW)	All
Kas-1	*A. thaliana*	Kashmir (IN)†	All
Lip-1	*A. thaliana*	Lipowiec (PL)†	All
Gr-1	*A. thaliana*	Graz (AU)†	All
Sv-0	*A. thaliana*	Svebolle (DK)†	All
Wil-1	*A. thaliana*	Wilma (LI)†	All
Bu-0	*A. thaliana*	Burghaun (FRG)†	All
Al-0	*A. thaliana*	Allerup (DK)†	All
S:60	*A. suecica*	Vännas (SW)	All
S:70	*A. suecica*	Söder Nyåker (SW)	All
S:90	*A. suecica*	Västanbäck (SW)	All
S:110	*A. suecica*	Ängebo (SW)	All
S:130	*A. suecica*	Strömsbruk (SW)	All
S:140	*A. suecica*	V Indal (SW)	All
S:150	*A. suecica*	Ytterhogdal (SW)	All
S:170	*A. suecica*	Los (SW)	All
S:223	*A. suecica*	Högsjö (SW)	All
S:240	*A. suecica*	Gålsjö (SW)	All
S:261	*A. suecica*	Hammarstrand (SW)‡	All
S:300	*A. suecica*	Sörfjärda (SW)§	All
S:330	*A. suecica*	Karlstad (SW)	All
S:354	*A. suecica*	Iisalmi (FI)*	All
S:361	*A. suecica*	Hanko (FI)*	All
S:81	*A. suecica*	Nordmaling (SW)	1 and 3
S:500	*A. suecica*	Helsinki (FI)	1 and 3
S:510	*A. suecica*	Helsinki (FI)	1 and 3
S:520	*A. suecica*	Artjärvi (FI)¶	1 and 3
S:530	*A. suecica*	Pålkäms (FI)	1 and 3
S:550	*A. suecica*	Pielavesi (FI)	1 and 3
S:560	*A. suecica*	Knaperåsen (SW)**	1 and 3
S:570	*A. suecica*	Oslättfors (SW)**	1 and 3
S:580	*A. suecica*	Oslättfors (SW)**	1 and 3
S:590	*A. suecica*	Hässleholm (SW)	1 and 3
S:600	*A. suecica*	Lund (SW)	1 and 3
S:700	*A. suecica*	Ulricehamn (SW)	1 and 3
S:370	*A. suecica*	Oulu (FI)	1 and 3
S:380	*A. suecica*	Helsinki (FI)	1 and 3
S:408	*A. suecica*	Axberg (SW)	1 and 3
S:412	*A. suecica*	Grytthyttan (SW)	1 and 3
S:420	*A. suecica*	Ramsberg (SW)	1 and 3
S:430	*A. suecica*	Ramsnäs (SW)	1 and 3
S:441	*A. suecica*	Ängelsberg (SW)	1 and 3
S:459	*A. suecica*	Garpenberg (SW)	1 and 3
S:460	*A. suecica*	Enviken (SW)	1 and 3
S:476	*A. suecica*	Bärby (SW)	1 and 3
S:485	*A. suecica*	Almunge (SW)	1 and 3
S:490	*A. suecica*	Hällen (SW)	1 and 3
S:231	*A. suecica*	Olofsfors (SW)	1 and 3
S:271	*A. suecica*	Stadsforsen (SW)	1 and 3
S:292	*A. suecica*	Ede (SW)	1 and 3
S:311	*A. suecica*	Stocktjärn (SW)	1 and 3
S:340	*A. suecica*	Kotka (FI)	1 and 3
S:221	*A. suecica*	Högsjö (SW)	1 and 3
S:182	*A. suecica*	Våxnan (SW)	1 and 3
S:163	*A. suecica*	Ytterhogdal (SW)	1 and 3
S:122	*A. suecica*	Friggesund (SW)	1 and 3

Except in the following cases, accessions have been collected by the authors: *Outi Savolainen, Oulo University; †Dr Goto, The SENDAI *Arabidopsis* Seed Stock Center; ‡Håkan Lindström, Tjälarne; §Svante Holm, Mitthögskolan; ¶Arrto Kurrto, Helsinki University; **Peter Stål, Gävle. SW, Sweden; FI, Finland; NO, Norway; IT, Italy; IN, India; PL, Poland; AU, Austria; DK, Denmark; LI, Lithuania; FRG, Germany.

agarose gel electrophoresis. The complete chloroplast sequence of *A. thaliana* (GenBank accession number NC_000932) and the software program OLIGO version 6 (Molecular Biology Insights) were used to design a total of 12 polymerase chain reaction (PCR) primer pairs (Table 2). These primer pairs were distributed over the chloroplast genome and amplified different classes of sequence (see Table 3). Seven of the primer pairs amplified mostly noncoding single copy regions (loci 1, 3, 4, 5, 6, 7 and 10), whereas four amplified predominantly noncoding inverted repeat regions (loci 8, 9, 11 and 12). The remaining primer pair amplified the middle of the adenosine triphosphatase alpha subunit gene (locus 2). Some of the 'noncoding' sequences also contain a minor amount of coding sequence (see Table 3). The primer pairs were all designed to amplify DNA sequences of approximately 400 base pairs (bp); the total length of the 12 sequences was 4288 bp. Sequencing was performed according to the following scheme (see also Table 1): I. Loci 1 and 3 were sequenced for the 25 *A. thaliana* and 48 *A. suecica* accessions plus one of the *A. arenosa* accessions in order to settle the issue of the maternal parent of *A. suecica* and to gain a first estimate of

Table 2 Primer sequences used to amplify the analysed sequences. The 3′-positions of the primer sequences in the chloroplast genome of *A. thaliana* (GenBank accession no. NC_000932) and the annealing temperatures used for PCR amplification are shown. Primer pairs 8, 9, 11 and 12 are located in inverted repeat regions and can therefore bind at two locations each.

