doi: 10.1111/j.1420-9101.2006.01217.x

The evolutionary history of the common chloroplast genome of *Arabidopsis thaliana* and *A. suecica*

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

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

Keywords:

Arabidopsis suecica;
Arabidopsis thaliana;
chloroplast;
DNA sequence polymorphism;
evolutionary history;
polyploidy;
population structure;
speciation.

Abstract

The evolutionary history of the common chloroplast (cp) genome of the allotetraploid *Arabidopsis suecica* and its maternal parent *A. thaliana* was investigated by sequencing 50 fragments of cpDNA, resulting in 98 polymorphic sites. The variation in the *A. suecica* sample was small, in contrast to that of the *A. thaliana* sample. The time to the most recent common ancestor (T_{MRCA}) of the *A. suecica* cp genome alone was estimated to be about one 37th of the T_{MRCA} of both the *A. thaliana* and *A. suecica* cp genomes. This corresponds to *A. suecica* having a MRCA between 10 000 and 50 000 years ago, suggesting that the entire species originated during, or before, this period of time, although the estimates are sensitive to assumptions made about population size and mutation rate. The data was also consistent with the hypothesis of *A. suecica* being of single origin. Isolation-by-distance and population structure in *A. thaliana* depended upon the geographical scale analysed; isolation-by-distance was found to be weak on the global scale but locally pronounced. Within the genealogical cp tree of *A. thaliana*, there were indications that the root of the *A. suecica* species is located among accessions of *A. thaliana* that come primarily from central Europe. Selective neutrality of the cp genome could not be rejected, despite the fact that it contains several completely linked protein-coding genes.

Introduction

The chloroplast (cp) genome has been widely used for studying phylogeny, evolution and ecology in plants (Clegg *et al.*, 1994; Olmstead & Palmer, 1994). The cp genome is very convenient to use from a population-genetic standpoint because of its absence of recombination (Palmer, 1987; Palmer *et al.*, 1988). It thus behaves as a single unit with one unique genealogy. The Y-chromosome and the mitochondrial genome which, as the cp genome, contains uniparentally inherited nonrecombining DNA, has been employed successfully in inferring the evolutionary and demographic history of humans using coalescent methods (Wilson & Balding, 1998; Pritchard *et al.*, 1999). The cp genome is well-suited for a similar approach to study the evolutionary history of plants (Ennos *et al.*, 1999; Provan *et al.*, 2001).

An area in which the application of cpDNA has been successful is in the analysis of parentage in allopolyploids (Widmer & Baltisberger, 1999; Ackerfield & Wen, 2003; Abbott & Lowe, 2004). An example illustrating multiple origins is provided by Segraves *et al.* (1999), who showed not only that *Heuchera grossulariifolia* has several cytoplasmic origins, but also that local populations sometimes host a mixture of cytoplasms of differing origin. This observation accords well with the current view of allopolyploids often being of multiple origin (Soltis & Soltis, 1993, 1995). In such cases, the time to the most recent common ancestor (T_{MRCA}) of the cp genome in the allopolyploid species may be as deep as the T_{MRCA} of the cp genome in its maternal species. The cp genome of an allopolyploid species that has a single origin is expected to be much less variable than the genome of its maternal species and may even be invariable. A low level of cp variation has, correspondingly, been interpreted as a sign that the sampled allopolyploid species is of single origin

Correspondence (present address): Mattias Jakobsson, Bioinformatics Program, Department of Human Genetics, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Ave, Ann Arbor, MI 48109-2218, USA.
Tel.: +734 615 9545; fax: +734 615 6553;
e-mail: mjakob@umich.edu

(Raybould *et al.*, 1991; Kochert *et al.*, 1996; Widmer & Baltisberger, 1999; Säll *et al.*, 2003; Ainouche *et al.*, 2004). Whether low levels of variation in polyploids is, in fact, because of a single polyploidization event, as suggested for *Spartina* by Ainouche *et al.* (2004), or multiple polyploidization events involving very closely related parents, is of course very difficult to know. However, in either of these cases, a sample of lineages from the allopolyploid species will coalesce much more recently in time than a sample of lineages from its maternal species.

Allopolyploids have received considerable attention recently for their specific possibilities in the study of genome evolution (Wendel, 2000; Bennett, 2004). The research has been focused on the nucleus and the specific conditions present in allopolyploids of two homologous genomes existing in parallel. However, the cytoplasmic genomes of allopolyploids also provide interesting opportunities for studying genome evolution. This is particularly true of species with a single origin which allows direct comparisons of different aspects of the cp genome of the allopolyploid and its maternal parent. In particular, the cp genome of an allopolyploid can be used to infer the maternal origin given a single origin of the species. If one assumes a single origin, the results from the cp are then informative about the entire species and not just the genomic component that comes from the maternal parent. However, if that assumption cannot be made, any information derived from the cp genome only concerns the genomic component that comes from the maternal parent. In the case of multiple origins of an allopolyploid, different parts of the genome (where the cp genome is one locus) will contain information about the different origins and the situation becomes much more complicated. Even if we assume a single origin of the allopolyploid species, there are some limitations of the conclusions that can be drawn from the cp genome. Most importantly, the cp genome is one locus and will therefore not provide statistical confidence in conclusions about the whole species. Moreover, conclusions that are based on one genealogy must be treated with caution as the genealogy of a single locus does not necessarily reflect the history of the species (Rosenberg & Nordborg, 2002).

Arabidopsis suecica (2n = 26) is an allotetraploid species for which a single origin has been proposed by investigating nuclear DNA markers (Jakobsson *et al.*, 2006) as well as cpDNA sequence (Säll *et al.*, 2003). As the model plant *A. thaliana* (2n = 10) is one of *A. suecica*'s parental species, the other being *A.* (*Cardaminopsis*) *arenosa* (2n = 16, 32; Hylander, 1957; Kamm *et al.*, 1995; O'Kane *et al.*, 1996), increasing attention is currently being directed at *A. suecica* (Pontes *et al.*, 2003; Säll *et al.*, 2003; Chen *et al.*, 2004; Wang *et al.*, 2004). In addition, several investigations using molecular techniques have recently identified *A. thaliana* as the maternal parent of *A. suecica* (Mummenhoff & Hurka, 1994; Price *et al.*, 1994; Comai *et al.*, 2000; Säll *et al.*, 2003). *Arabidopsis suecica* is highly, although not completely, selfing (Säll *et al.*, 2004). It is found mainly in central Sweden and southern Finland (Hultén, 1971; T. Säll, unpublished data). As its present breeding range was covered by ice during the last glaciation, which retreated some 10 000 years ago (Andersen & Borns, 1994), there are two possible temporal scenarios for its origin. If the species is <10 000 years old, it may have originated in the wake of the retreating ice in Sweden and Finland, as Hultgård (1987) and Suominen (1994) have proposed. If, on the other hand, it is more than 10 000 years old, it must have originated somewhere south of the ice cover and have spread north when the ice retreated, this being followed by extinctions of all its populations outside Sweden and Finland.

In this article we will: (i) quantify and compare the levels of genetic variation in *A. thaliana* and *A. suecica* cpDNA; (ii) investigate the geographic distribution of the cpDNA variation of the two species; attempt to (iii) date the MRCA of the *A. suecica* cp, which, in the case of a single origin, is the lower bound on the time-of-origin of *A. suecica*; (iv) identify which *A. thaliana* accessions in a worldwide sample that are the most closely related to the *A. suecica* accessions. To be able to analyse the different questions in sufficient depth, we have based our studies on three partly overlapping datasets. In particular a large number of *A. thaliana* accessions from Sweden have been studied.

Material and methods

Plant material and experimental design

A total of 150 plants from *A. thaliana* and *A. suecica* were used in the study (Tables 1 and 2). Fresh young leaves from greenhouse grown plants were harvested for DNA extraction, using the Plant DNeasy kit from Qiagen (Hilden, Germany). Polymerase chain reaction (PCR)/sequencing primers (Table S1) were designed using the OLIGO (v.6.3) software (Molecular Biology Insights, Cascade, CO, USA) for amplifying a total of 50 fragments from the cp genome. All of the primers amplified fragments residing outside the two large repeated regions, IRA and IRB, present in the *A. thaliana* cp genome (Sato *et al.*, 1999). PCR reactions were performed in a 25 μL mixture containing 0.5 ng of template DNA, a 1x PCR reaction buffer (Applied Biosystems, Foster City, CA, USA), 2.5 mM $MgCl_2$, 0.4 μM of each primer (DNA technology A/S), 200 mM of each dNTP (Amersham Pharmacia Biotech, Uppsala, Sweden) and 0.75 units of AmpliTaq Gold (Applied Biosystems). The PCR products were purified using the QIAquick 96 PCR Purification Kit from Qiagen. Both strands were sequenced. Sequencing was performed using labelled dye-terminators from Beckman (Fullerton, CA, USA) (CEQ Dye Terminator Cycle Sequencing Quick Start Kit). The sequencing was carried out on a Beckman CEQ 2000 sequencer, using short capillary arrays (CEQ Separation

Table 1. Names and locations of the accessions in datasets TS23 and TS113. The species is indicated by a *t* for *Arabidopais thaliana* and an *s* for *Arabidopais suecica*. All 113 accessions were included in dataset TS113 and the accessions indicated by 1 under TS23 were included in dataset TS23.