Primer number [forward (F)/reverse (R)]	Primer sequence	Annealing temperature (°C)	3′-position
1F	5′-ATAGAACTTT CTCAGCAATT C-3′	58	8259
1R	5′-TAAATTAACC TTTTGTCGAA C-3′		8605
2F	5′-GCGCGAGGTA TTGTAACGTA G-3′	58	10764
2R	5′-AAACGCCTTG GCTAACCCTAT-3′		11115
3F	5′-TTTGCTTCAA CCCGTCAACT A-3′	64	32290
3R	5′-TCAACCATTT CCGAACACCT T-3′		32667
4F	5′-AATGATAATC AAATCGCACC A-3′	60	44469
4R	5′-AATGTTACGC CTTCAACCAC T-3′		44840
5F	5′-TTGTGTCGAT CTTGTCCTTCT-3′	60	63084
5R	5′-CTTCTTTGTC TGATTCGAGG G-3′		63456
6F	5′-GTCATTTACC CTGTTAGTCC G-3′	60	69327
6R	5′-GAAATACAAG ACAGCCAATCC-3′		69679
7F	5′-GGGGATAGGC TGGTTCACTT-3′	64	76617
7R	5′-AAATGCTCAA CACCCACGTA A-3′		76983
8F	5′-ATCTCGCACG GCTCCTAAGT-3′	64	95785
8R	5′-TTACGGGTAG TTCCTGCAAA G-3′		96160
9F	5′-AACGCCCTTG TTGACGAT-3′	64	98582
9R	5′-CTAGTTACTC TTCGGGACGGA-3′		98935
10F	5′-TTTTGATTTC TCTTGAGCAAT-3′	60	113947
10R	5′-TTCCTAAGAG CAGCGTGTCT A-3′		114309
11F	5′-TCGGTGTAGG TTCGGGATAA-3′	64	129818
11R	5′-GATAGCGATA GCGGACTCAA A-3′		130195
12F	5′-CCGCTTTGAA ATCGTCC-3′	64	144471
12R	5′-ATTCCAGTTG ACCGAGCCTAA-3′		144821

Table 3 Number of substitutions and insertions/deletions (indels) detected between *A. arenosa* (one accession; all 12 loci) and *A. thaliana* (25 accessions; all 12 loci) and within *A. thaliana*. For each locus, the macro region and whether the locus is part of a coding region are indicated.

			Variation *A. thaliana–A. arenosa*			Variation within *A. thaliana*		
Locus	Macro region	Coding/noncoding*	Substitutions	Indels	Total	Substitutions	Indels	Total
1	LSC	Noncoding	11	2	13	4	2	6
2	LSC	Coding	2	0	2	1	0	1
3	LSC	Noncoding	7	3	10	0	1‡	1
4	LSC	Noncoding (12 bp coding)	5	1	6	0	0	0
5	LSC	Noncoding	16	1 + 1†	18	2	1§ + 1¶	4
6	LSC	Noncoding (66 bp coding)	6	0	6	2	0	2
7	LSC	Noncoding	13	3	16	0	1‡	1
8	IRA	Noncoding	1	0	1	1	0	1
9	IRA	Noncoding (142 bp coding)	0	0	0	0	0	0
10	SSC	Noncoding (17 bp coding)	11	3 + 1†	15	2	1 + 1‡	4
11	IRB	Noncoding	3	2	5	0	0	0
12	IRB	Noncoding	1	0	1	0	0	0
Total			76	17	93	12	8	20

LSC, large single copy region; IRA, inverted repeat region A; SSC, small single copy region; IRB, inverted repeat region B. *If the noncoding loci cover a small part of coding sequence, then the number of coding base pairs (bp) is noted in parentheses. †Large indel (>70 bp). ‡Mononucleotide repeat with three length variants. §Dinucleotide repeat with five length variants. ¶Mononucleotide repeat with four length variants.

the number of maternal origins. II. As no variation was found within *A. suecica* at these two loci, the remaining 10 loci were sequenced for the 25 *A. thaliana* accessions plus 15 of the *A. suecica* accessions and one *A. arenosa* accession in order to investigate which *A. thaliana* accession is most similar to *A. suecica*. III. In a previous paper (Lind-Halldén *et al.*, 2002), we found a high level of variation among nuclear markers in *A. arenosa*. We therefore sequenced four loci (1, 2, 8 and 11) from an additional six *A. arenosa* accessions so as to provide a limited estimate of the amount of cpDNA variation in this species. The loci were chosen to represent the different classes of sequence, but were otherwise selected randomly.

PCR and sequencing

The optimal annealing temperature of each of the 12 primer pairs was determined by testing with four different temperatures (ranging from 58 to 64 °C). PCR reactions were performed in a total reaction volume of 25 μL containing 5 ng of template DNA, 1× PCR reaction buffer (Applied Biosystems, Foster City, CA, USA), 2.5 mM $MgCl_2$, 0.4 μM of each primer (DNA technology A/S, Aarhus, Denmark), 200 μM of each dNTP (Amersham Pharmacia Biotech, Little Chalfont, UK) and 0.75 units of AmpliTaq Gold (Applied Biosystems). The PCR programme consisted of an initial denaturation step of 9 min at 95 °C, followed by 30 cycles with 1 min at 96 °C, 1 min at the appropriate annealing temperature (Table 2) and 2 min at 72 °C. This was then followed by a final elongation step of 10 min at 72 °C. The PCR products were purified using a QIAquick 96 PCR Purification Kit from Qiagen (Hilden, Germany). Sequencing of both strands was performed using labelled dye-terminators from Applied Biosystems (Big-Dye Terminator Cycle Sequencing Kit). The sequencing mix supplied with the kit (ReadyReaction mix) was diluted four times with 80 mM Tris–HCl pH 9.0 and 2 mM $MgCl_2$. Twenty nanograms of primary PCR product was used as template DNA in the sequencing reactions. Everything else was done according to the supplier's protocol. Unincorporated dye terminators were removed from the sequencing reactions using gel-filtration (DyeEx 96 Kit from Qiagen). DNA sequencing was carried out on ABI 310 sequencers using POP 6 polymer and a short capillary (47 cm). The DNA sequences of the two strands were aligned and edited using SEQUENCE NAVIGATOR software (Applied Biosystems). All DNA sequences from the same locus were then aligned and each substitution and insertion/deletion (indel) was double-checked on the electropherograms. The differences were then scored, with substitutions and indels being treated separately. The indels were recorded as present or absent unless tandem repeats were involved, in which case all different length variants were scored. The sequences have been submitted to GenBank under the following accession numbers: AY161952–AY162026, AY163906–AY164256, AY167485–AY167559, AY167907–AY167921 and AY170141–AY170206.