Name	S.	Location	Reg.	lat.	lon.	TS23	Name	S.	Location	Reg.	lat.	lon.	TS23
Ag-0	*t*	Argentat	FR	45.05N	01.55E		Ri-0	*t*	Richmond BC	CA	49.50N	123.00W	
An-1	*t*	Antwerpen	BE	51.13N	04.24E		Rou-0	*t*	Rouen	FR	49.27N	01.05E	
BENSHEIM	*t*	Bensheim	GE	49.40N	08.36E		Rsch-0	*t*	Rschew/Sterize	RU	56.15N	34.19E	
Bl-1	*t*	Bologna	IT	44.30N	11.20E		S-96	*t*	Mühlen	PL	53.54N	20.21E	
Bla-2	*t*	Blanes	SP	41.40N	02.47E		Sei-0	*t*	Seis am Schlern	IT	46.32N	11.33E	
Blh-1	*t*	Bulhary	CZ	48.50N	16.45E		Sf-1	*t*	San Feliu	SP	41.07N	03.01E	
Bu-0	*t*	Burghaun	GE	50.42N	09.43E	1	SN-(5)-1	*t*	Unknown	CZ	49.00N	17.00E	
Can-0	*t*	Las Palmas	CI	28.16N	14.00W		Stv-0	*t*	Stabowa	RU	52.00N	36.00E	
CAP-VERDE	*t*	Cape-Verde Isl.	CVI	15.01N	23.36W		Su-0	*t*	Southport	GB	53.39N	03.00W	
Cha-0	*t*	Champex	SL	46.02N	07.06E		Sv-0	*t*	Svebolle	DK	55.38N	11.16E	1
Cnt-1	*t*	Canterbury	GB	51.17N	01.04E		T1	*t*	Vänersborg†	CSW	58.23N	12.19E	
Co-1	*t*	Coimbra	PO	40.13N	08.25W		T10	*t*	Lilla Edet†	CSW	58.06N	12.09E	1
Col-0	*t*	Reference	U	–	–		T104	*t*	Nurmes‡	CFI	63.32N	29.10E	1
DA-(1)-12	*t*	Unknown	CZ	49.00N	17.00E		T110	*t*	Liarum†	SSW	55.57N	13.51E	
DHV230.004	*t*	Maidstone, Kent	GB	51.16N	00.31E		T120	*t*	Kävlinge†	SSW	55.48N	13.07E	
DIJON-G	*t*	Dijon	FR	47.19N	05.01E		T140	*t*	Älmtamåla†	SSW	56.27N	15.30E	
EDEN	*t*	Eden*	NSW	62.53N	18.11E		T150	*t*	Vimmerby*	CSW	57.40N	15.51E	
EDSÄTER	*t*	Edsäter*	NSW	62.54N	18.22E		T160	*t*	Västervik*	CSW	57.45N	16.40E	1
EKOH-2	*t*	Kristianstad†	SSW	56.02N	14.26E		T170	*t*	Lund†	SSW	55.42N	13.11E	
ENKHEIM-1	*t*	Frankfurt	GE	50.09N	08.45E		T181	*t*	Lund†	SSW	55.42N	13.11E	
ENKVEIN	*t*	Frankfurt	GE	50.08N	08.44E		T190	*t*	Lund†	SSW	55.42N	13.11E	
Er-0	*t*	Erlangen	GE	49.35N	11.00E		T20	*t*	Tollarp†	SSW	55.56N	13.59E	1
Est-0	*t*	Unknown	ES	58.59N	23.28E		T200	*t*	Hovdala†	SSW	56.06N	13.43E	
Fl-1	*t*	Unknown	CFI	63.00N	25.00E		T30	*t*	Vinslöv†	SSW	56.05N	13.58E	
Fr-5	*t*	Frankfurt	GE	50.08N	08.39E		T340	*t*	Höör†	SSW	55.56N	13.32E	1
Ge-1	*t*	Genève	SL	46.12N	06.08E		T350	*t*	Klevshult†	CSW	57.21N	14.05E	1
Gr-1	*t*	Graz	AU	47.04N	15.25E		T360	*t*	Mantorp†	CSW	58.21N	15.17E	
Hel-1	*t*	Helsinki	SFI	60.20N	25.00E		T370	*t*	Kungs-Husby†	CSW	59.32N	17.15E	
Hl-0	*t*	Hilversum	NL	52.13N	05.10E		T380	*t*	Stavsnäs-Värmdö†	CSW	59.17N	18.41E	
Hodja-Obi-G.	*t*	Hodja-Obi-Garm	TA	38.43N	69.41E		T40	*t*	Hässleholm†	SSW	56.10N	13.46E	
HÖRBY-1	*t*	Hörby†	SW	55.51N	13.39E		T400	*t*	Flyinge†	SSW	55.45N	13.21E	
Jl-1	*t*	Vranov u Brno	CZ	49.19N	16.38E		T410	*t*	Revinge†	SSW	55.44N	13.27E	
Kas-1	*t*	Kashmir	IN	34.36N	74.48E	1	T420	*t*	Holmby†	SSW	55.45N	13.23E	
Kn-0	*t*	Kaunas	LI	54.54N	23.55E		T440	*t*	Algutsrum†	SSW	56.41N	16.31E	
Ko-3	*t*	Unknown	DK	55.40N	11.30E		T50	*t*	Kristianstad†	SSW	56.02N	14.14E	
KONDARA	*t*	Khurmatov	TA	38.00N	70.00E		T70	*t*	Lund†	SSW	55.43N	13.12E	
Le-0	*t*	Leiden	NL	52.10N	04.29E		T700	*t*	Anten†	CSW	57.59N	12.25E	1
LIMEPORT	*t*	Friedensville, PA	USA	41.00N	79.30W		T81	*t*	Karhumäki‡	RU	62.55N	34.25E	1
Lip-0	*t*	Lipowiec	PL	50.05N	19.27E	1	T93	*t*	Tvärminne‡	CFI	59.51N	23.19E	1
Ll-0	*t*	Llagostera	SP	41.50N	02.53E		Ta-0	*t*	Tabor	CZ	49.25N	14.40E	
LÖVIK	*t*	Lövik*	NSW	62.48N	18.05E		Tadjikistan	*t*	Sorbo	TA	38.49N	69.28E	
Mt-0	*t*	Martuba/Cyrenaika	LB	32.14N	22.42E		Tol-0	*t*	Toledo OH	USA	42.00N	83.00W	
MUHLEN	*t*	Mühlen	PL	53.54N	20.21E		TRÄDG-3	*t*	Trädgårdslund	SSW	56.08N	14.20E	
Na-1	*t*	Nantes	FR	47.13N	01.34W		Ts-1	*t*	Schwiegerhausen	GE	51.41N	10.12E	
Nok-0	*t*	Noordwijk	NL	52.14N	04.25E		Tsu-1	*t*	Unknown	JP	34.43N	136.30E	
Oy-0	*t*	Ostese	NO	60.23N	06.12E	1	Tu-0	*t*	Turin	IT	45.04N	07.40E	
Per-1	*t*	Perm	RU	58.00N	56.14E		Ty-0	*t*	Taynuilt	GB	56.25N	05.14E	
PETROGR	*t*	Petergof	RU	59.52N	29.54E		Wa-1	*t*	Warschau	PL	52.15N	21.00E	
Pl-0	*t*	Unknown	U	-	-		Wil-1	*t*	Wilna	LI	55.00N	25.00E	1
Pla-0	*t*	Playa de Aro	SP	41.48N	03.03E		Ws-0	*t*	Wassilewskija	RU	52.14N	29.48E	
Pn-0	*t*	Pontivy	FR	48.03N	02.58W		Yo-0	*t*	Yosemite Nat.P.	USA	38.00N	120.00W	
Po-1	*t*	Poppelsdorf	GE	50.43N	07.04E		S130	*s*	Strömsbruk†	NSW	61.53N	17.19E	1
Ra-0	*t*	Randan	FR	46.01N	03.21E		S150	*s*	Ytterhogdal†	NSW	62.10N	14.56E	1
RDL-1	*t*	Unknown	NL	52.00N	05.00E		S261	*s*	Hammarstrand§	NSW	63.07N	16.22E	1

Table 1 Continued.