Statistical analysis

We calculated the number of segregating sites (K) and the average number of pairwise differences (Π). Both K and Π were based on substitution differences only (indels were not included). Variation at the nucleotide level was estimated using first the 'Watterson' estimator $\theta_W = K/aL$, where L is the length of the sequence and $a = \Sigma(1/x)$ and $1 \leq x \leq n - 1$, and, secondly, through the average number of pairwise differences per bp, $\pi = \Pi/L$ (see, e.g. Li, 1997). To test for selection we compared θ_W to π using the 'Tajima test' (Tajima, 1989).

Haplotypes were created for the substitutions, the indels and a combination of the two. To visualize the relationship among haplotypes, two networks were constructed as described by Bandelt *et al.* (1999), one for the substitutions only and one for the combined data (Figs 1 and 2). These networks were constructed as follows: all haplotypes differing by single differences were connected by a single step. This was repeated for an increasing number of differences until all haplotypes were connected by their minimum distance. Only primary connections were drawn in the network. When two or more connections appeared which were equidistant from one haplotype, all connections with the same distance were drawn in the network.

Isolation by distance was investigated using the Spearman rank correlation (see, e.g. Sokal & Rohlf, 1995) and the Mantel test (Mantel, 1967). Pairwise geographical distance between accessions (i and j) was calculated as: $Dist_{ij} = X_i \times X_j + Y_i \times Y_j + Z_i \times Z_j$, where $X_i = \cos(lat_i) \times \cos(lon_i)$, $Y_i = \cos(lat_i) \times \sin(lon_i)$ and $Z_i = \sin(lat_i)$, and latitudes (lat) and longitudes (lon) are expressed as radians.

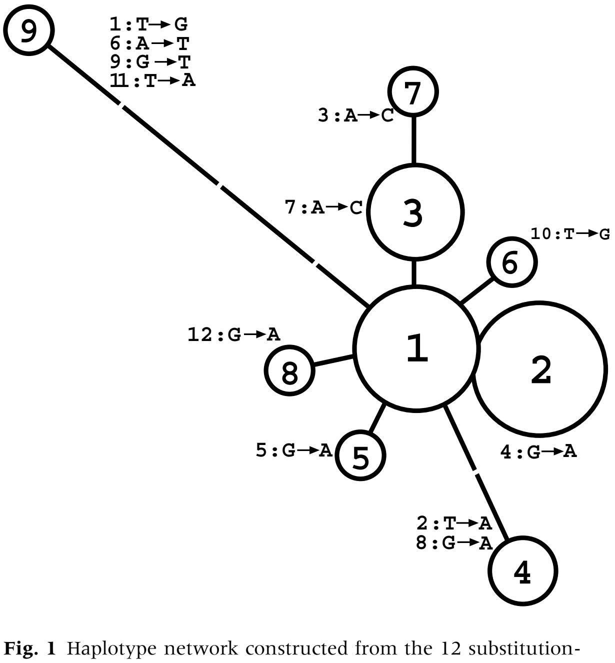

Fig. 1 Haplotype network constructed from the 12 substitution-variable sites within *A. thaliana*. The substitution haplotypes are given in the circles and correspond to those in Table 5, with the area of each circle being proportional to the number of accessions in each haplotype. Haplotype 1 has the following alleles in the 12 polymorphic positions: T, T, A, G, G, A, A, G, G, T, T, G. The substitutions shown in the figure are noted relative to haplotype 1.

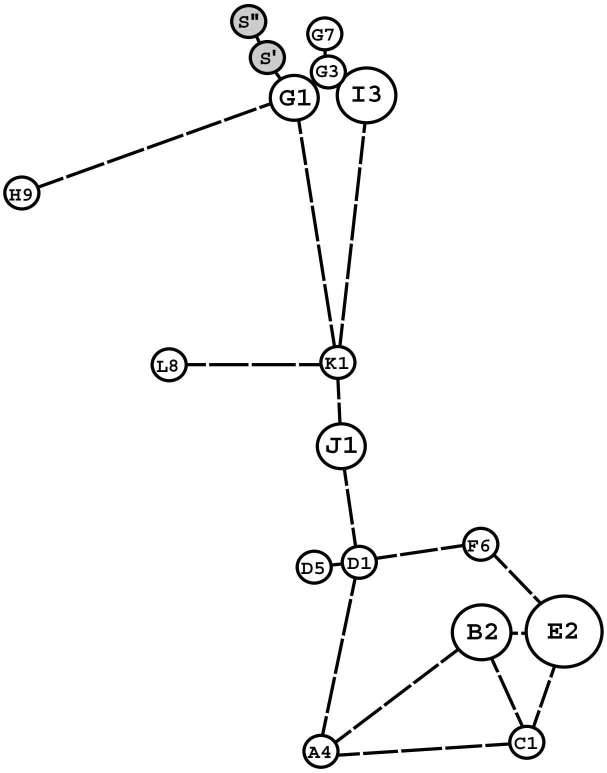

Fig. 2 Haplotype network constructed for *A. thaliana* from both substitutions and insertions/deletions. Haplotype names correspond to those in Table 5 and the areas of the circles are proportional to the number of accessions in each haplotype. Distances between circle midpoints connected by lines are proportional to the number of differences between haplotypes. The exact numbers of differences between haplotypes are shown in Table 6. The two *A. suecica* haplotypes are indicated by S′ and S″.