Name	Species	Location	Region	Latitude	Longitude	TS23	Name	Species	Location	Region	Latitude	Longitude	TS23
S300	*s*	Sörfjärda[¶]	NSW	62.02N	17.27E	1	S60	*s*	Vännas[†]	NSW	63.55N	19.46E	1
S354	*s*	Iisalmi[‡]	CFI	63.43N	27.12E	1	S90	*s*	Västanbäck[†]	NSW	63.47N	17.05E	1
S361	*s*	Hanko[‡]	SFI	59.53N	23.07E	1							

Except in the following cases, the accessions were acquired from the Nottingham Seed Stock Centre, The SENDAI *Arabidopsis* Seed Stock Center and LEHLE: *Magnus Nordborg; [†]collected by the authors; [‡]Outi Savolainen, Oulu University; [§]Håkan Lindström, Tjälarne; [¶]Svante Holm, Mitthögskolan. AU, Austria; BE, Belgium; CA, Canada; CFI, central Finland; CI, Canary Islands; CSW, central Sweden; CVI, Cape Verde Islands; CZ, Czech Republic; DK, Denmark; ES, Estonia; FR, France; GB, Great Britain; GE, Germany; IN, India; IT, Italy; JP, Japan; LB, Libya; LI, Lithuania; NSW, north Sweden; NL, Netherlands; NO, Norway; PL, Poland; PO, Portugal; RU, Russia; SFI, south Finland; SL, Switzerland; SP, Spain; SSW, south Sweden; TA, Tajikistan; USA, United States of America and U, Unknown.

Capillary Array). The sequences obtained were aligned for each fragment and the polymorphisms were scored using the PHRED quality values (Phred-phrap package from CodonCode, Dedham, MA, USA) and the SEQUENCHER software from GeneCode (Ann Arbor, MI, USA). All polymorphic sites were inspected visually and were verified using SEQUENCHER. We also re-sequenced more that 40% of the 50 fragments (all fragments that showed variation in *A. suecica* were re-sequenced) to get an estimate on the amount of sequencing errors in the data. No discrepancies were found. If we (conservatively) assume that our sequencing error rate was 10^{-4} (Hill *et al.*, 2000), then an error rate of approximately 10^{-8} is expected for our approach where polymorphisms were confirmed on both strands. This means that for dataset TS23 (described below), where a total of approximately 10 kb DNA were sequenced in 23 accessions, we expect approximately 2×10^{-3} polymorphisms because of sequencing errors. All polymorphic sites within *A. thaliana* were submitted to The Arabidopsis Information Resource (TAIR:http://www.arabidopsis.org) with the following TAIR (polymorphism) accession numbers: 1005468356-1005468599. The term 'locus' will be used to refer to a polymorphic site and the term 'allele' will be used for the different states present at a polymorphic site.

Classification of polymorphisms and analysis of variation

The polymorphic sites detected were divided into three classes: single nucleotide polymorphisms (SNPs), insertions/deletions (indels) and simple sequence length polymorphisms (SSLPs). Polymorphic sites in nonrepetitive DNA, at which sets of bases were inserted (or deleted) in different accessions were classified as indels, polymorphic sites detected in short stretches of repeated DNA being termed SSLPs. This division was made as SSLPs can be expected to evolve at a different rate from SNPs and indels (see e.g. Li, 1997).

We computed two estimators of the well known population genetic parameter Θ ($\Theta = 2\mu N_e$, where μ is the mutation rate per generation and N_e is the effective population size). As *A. thaliana* and *A. suecica* are highly selfing (Säll *et al.*, 2004), we can assume that both species are sufficiently highly selfing that the rate of coalescence is twice that of an equivalent outbreeding population (Nordborg & Donnelly, 1997), hence the '2' in the expression for Θ. The parameter Θ can be interpreted as the expected number of mutations separating a sample of two sequences. The average number of pairwise differences (Π) and the 'Watterson' estimator Θ_W were computed for the SNPs (for details on computing Π and Θ_W see e.g. Li, 1997). The estimator Θ_W is derived from the number of differences in the sample (and correcting for sample size) in contrast to Π which estimates Θ using the allele frequencies of the sample. Tajima's test is commonly used for testing whether a locus is under selection or not. The test statistic D was computed according to the procedure described in Tajima (1989), that is, D equals the difference between Π and Θ_W divided by the square root of the variance of the difference between Π and Θ_W. The significance of D was evaluated by comparing it to a simulated distribution of D (Fu & Li, 1993) that was obtained using standard coalescent simulations (e.g. Nordborg, 2001). The Hudson & Kaplan (1985) method, also known as the 'four gamete test', was used to detect homoplasies at biallelic loci (SNPs and indels). For a pair of (biallelic) loci, one simply determines if all four possible combinations of the alleles are present among the sampled sequences. If all four allele-combinations are observed, this is taken as evidence of homoplasy. Clearly this method will underestimate the true number of homoplasies.

Datasets

Three datasets were generated. The first dataset, TS23, was generated by sequencing a total of 50 fragments (Table S1) in 15 *A. thaliana* and eight *A. suecica* accessions (Table 1). This dataset was generated primarily to date the T_{MRCA} of the *A. suecica* cp, but was also used to evaluate the proposed single origin of the cp genome of *A. suecica* (Säll *et al.*, 2003). The results of TS23 were also used to select highly variable fragments to be sequenced for the second and third set of accessions. The TS23 dataset resulted in a total of 98 polymorphisms (35 SNPs,

Table 2 Names and locations of the 45 *Arabidopsis suecica* accessions in dataset S45.

Name	Location	Region	Latitude	Longitude
S116	Ängebo	NSW	61.58N	16.20E
S122	Friggesund	NSW	61.54N	16.33E
S130	Strömsbruk	NSW	61.53N	17.19E
S141	V Indal	NSW	62.36N	17.02E
S150	Ytterhogdal	NSW	62.10N	14.56E
S163	Ytterhogdal	NSW	62.11N	14.53E
S171	Los	NSW	61.44N	15.10E
S182	Våxnan	NSW	61.41N	15.03E
S231	Olofsfors	NSW	63.34N	19.25E
S261	Hammarstrand	NSW	63.07N	16.22E
S271	Stadsforsen	NSW	62.58N	16.40E
S292	Ede	NSW	62.03N	16.51E
S300	Sörfjärda	NSW	62.02N	17.27E
S311	Stocktjärn	NSW	63.47N	20.12E
S331	Karlstad	CSW	59.23N	13.28E
S340	Kotka*	SFI	60.29N	26.55E
S354	Iisalmi*	CFI	63.43N	27.12E
S361	Hanko*	SFI	59.53N	23.07E
S370	Oulu*	CFI	65.01N	25.28E
S408	Axberg	CSW	59.22N	15.13E
S412	Grytthyttan	CSW	59.42N	14.32E
S420	Ramsberg	CSW	59.46N	15.18E
S430	Ramsnäs	CSW	59.47N	16.11E
S441	Ängelsberg	CSW	59.57N	16.01E
S459	Garpenberg	CSW	60.19N	16.12E
S460	Enviken	CSW	60.48N	15.48E
S476	Bärby	CSW	59.51N	17.49E
S485	Almunge	CSW	59.53N	18.03E
S490	Hällen	CSW	59.50N	17.34E
S500	Helsinki†	SFI	60.12N	25.03E
S510	Helsinki†	SFI	60.11N	24.59E
S520	Artjärvi†	SFI	60.44N	26.10E
S530	Pålkäms†	CFI	61.15N	24.20E
S540	Kiuruvesi†	CFI	63.37N	26.31E
S550	Pielavesi†	CFI	63.21N	26.57E
S560	Knaperåsec‡	CSW	60.39N	17.16E
S570	Oslättfors‡	CSW	60.46N	16.57E
S580	Oslättfors‡	CSW	60.46N	16.59E
S590	Hässleholm	SSW	56.09N	13.46E
S60	Vännäs	NSW	63.55N	19.46E
S600	Lund	SSW	55.42N	13.11E
S700	Ulricehamn	CSW	57.47N	13.25E
S71	Söder Nyåker	NSW	63.43N	19.21E
S81	Nordmaling	NSW	63.35N	19.28E
S90	Västanbäck	NSW	63.47N	17.05E

Except in the following cases, the accessions were collected by the authors: *Outi Savolainen, Oulu University; †Arto Kurtto, Helsinki University; ‡Peter Stål, Gävle.

CFI, central Finland; CSW, Sweden; NSW, north Sweden; SFI, south Finland; SSW, south Sweden.

12 indels and 51 SSLPs, see Table 3). Table 3 lists the exact location in the cp genome of each polymorphism detected, whether polymorphisms were located in coding regions or not and whether polymorphisms in coding regions were synonymous or nonsynonymous. Of these 98 polymorphisms, 11 overlap the polymorphisms earlier reported in Säll *et al.* (2003).

The second dataset, TS113, was generated by sequencing nine of the 50 fragments studied in dataset TS23 (Table S1) in 90 additional *A. thaliana* accessions. This much larger set of *A. thaliana* accessions was used to determine which of the *A. thaliana* accessions were most similar to the *A. suecica* accessions. In this large set of *A. thaliana* accessions we were also able to assess the level of population structure in *A. thaliana*. Dataset TS113 contained 105 (90 + 15) *A. thaliana* and eight *A. suecica* accessions (Table 1), in which a total of 27 polymorphisms were found (nine SNPs, five indels and 13 SSLPs).

The third dataset, S45, was generated by sequencing two of the 50 fragments in an additional 37 *A. suecica* accessions (Table 2; only three fragments were variable for the eight *A. suecica* accessions sequenced in TS23; Table S1). These additional *A. suecica* accessions, which cover the entire range of distribution of the species, were investigated to detect (and if possible, evaluate) population structure and isolation-by-distance in *A. suecica*. Dataset S45 contained two polymorphic sites (one indel and one SSLP) in a total of 45 *A. suecica* accessions. We attempted to sequence the third fragment, known to be variable for *A. suecica*, but we did not include data on this fragment in dataset S45 because of typing difficulties, presumably because of a very long stretch of repeated DNA in this particular fragment.