Results

Differences between *A. thaliana* and *A. arenosa*

The differences between *A. thaliana* (25 accessions, 12 loci) and *A. arenosa* (one accession, 12 loci) are summarized in Table 3. A total of 76 fixed nucleotide substitutions were found between these two species. Comparing *A. arenosa* with the individual accessions of *A. thaliana*, the number of differences varied between 76 and 80 because of variation in *A. thaliana* (see Table 3 and below). Calculated across all 12 loci (4288 bp), this corresponds to 0.018 fixed substitutions per position. Two of the substitutions occurred within coding sequence and five within repeated regions. The remaining 69 fixed substitutions were located within the noncoding, single copy regions (2411 bp; 0.029 substitutions per position). These 69 substitutions consisted of 24 transitions (34.8%) and 45 transversions (65.2%). To test for substitution heterogeneity among the sequences, we performed an ordinary chi-square test. The result showed significant heterogeneity when all 12 loci were tested ($\chi^2_{11} = 49.1$; $P < 0.0001$). On the other hand, no significant heterogeneity was detectable when the loci classified as coding and/or from repeated regions were omitted ($\chi^2_6 = 9.8$; $P = 0.13$). As would be expected, the two groups of sequences (coding and/or repeated and noncoding) showed significant heterogeneity ($\chi^2_1 = 32.9$; $P < 0.0001$).

In addition, 17 indels separated *A. thaliana* and *A. arenosa*. Two of these were considerably larger than the rest: one was situated in locus 10 (85 bp), and the other, in locus 5, was about 100 bp and complex because it was located in a repetitive region. In both cases, *A. thaliana* carried the shorter variant. All of the remaining 15 indels were shorter than 9 bp and in eight cases *A. thaliana* carried the shorter variant. The numbers of substitution and indel differences were found to be positively correlated ($r = 0.78$; $P = 0.003$) for all 12 loci using the Spearman rank correlation coefficient, whereas when the coding and repeated sequences were omitted no significant correlation was found ($r = 0.50$; $P = 0.25$).

Variation within *A. thaliana*

The 25 *A. thaliana* samples showed a limited amount of genetic variation (see Table 3): 20 variable sites were identified, of which 12 were substitutions and eight were indels. Ten of the 12 substitutions were singletons, i.e. the variant occurred in only one accession. Of the two nonsingleton substitutions, one occurred in locus 1 (32% A and 68% G) and the other in locus 5 (20% C and 80% A). Only two of the variable sites were located in the coding and repeated regions. When coding sequences, repeated sequences and sites with alignment gaps were omitted (leaving 2362 bp), the average number of pairwise differences (Π) was 1.43, and Π per bp (π) was 0.61×10^{-3}. The number of segregating sites (K) was 10 and the 'Watterson' estimator calculated per bp was $\theta_W = 1.13 \times 10^{-3}$. Table 4 shows the results for each locus. The observation of $\theta_W > \pi$ was consistent for all variable loci. When Tajima's test of neutrality (Tajima, 1989) was applied to all loci, the observed 'Tajima's D' was −1.52, which is nonsignificant.

Five of the eight indel sites were detected in short stretches of sequence containing tandem repeats of mono- or dinucleotides, i.e. microsatellites. One dinucleotide (AT) repeat sequence was found in five different length variants among the *A. thaliana* accessions (6, 7, 8, 9 and 11 repeat units). One mononucleotide (T) repeat had four length variants (10, 11, 12 and 13 repeat units), whereas three mononucleotide (A) repeats occurred in three length variants each (8, 9 and 10; 9, 10 and 11; 11, 12 and 13 repeat units). Thus the tandem repeats appeared relatively variable. To investigate such repeats further, we identified all additional repeats with more than five repeat units

Table 4 Analysis of the substitution variation within *A. thaliana* (12 loci sequenced for each of 25 accessions).

Locus	Sequence length in base pairs (bp), L (L_e)*	Average number of pairwise differences per bp, π†	Watterson estimator, θ_W‡
1	350 (339)	2.05×10^{-3}	3.12×10^{-3}
2	348	0.23×10^{-3}	0.76×10^{-3}
3	372 (370)	0	0
4	369	0	0
5	370 (361)	1.14×10^{-3}	1.47×10^{-3}
6	349	0.46×10^{-3}	1.52×10^{-3}
7	355 (353)	0	0
8	373	0.21×10^{-3}	0.71×10^{-3}
9	349	0	0
10	341 (316)	0.51×10^{-3}	1.68×10^{-3}
11	372	0	0
12	340	0	0

*L = total sequence length (L_e = sequence length excluding sites with gaps). If only one number is given, then $L_e = L$.

†$\pi = \Pi / L_e$, where Π is the average number of pairwise differences.

‡$\theta_W = K / (a \times L_e)$, where K is the number of segregating sites and $a = \Sigma(1/x)$ from $1 \le x \le n - 1$.

among the 12 loci. We found 22 invariable mononucleotide repeats (14 of these had six repeat units, five had seven repeat units, two had eight repeat units and one had nine repeat units). Accordingly, all repeats of 10 repeat units or more were alleles at variable microsatellite loci in the investigated accessions. There were eight cases where an accession carried a nine repeat unit allele at a variable microsatellite locus. Thus 24% of all repeats with nine repeat units are located at variable microsatellite loci. For eight repeat units, 25% represent alleles at variable microsatellite loci.