Clustering

For the accessions in each dataset, pairwise allele-sharing distances, *d*, were computed for SNPs and indels using the following expression:

$$d_{i,j} = \sum_{l} b_{i,j,l}/c_{i,j}, \qquad (1)$$

where $b_{i,j,l} = 1$ if $a_{i,l} = a_{j,l}$ and $b_{i,j,l} = 0$ if $a_{i,l} \neq a_{j,l}$, $a_{i,l}$ is the allele of the ith accession at the lth site and $c_{i,j}$ is the number of comparisons made for accession i and j. For the SSLPs, the pairwise average squared number of differences, $\delta\mu^2$ (Goldstein *et al.*, 1995) was computed. For closely related accessions in which the T_{MRCA} of the accessions is reasonably small, $\delta\mu^2$ is expected to scale linearly with time (Goldstein *et al.*, 1995; Slatkin, 1995). On the basis of these pairwise distance matrices, we obtained unrooted Neighbour-Joining trees (NJ-tree; Saitou & Nei, 1987), using the computer program NEIGHBOUR (Felsenstein, 2004, PHYLIP version 3.6, distributed by the author in question, Department of Genome Sciences, University of Washington, Seattle, WA, USA). Dataset TS113 contained a relatively small number of SNPs and indels (see Results), so in order to gain as much resolution as possible for this dataset we also constructed a combined distances matrix of SNPs, indels and SSLPs by adding the pairwise distances computed individually for each polymorphism type. For

Table 3 Summary of all the polymorphisms present in dataset TS23. The position indicates where the polymorphism is located in the *A. thaliana* chloroplast genome. If the polymorphism was located in a protein-coding region, it is noted whether the polymorphism is synonymous (S) or nonsynonymous (NS). For indels and SSLPs, the leftmost position is indicated. The type of polymorphism is indicated as being a substitution (s), an indel (i), a mononucleotide repeat (m), a dinucleotide repeat (d), a trinucleotide repeat (t) or a compound repeat (c). The possibilities are shown as being 1 or 0 for indels (present or not present). For the SSLPs the number of repeat units is noted, missing values being designated as mv.

Position	112	173	206	215	290	312	3013	3046	4114	4345	4353	4690	4747	7905	8284	8377	8403	8520	14818	14825	20227	20231	20245	25814	28673	30287	31017	31159	31247	31287	31309	31337	32290	32370
Protein coding							S	S																S										
Polymorphism type	m	m	t	s	m	i	s	s	m	m	m	d/t	d/t	m	s	i	i	s	m	m	s	m	i	s	m	m	m	m	i	s	d	s	m	m
Insertion/SSLP sequence	A	T	‡		A	CTCT			T	C	A	AT/AAT	AT/ATT	A		AAAACA	TTTAT		C	T		T	TATTA		T	A	T	A	§		AT		A	A
Accession																																		
Col-0	15	8	13	A	7	1	G	G	12	8	10	17	8	13	T	1	0	G	7	9	T	8	0	C	13	13	13	9	1	G	9	C	7	10
T160	15	7	13	A	7	1	G	A	12	8	10	mv	mv	13	T	1	1	A	7	9	T	8	mv	C	13	mv	13	9	0	G	9	C	7	10
T81	14	16	15	A	8	1	G	G	11	mv	mv	mv	mv	12	T	0	1	G	8	9	C	8	0	G	11	mv	13	10	1	G	9	C	7	9
T93	15	8	13	A	7	1	G	G	12	7	10	mv	mv	13	T	0	1	G	7	9	T	8	0	C	11	10	13	10	1	G	9	C	7	11
T104	15	8	13	A	7	1	G	G	11	8	9	17	8	12	T	1	0	G	7	9	T	8	0	C	13	14	12	9	1	G	9	C	7	10
Oy-0	15	8	13	A	7	1	G	G	12	8	15	17	8	12	T	0	1	G	7	9	T	9	0	C	10	10	12	10	1	G	9	C	7	11
T10	16	7	13	A	7	0	A	A	12	8	10	17	8	12	T	1	0	A	7	9	T	8	0	C	18	14	13	9	0	G	9	C	7	10
T20	14	22	13	A	9	0	G	G	11	7	11	15	8	12	T	0	1	G	7	9	T	8	0	G	10	14	13	10	1	G	9	G	5	9
T700	15	7	13	A	7	0	A	A	12	mv	mv	17	8	12	T	1	0	A	7	9	T	8	0	C	18	14	13	9	0	G	9	C	7	10
T340	14	20	13	A	9	0	G	G	11	7	11	15	8	12	T	0	1	G	7	9	T	8	0	G	10	14	13	10	1	G	9	G	5	10
T350	15	7	13	A	7	1	G	A	12	8	10	17	10	13	T	1	1	A	8	9	T	8	0	C	14	14	13	9	0	G	9	C	7	10
Kas-1	15	8	13	A	7	1	G	G	12	8	10	16	8	12	T	1	1	G	8	8	T	8	0	C	13	14	13	9	1	G	9	C	7	10
Lip-0	mv	8	13	C	7	1	G	G	12	8	10	18	8	13	T	1	1	G	7	9	T	8	0	C	11	14	13	9	1	G	9	C	7	10
Sv-0	13	9	13	A	7	1	G	G	11	mv	mv	15	8	12	T	0	0	G	8	9	T	8	0	C	13	12	13	10	1	G	9	C	7	11
Wil-1	13	15	13	A	7	1	G	G	10	9	10	15	8	12	G	0	1	G	7	9	T	8	1	C	11	14	12	10	1	C	10	C	7	9
Bu-0	14	21	13	A	9	0	G	G	10	7	11	15	mv	12	T	0	1	G	7	9	T	8	0	G	10	14	13	10	1	G	9	G	5	10
S60	12	19	13	A	8	1	G	G	12	8	10	15	8	12	T	0	1	G	7	9	T	8	0	G	11	14	13	10	1	G	12	C	7	9
S90	12	19	13	A	8	1	G	G	12	8	10	15	8	12	T	0	1	G	7	9	T	8	0	G	11	14	13	10	1	G	9	C	7	9
S130	12	19	13	A	8	1	G	G	12	8	10	15	8	12	T	0	1	G	7	9	T	8	0	G	11	mv	13	10	1	G	9	C	7	9
S150	12	19	13	A	8	1	G	G	12	8	10	15	8	12	T	0	1	G	7	9	T	8	0	G	11	mv	13	10	1	G	12	C	7	9
S261	12	19	13	A	8	1	G	G	12	8	10	15	8	12	T	0	1	G	7	9	T	8	0	G	11	14	13	10	1	G	9	C	7	9
S300	12	20	13	A	8	1	G	G	12	8	10	15	8	12	T	0	1	G	7	9	T	8	0	G	11	14	13	10	1	G	9	C	7	9
S354	12	19	13	A	8	1	G	G	12	8	10	15	8	12	T	0	1	G	7	9	T	8	0	G	11	mv	13	10	1	G	9	C	7	9
S361	12	19	13	A	8	1	G	G	12	8	10	15	8	12	T	0	1	G	7	9	T	8	mv	G	11	14	mv	10	1	G	9	C	7	9

Table 3 Continued

Position	36693	42409	42475	42550	42590	45196	49949	49959	49972	50846	50858	56730	56758	56767	58592	58611	59387	59540	63122	63164	63178	63231	63358	66308	66700	67059	67164	68340	69368	70189	71178	71179	76735	77830
Protein coding					NS																													
Polymorphism type	m	s	s	i	s	m	m	m	s	m	i	s	m	s	s	m	m	s	d	m	s	i	s	m	m	s	m	m	s	m	m	s	m	s
Insertion/SSLP sequence	A			AGAAT		T	T	T		A	AAAG		T			T	T		AT	T		GATTC		T	T		G	A		A	T		A	
Accession																																		
Col-0	10	C	T	1	G	11	8	13	A	11	1	T	9	A	G	9	11	G	8	13	A	0	A	13	16	T	13	11	G	13	11	T	12	G
T160	10	C	T	1	G	11	8	13	A	11	1	T	9	A	G	9	11	G	8	12	A	0	A	12	11	T	10	12	G	12	mv	mv	11	G
T81	10	C	T	1	G	16	9	13	A	11	0	T	10	A	T	9	12	G	9	12	A	0	A	13	12	T	9	mv	G	9	10	T	12	G
T93	10	C	T	0	G	16	8	13	A	11	1	T	9	T	T	9	10	G	8	11	A	0	A	13	10	T	10	12	G	14	11	T	12	G
T104	10	C	T	1	G	11	8	13	A	11	1	T	9	A	G	9	10	G	6	12	A	0	A	13	15	T	11	12	G	13	11	T	11	G
Oy-0	10	C	T	0	G	16	8	13	A	11	1	T	9	T	T	9	10	G	8	12	A	0	A	13	11	T	13	mv	G	14	12	T	13	G
T10	10	C	T	1	A	11	8	13	A	11	1	T	9	A	G	9	mv	mv	8	11	A	0	A	13	11	T	9	12	G	12	11	T	11	G
T20	10	C	T	1	G	16	9	13	T	10	0	T	9	A	T	9	11	mv	9	12	A	0	C	13	11	T	11	12	G	8	11	T	12	G
T700	10	C	T	1	A	11	8	13	A	11	1	T	9	A	G	9	11	G	8	11	A	0	A	13	11	T	9	12	G	12	mv	mv	11	G
T340	10	C	T	1	G	16	9	13	T	10	0	T	9	A	T	9	11	G	9	12	A	0	C	13	11	T	10	12	G	8	11	T	12	G
T350	10	C	T	1	G	11	mv	mv	A	11	1	T	9	A	G	9	11	G	8	12	A	0	A	13	10	T	10	7	G	mv	11	T	11	G
Kas-1	10	C	T	1	G	10	8	13	A	11	1	T	9	A	G	9	11	G	8	12	A	0	A	13	14	T	11	mv	G	13	11	T	12	A
Lip-0	10	T	T	1	G	11	mv	mv	A	11	1	T	9	A	G	9	11	G	7	11	A	0	A	14	16	C	11	12	G	13	mv	mv	12	G
Sv-0	11	C	T	0	G	mv	8	12	A	10	1	T	9	A	T	8	10	G	8	10	A	0	A	13	13	T	9	12	G	13	11	A	13	G
Wil-1	10	C	C	1	G	mv	9	13	A	11	1	T	9	A	T	9	11	A	9	12	T	0	A	13	16	T	10	mv	T	13	11	T	12	G
Bu-0	10	mv	T	1	mv	16	9	13	T	10	0	T	9	A	T	9	11	G	9	12	A	0	C	13	mv	T	10	12	G	8	11	T	12	G
S60	10	C	T	1	G	16	mv	mv	A	11	0	G	9	A	T	9	11	G	11	12	A	0	A	13	12	T	10	mv	G	9	11	T	12	G
S90	10	C	T	1	G	16	mv	mv	A	11	0	G	9	A	T	9	11	G	11	12	A	1	A	13	12	T	10	12	G	9	11	T	12	G
S130	10	C	T	1	G	16	9	13	A	11	0	G	9	A	T	9	11	G	11	12	A	1	A	13	12	T	10	mv	G	9	mv	mv	12	G
S150	10	C	T	1	G	16	9	13	A	11	0	G	9	A	T	9	11	G	11	12	A	0	A	13	12	T	10	12	G	9	11	T	12	G
S261	10	C	T	1	G	16	9	13	A	11	0	G	9	A	T	9	11	G	11	12	A	1	A	13	12	T	10	12	G	9	11	T	12	G
S300	10	C	T	1	G	16	mv	mv	A	11	0	G	9	A	T	9	11	G	11	12	A	0	A	13	12	T	10	12	G	mv	11	T	12	G
S354	10	C	T	1	G	16	9	13	A	11	0	G	9	A	T	9	11	G	11	12	A	0	A	13	12	T	10	12	G	9	11	T	12	G
S361	10	C	T	1	G	16	9	13	A	11	0	G	9	A	T	9	11	G	11	12	A	1	A	13	12	T	10	12	G	9	11	T	12	G