Arabidopsis suecica, variation and maternal origin

The *A. suecica* accessions which were sequenced for all 12 loci were identical to each other with respect to substitutions. When *A. suecica* was compared with the individual *A. thaliana* sequences, the number of differences varied between 0 and 4 bp. When *A. suecica* was compared with the *A. arenosa* sequence, 76 substitutions and 17 indels were found. The same pattern was true for the 33 *A. suecica* accessions which were sequenced for loci 1 and 3 (720 bp): they were all identical, differed from *A. thaliana* by 0 or 1 bp and from *A. arenosa* at 23 sites (18 substitutions and five indels). As chloroplasts are maternally inherited, these data identify *A. thaliana* as the maternal parent of *A. suecica*. The only variation within *A. suecica* was one 5 bp indel in locus 5. The extra 5 bp were present in five of 15 (33%) of the *A. suecica* accessions (S:90, S:130, S:170, S:261 and S:361; see Table 1); thus the two variants were found in both Sweden and Finland. All of the 25 *A. thaliana* accessions had the shorter variant.

Haplotype pattern in *A. thaliana* and the origin of *A. suecica*

Based on the 12 substitutions, a total of nine substitution-haplotypes were identified in *A. thaliana* (Table 5). *Haplotype 1* was central and the other haplotypes could be seen as radiating from this haplotype (Fig. 1). *Haplotypes 2, 3, 5, 6* and *8* differed at only one position compared with *haplotype 1* and at two positions compared with each other. *Haplotype 7* was distinguished from *haplotype 3* by an additional substitution. *Haplotype 4* was distinguished from *haplotype 1* by two substitutions, whereas *haplotype 9* differed by four. As only two alleles occurred at each

Table 5 Haplotypes based on substitutions and insertions / deletions (indels) detected in the 25 *A. thaliana* accessions (20 variable sites in total).

Haplotype	Accessions
Substitution	
1	T:81, T:93, T:104, Oy-0, Ct-1, T:40, Lip-1
2	T:1, T:10, T:160, T:350, T:360, T:370, T:380, T:700
3	T:50, T:20, T:340, Bu-0
4	Al-0
5	T:70
6	Kas-1
7	Gr-1
8	Sv-0
9	Wil-1
Indel	
A	Al-0
B	T:1, T:700, T:10
C	T:104
D	T:70, Lip-1
E	T:160, T:350, T:360, T:370, T:380
F	Kas-1
G	T:81, Ct-1, T:20, Gr-1
H	Wil-1
I	T:50, T:340, Bu-0
J	T:93, T:40
K	Oy-0
L	Sv-0
Substitution and indel	
A4	Al-0
B2	T:1, T:700, T:10
C1	T:104
D1	Lip-1
D5	T:70
E2	T:160, T:350, T:360, T:370, T:380
F6	Kas-1
G1	T:81, Ct-1
G3	T:20
G7	Gr-1
H9	Wil-1
I3	T:50, T:340, Bu-0
J1	T:93, T:40
K1	Oy-0
L8	Sv-0

locus, all differences were additive. All of the *A. suecica* accessions were *haplotype 1*.

Based on the eight indel sites, a total of 12 indel-haplotypes were identified (Table 5). The larger number of indel- than substitution-haplotypes was the result of the occurrence of multiple alleles at six of the eight indel loci, which also meant that differences among the indel-haplotypes were not always additive. The largest number of differences for indels was 8, whereas the average number was 4.1.

The pairwise substitution and indel differences among accessions were positively correlated ($r = 0.35$; $P < 0.0002$) using the Spearman correlation coefficient and applying the Mantel test. Four indel-haplotypes unambiguously coincided with substitution-haplotypes (*A* and *4*; *F* and *6*; *H* and *9*; *L* and *8*). Among the remaining haplotypes, a hierarchical pattern emerged, with indel-haplotypes nested within substitution-haplotypes more often than vice versa. Only two indel-haplotypes contained more than one substitution-haplotype (*haplotype D* contained *haplotype 1* plus *haplotype 5*; *haplotype G* contained parts of *haplotypes 1* and *3* plus *haplotype* 7). On the other hand, three of the substitution-haplotypes were divided between eight of the indel-haplotypes (*haplotype 2* was divided between *haplotypes B* and *E*; *haplotype 1* was divided between *haplotype C, D, G, J* and *K*; *haplotype 3* was divided between *haplotypes G* and *I*). Based on all the 20 variable sites (substitutions and indels), a total of 15 haplotypes were identified (Table 5). The number of differences for all combinations of the haplotypes for the combined data is shown in Table 6.

When the haplotypes based on both substitutions and indels were investigated, two major groups appeared (Fig. 2). One group contained haplotypes *B2* and *E2* (eight accessions in total, all of which came from south central Sweden). The other group was made up of haplotypes *G1*, *G3*, *G7* and *I3*. This group contains seven *A. thaliana* and all of the *A. suecica* accessions, with the *A. thaliana* accessions coming from Germany, Austria, Italy, Finland and the southernmost part of Sweden (Scania). With one exception (J1), the rest of the haplotypes all contained one accession. The Columbia sequence represented in GenBank (accession number NC_000932) fell into haplotype *A4*. Genetic differences and geographical distances were compared, again using the Spearman correlation coefficient and the Mantel test. No significant association was found ($r = 0.13$; $P = 0.12$) when the whole body of material was investigated; however, if the Swedish accessions alone were analysed, a significant positive correlation was demonstrated ($r = 0.28$; $P < 0.0002$).

Table 6 Pairwise differences (both substitutions and insertions/deletions) between the 15 haplotypes of *A. thaliana* from Table 5.

	A4	B2	C1	D1	D5	E2	F6	G1	G3	G7	H9	I3	J1	K1	L8
A4		5	5	5	6	6	5	9	10	11	12	9	6	7	8
B2			3	4	5	2	5	9	10	11	12	9	5	6	7
C1				4	5	3	4	7	8	9	10	7	6	5	7
D1					1	4	3	7	8	9	10	7	3	5	8
D5						5	4	8	9	10	11	8	4	6	9
E2							3	8	9	10	11	8	5	4	8
F6								7	8	9	10	7	4	4	8
G1									1	2	6	2	6	6	7
G3										1	7	1	7	7	8
G7											8	2	8	8	9
H9												8	9	9	11
I3													7	7	8
J1														2	5
K1															4
L8															

Discussion

Differences between *A. thaliana* and *A. arenosa*

In this study, we found that the two parental species of *A. suecica* differed by 0.018 fixed nucleotide substitutions per bp when calculated across all 12 loci and 0.029 substitutions per bp in noncoding single copy DNA alone. In order to estimate the rate of substitution per year an estimate of divergence time is needed. Kuittinen & Aguadé (2000) estimated the time of divergence between *A. thaliana* and *A. arenosa* to be 3.8–5.8 million years ago (MYA) using Rorippa pollen records as the fossil reference point and sequence data from the Chalcone Isomerase gene. Koch *et al.* (2000) reported 5.1–5.4 MY based on the same fossil reference point plus sequences from two other genes. If we assume that the divergence was 5 MYA, then the substitution rate becomes 1.8×10^{-9} per position and year $[0.018/(2 \times 5 \times 10^{6})]$ for all loci and 2.9×10^{-9} for noncoding single copy sequence. The synonymous substitution rates in plant mitochondrial, chloroplast and nuclear genes have been found to occur in the approximate ratio of 1 : 3 : 12 (Li, 1997), with a substitution rate in cpDNA in the range $1.0–1.5 \times 10^{-9}$. This is only slightly lower than the rate observed here. More recently, Yang *et al.* (1999) found a substitution rate of $0.5–0.7 \times 10^{-9}$ in mitochondrial DNA among a number of Brassicaceae species (including *A. thaliana*), whereas Koch *et al.* (2000) reported a nuclear synonymous substitution rate of 1.5×10^{-8} among *Arabidopsis* and *Arabis* species. The overall pattern among the different genome components of *A. thaliana* and its relatives thus appears to be compatible with Li's (1997) 'rule of thumb' ratio.

Variation within *A. thaliana*

We found the average level of interpopulation variation among the *A. thaliana* noncoding single copy sequences in the *A. thaliana* cp genome to be low ($\pi = 0.61 \times 10^{-3}$). Bergelson *et al.* (1997) reported an average π equal to

1.4×10^{-3} in mainly noncoding nuclear sequences. Other estimates of silent nucleotide diversity in nuclear sequences of *A. thaliana* tend to be higher from 5×10^{-3} to 15×10^{-3} (see Aguadé, 2001 and references therein). The ratio between the π-values of the present study and Bergelson *et al.* is slightly lower than Li's (1997) 1 : 4 ratio for cp : nuclear sequence (see above); however, when our result is compared with those of the other studies, the ratio is much greater. Our observation of high levels of variability in mononucleotide repeats indicates that these may be useful for improving resolution in studies of genetic variation. Similar observations have been made in other species, such as *Abies alba* (Vendramin *et al.*, 1999), *Picea abies* (Vendramin *et al.*, 2000) and six species of *Glycine* (Powell *et al.*, 1995). On the other hand, Provan *et al.* (1999) found a lower level of variation in mononucleotide repeats in cpDNA of *Pinus torreyana*, but their estimate of the relevant mutation rate (5×10^{-5}) was still considerably higher than the point mutation rate.

The largest pairwise difference between *A. thaliana* accessions observed in our study was six substitutions in 4288 positions. Assuming (as above) that *A. thaliana* originated 5 MYA thus gives an estimate of approximately 400 000 YA ($6/76 \times 5 \times 10^6$) for the deepest divergence among the *A. thaliana* accessions investigated here. Another way to estimate the divergence time is to use the average pairwise difference for the *A. thaliana* dataset (1.59), which is expected to correspond to half the total divergence time (Kingman, 1982). Using this approach, the divergence time is approximated at 200 000 YA ($1.59/76 \times 2 \times 5 \times 10^6$). These estimates indicate a much shorter time of divergence than observations of nuclear DNA sequences. Koch *et al.* (2000) estimated a divergence time of 1.5 MYA and a divergence time of 1.1 MYA can be inferred using the data presented by Miyashita *et al.* (1998). The data for synonymous substitutions in Kawabe *et al.* (2000), on the other hand, indicate a divergence time of less than 400 000 YA. These differences may of course be due to the different samples of accessions used, but may also reflect different coalescence times for nuclear sequences and the cp genome.

The *A. thaliana* haplotypes reveal a number of striking features. One is that indel-haplotypes are contained within substitution-haplotypes to a large extent rather than vice versa. Our interpretation of this is that the indels in general are younger than the substitutions. Another interesting feature is that there are signs of isolation by distance within Sweden, primarily separating the Scanian accessions from the south central accessions. Isolation by distance is not usually found in *A. thaliana* with the exception of effects because of very large distances, such as when Asian and European accessions are compared (Sharbel *et al.*, 2000). Our results indicate that, geographical structure may also play a role on a smaller scale, at least for cytoplasmic DNA. Finally, the placing of *A. suecica* among the Scanian and central European accessions (plus one Finnish accession), rather than among the other Swedish accessions, is interesting. This pattern is more compatible with the formation of *A. suecica* in Europe south of the Fennoscandian peninsula than with a formation in Finland or Sweden. The presumed late entry of *A. arenosa* into Sweden and Finland, which could be hard to reconcile with the present range of *A. suecica* if the last species was formed *in situ* (Mummenhof & Hurka, 1995; Lind-Halldén *et al.*, 2002), is no longer a problem given the alternative scenario.

Single origin of *A. suecica*

The most striking observation in our study is the very low level of cpDNA variation within *A. suecica*. It should be pointed out that the level of genetic variation in *A. thaliana* found in this study is sufficient to allow detection of multiple origins in *A. suecica*. Given that *A. suecica* is endemic to Sweden and Finland and that our sample of 48 populations covers the entire range of the species, our results indicate very strongly that *A. suecica* effectively has a single origin, at least with respect to the maternal parent. Studies of nuclear variation corroborate our conclusion. Lind-Halldén *et al.* (2002) found that *A. suecica* showed much lower levels of variation than either of its parental species for nuclear markers and Hagenblad & Nordborg (personal communication) found no variation when approximately 2000 bp of nuclear sequence from four accessions of *A. suecica* were sequenced. Thus *A. suecica* can be added to the list of allopolyploids with a putative single origin, along with *Arachis* and *Spartina*. As pointed out earlier, this makes it particularly suited for determining the kind and pace of molecular changes that occur during polyploid genome evolution. The drawback is that the intrinsically low level of genetic variation may impair certain types of investigations, such as genetic mapping.

Dating the origin of *A. suecica*

The 15 accessions of *A. suecica* for which we sequenced all 12 loci were identical with respect to substitution differences. Under the assumption that the cp genome does not recombine, the 15 accessions will have a genealogy that can be represented by a single tree. If we consider noncoding single copy sequences only and assume that the substitution rate found between *A. thaliana* and *A. arenosa* is also true for *A. suecica*, then the probability of not observing any variation will be $[1 - 2.9 \times 10^{-9}]^{2411 \times T \times G}$, where T is the time to the common ancestor and G is the total length of the tree scaled in units of T. The value of G is of course dependent on the structure of the tree; here it must be between the two extremes 15 and 2. It has been suggested (Hultgård, 1987; Suominen, 1994) that T is 10 000 years. Using the above formula, we find that the probability of our observation is 0.35 for $G = 15$ and 0.87 for $G = 2$. Our observation of no substitution differences among in the samples is thus fully compatible with an origin at 10 000 YA. The same formula can be used to calculate the upper limit of a 95% probability interval, which is 29 000 and 216 000 YA, respectively. One variable site was detected in *A. suecica*, a 5-bp indel with a maximum difference of one between accessions and an average of approximately 0.5. Above, it was estimated that the deepest divergence among the investigated *A. thaliana* accessions occurred 200 000–400 000 YA. The maximum pairwise difference among the *A. thaliana* accessions with respect to indels is 8, with an average of 4.37 for the noncoding sequences. Assuming a similar rate of change in indels between *A. thaliana* and *A. suecica*, the estimate of *A. suecica*'s time of origin becomes 25 000–50 000 YA using the maximum values and 23 000–46 000 YA using averages. It is very likely that our sample of 15 accessions has a common ancestor that is close to the origin of the species. To illustrate this, under strict neutrality a random sample of 15 individuals is expected to cover 93% of the time to the common ancestor of all individuals living today, irrespective of population size (Li, 1997). These estimates are naturally very rough, but the possibility that *A. suecica* formed more than 20 000 YA is consistent with the observation that the species is most similar to the *A. thaliana* accessions originating in central Europe. The inland ice reached northern Germany at that time (Andersen & Borns, 1994) and thus a formation in Sweden or Finland would not have been possible.

Acknowledgments

We thank M. Dehlin, N. Goto, S. Holm, A. Kurrto, H. Lindström, M. Nordborg, O. Savolainen and P. Stål for help with material, M. Sterner and L. Hall for technical assistance and M. Nordborg for suggestions on the analyses. We thank B. O. Bengtsson and K. E. Slack for comments on the manuscript. This work was supported by the Crafoord foundation, the Erik-Philip Sörensen foundation, the Tryggers foundation and Magnus Bergvall's foundation.

References

Aguadé, M. 2001. Nucleotide sequence variation at two genes of the phenylproaonid pathway, the FAH1 and F3H genes, in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **18**: 1–9.

Andersen, B.G. & Borns, H.W. Jr 1994. *The Ice Age World.* Scandinavian University Press, Oslo.

Bandelt, H.-J., Forster, P. & Röhl, A. 1999. Median-joining network for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**: 37–48.

Bergelson, J., Stahl, E., Dudek, S. & Kreitman, M. 1997. Genetic variation within and among populations of *Arabidobsis thaliana*. *Genetics* **148**: 1311–1323.

Comai, L., Tyagi, A.P., Winter, K., Holmes-Davis, R., Reynolds, S.H., Stevens, Y. & Byers, B. 2000. Phenotypic instability and rapid gene silencing in newly formed Arabidobsis allotetraploids. *Plant Cell* **12**: 1551–1567.

Gaut, B.S. & Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci.* **94**: 6809–6814.

Hultén, E. 1971. *Atlas of the Distribution of Vascular Plants in Northwestern Europe*. Generalstabens litografiska anstalts förlag, Stockholm.

Hultgård, U.-M. 1987. *Parnassia palustris* L. in Scandinavia. *Symbolae Bot. Upsalienses* **28**: 1–128.

Hylander, N. 1957. *Cardaminopsis suecica* (Fr.) Hiit., a northern amphidiploid species. *Bull. Jardin Bot. Bruxelles.* **27**: 591–604.

Kamm, A., Galasso, I., Schmidt, T. & Heslop-Harrison, J.S. 1995. Analysis of a repetitive DNA family from *Arabidobsis arenosa* and relationships between *Arabidopsis* species. *Plant Mol. Biol.* **27**: 853–862.

Kawabe, A., Yamane, K. & Miyashita, N.T. 2000. DNA polymorfism at the cytosolic phosphoglucose isomerase (*PgiC*) locus of the wild plant *Arabidopsis thaliana*. *Genetics* **156**: 1339–1347.

Kingman, J.F.C. 1982. On the genealogy of large populations. *J. Appl. Prob.* **19A**: 27–43.

Koch, M.A., Haubold, B. & Mitchell-Olds, T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis* and related genera (Brassicaceae). *Mol. Biol. Evol.* **17**: 1483–1498.

Kochert, G., Stalker, H.T., Gimenes, M., Galgaro, L., Lopes, C.R. & Moore, K. 1996. RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, *Arachis hypogaea* (Leguminosae). *Am. J. Bot.* **83**: 1282–1291.

Kuittinen, H. & Aguadé, M. 2000. Nucleotide variation at the CHALCONE ISOMERASE locus in *Arabidobsis thaliana*. *Genetics* **155**: 863–872.

Lagercrantz, U. 1998. Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that *Brassica* genomes evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* **150**: 1217–1228.

Lagercrantz, U. & Lydiate, D.J. 1996. Comparative genome mapping in Brassica. *Genetics* **144**: 1903–1910.

Li, W.-H. 1997. *Molecular Evolution.* Sinauer Associates Inc. Publishers, Sunderland, MA.

Lind-Halldén, C., Säll, T. & Halldén, C. 2002. Genetic variation in *Arabidopsis suecica* and its parental species *A. arenosa* and *A. thaliana*. *Hereditas* **136**: 45–50.

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

Mesicek, J. 1970. Chromosome counts in *Cardaminopsis arenosa* Agg. (Cruciferae). *Preslia* **42**: 225–248.

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

Mummenhof, K. & Hurka, H. 1994. Subunit polypeptide composition of Rubisco and the origin of allopolyploid *Arabidobsis suecica* (Brassicaceae). *Biochem. Syst. Ecol.* **22**: 807–812.

Mummenhof, K. & Hurka, H. 1995. Allopolyploid origin of *Arabidobsis suecica* (Fries) Norrlin: evidence from chloroplast and nuclear genome markers. *Bot. Acta* **108**: 449–456.

O'Kane, S.L., Schaal, B.A. & Al-Shehbaz, I.A. 1996. The origin of *Arabidobsis suecica* (Brassicaceae) as indicated by nuclear rDNA sequences. *Syst. Bot.* **21**: 559–566.

Palmer, J.D. 1987. Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. *Am. Naturalist* **130**: S6–S29.

Palmer, J.D., Jansen, R.K., Michaels, H.J., Chase, M.W. & Manhart, J.R. 1988. Chloroplast DNA variation and plant phylogeny. *Ann. Missouri Bot. Garden* **75**: 1180–1206.

Powell, W., Morgante, M., Andre, C., Mcnicol, J.M., Machray, G.C., Doyle, J.J., Tingey, S.V. & Rafalski, J.A. 1995. Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. *Curr. Biol.* **5**: 1023–1029.

Price, R.A., Al-Shebaz, I.A. & Palmer, J.D. 1994. Systematic relationships of *Arabidopsis*: a molecular and morphological perspective. In: *Arabidopsis* (E. Meyerowitz & C. Somerville, eds), pp. 7–19. Cold Spring Harbor Laboratory Press, Cold Spring, New York.

Provan, J., Soranzo, N., Wilson, N.J., Goldstein, D.B. & Powell, W. 1999. A low mutation rate for chloroplast microsatellites. *Genetics* **153**: 943–947.

Ramsey, J. & Schemske, D.W. 1998. Pathways, mechanisms and rates of polyploid formation in flowering plants. *Ann. Rev. Ecol. Syst.* **29**: 467–501.

Ramsey, J. & Schemske, D.W. 2002. Neopolyploidy in flowering plants. *Ann. Rev. Ecol. Syst.* **33**: 589–639.

Raybould, A.F., Gray, A.J., Lawrence, M.J. & Marshall, D.F. 1991. The evolution of *Spartina anglica* CE Hubbard (Gramineae) – origin and genetic variability. *Biol. J. Linn. Soc.* **43**: 111–126.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E. & Tabata, S. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* **6**: 283–290.

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

Sokal, R.R. & Rohlf, F.J. 1995. *Biometry*, 3rd edn. Freeman, New York, USA.

Soltis, D.E. & Soltis, P.S. 1993. Molecular data and the dynamic nature of polyploidy. *Crit. Rev. Plant Sci.* **12**: 243–273.

Soltis, D.E. & Soltis, P.S. 1995. The dynamic nature of polyploid genomes. *Proc. Natl. Acad. Sci. USA* **92**: 8089–8091.

Suominen, J. 1994. Ruotsinpitkäpalon, *Arabidobsis suecica*, syntyseudusta. *Lutukka* **10**: 77–84.

Tajima, F. 1989. Statistical test for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

The Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *A. thaliana*. *Nature* **408**: 796–815.

U, N. 1935. Genome analysis in Brassica with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jap. J. Bot.* **7**: 389–452.

Vendramin, G.G., Anzidel, M., Madaghhiele, A., Sperisen, C. & Bucci, G. 2000. Chloroplast microsatellite analysis reveals the presence of population subdivision in Norway spruce (*Picea abies* K.). *Genome* **43**: 68–78.

Vendramin, G.G., Degen, B., Petit, R.J., Anzidel, M., Madaghhiele, A. & Ziegenhagen, B. 1999. High levels of variation at *Abies alba* chloroplast microsatellite loci in Europe. *Mol. Ecol.* **8**: 1117–1126.

Yang, Y.-W., Lai, K.-N., Tai, P.-Y. & Li W.-H. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J. Mol. Evol.* **48**: 597–604.

Received 23 October 2002; revised 28 February 2003; accepted 19 March 2003