Table 3 Continued

Position	77894	81879	82454	82601	83880	83986	96023	99364	108115	112698	112708	113220	113236	113265	113339	114027	114168	114256	115009	115647	119064	119131	119145	119153	119205	120675	120786	124152	124321	124355
Protein coding					NS														NS											
Polymorphism type	i	s	m	m	s	m	s	m	m	d	m	s	s	s	t/d	s	i	s	c	m	m	i	t	s	i	m	m	s	s	s
Insertion/SSLP Sequence	†		T	T		T		A	A	AT	T				TA/TAA		*		T/A	T	T	AAAA	TAT		A	A	A			
Accession																														
Col-0	0	A	11	10	C	8	T	17	10	5	9	T	G	C	5	T	1	G	9/6	10	11	0	11	A	0	11	10	A	T	G
T160	1	A	10	10	C	8	T	mv	9	5	9	T	G	C	5	T	1	G	9/6	9	12	1	11	A	0	11	10	C	T	G
T81	0	A	11	9	C	9	T	17	9	6	8	A	G	C	5	T	0	G	10/5	10	9	0	9	A	0	11	10	A	T	G
T93	0	T	12	11	T	8	T	11	9	5	9	T	G	C	5	T	1	G	9/6	10	10	0	8.7	A	0	12	10	A	T	G
T104	1	A	11	9	C	8	T	17	9	5	9	T	G	C	5	T	1	G	9/6	10	10	0	11	A	0	11	10	A	T	G
Oy-0	0	T	12	10	C	8	T	11	9	5	9	T	G	C	5	T	1	G	9/6	10	10	0	8.7	A	0	12	10	A	T	C
T10	1	A	11	10	C	8	T	17	9	5	9	T	G	C	7	T	1	G	9/6	10	mv	mv	mv	mv	mv	12	10	A	T	G
T20	0	A	12	10	C	9	T	17	9	6	9	A	G	C	5	T	0	G	9/6	mv	10	0	9	A	1	11	10	A	G	G
T700	1	A	10	10	C	8	T	17	10	5	9	T	G	C	7	T	1	G	9/6	10	12	1	11	A	0	12	10	A	T	G
T340	0	A	12	10	C	9	T	17	9	6	10	A	G	C	5	T	0	G	9/6	10	10	0	9	A	1	11	10	A	G	G
T350	1	A	10	10	C	8	T	17	9	5	10	T	G	C	5	T	1	G	9/6	10	mv	mv	mv	mv	mv	11	10	A	T	G
Kas-1	0	A	10	10	C	9	G	17	9	5	8	T	G	C	5	T	1	G	10/5	10	11	0	11	A	0	11	10	A	T	G
Lip-0	0	A	11	10	C	mv	T	17	9	5	9	T	G	C	5	T	1	G	9/6	10	mv	mv	mv		mv	12	9	A	T	G
Sv-0	0	T	11	mv	C	8	T	mv	10	6	8	T	A	T	5	T	1	A	9/6	10	9	0	11	C	0	11	10	A	T	G
Wil-1	0	A	11	mv	C	8	T	17	9	5	8	A	G	C	5	A	1	G	9/6	10	10	0	11	A	0	11	11	A	T	G
Bu-0	0	A	12	11	C	9	T	17	9	6	10	A	G	C	5	T	0	G	9/6	10	10	0	9	A	1	11	10	A	G	G
S60	0	A	12	10	C	9	T	13	9	6	10	A	G	C	5	T	0	G	9/6	10	10	0	9	A	0	11	10	A	T	mv
S90	0	A	12	10	C	9	T	13	9	6	10	A	G	C	5	T	0	G	9/6	10	mv	mv	mv	A	mv	11	10	A	T	G
S130	0	A	12	10	C	9	T	13	9	6	10	A	G	C	5	T	0	G	9/6	10	mv	mv	mv	A	mv	11	10	A	T	G
S150	0	A	12	10	C	9	T	13	9	6	10	A	G	C	5	T	0	G	9/6	10	10	0	9	A	0	11	10	A	T	G
S261	0	A	12	10	C	9	T	13	9	6	10	A	G	C	5	T	0	G	9/6	10	10	0	9	A	0	11	10	A	T	G
S300	0	A	12	mv	C	9	T	13	9	6	10	A	G	C	5	T	0	G	9/6	mv	10	mv	mv	A	mv	11	10	A	T	G
S354	0	A	12	10	C	9	T	13	9	6	10	A	G	C	5	T	0	G	9/6	10	10	0	9	A	0	11	10	A	T	G
S361	0	A	12	mv	C	9	T	13	9	6	10	A	G	C	5	T	0	G	9/6	10	10	0	9	A	0	11	10	A	T	G

*AATTTTAAAAATCAATTGGAAAT.
†TTTTTTTTCTA.
‡TTA/AAT.
§ATTCTCA.

this combined distance matrix we chose to give SNPs and indels a weight of 10 relative to SSLPs. This (arbitrary) weighting was carried out to compensate, at least to some degree, for the expected differences in mutation rate of these polymorphism types. We also tested other weights (e.g. 2 and 100) for which the results were almost identical to the results acquired when only using the SSLPs and the SNPs and indels, respectively. From this combined pairwise distance matrix we obtained an unrooted NJ-tree. Finally, an unrooted NJ-tree was obtained in a similar way for all 32 Swedish accessions of *A. thaliana* and all eight *A. suecica* accessions in dataset TS113.

Population structure

F_{ST} was computed, based on the groups in question, from the total average pairwise difference for all pairs of accessions, $\hat{\Pi}_{total}$ and the mean of the average pairwise difference within each group, $\hat{\Pi}_{within}$ (Hudson *et al.*, 1992):

$$F_{ST} = 1 - \left(\frac{1}{K} \sum_{i=1}^{K} \hat{\Pi}_{within_i} \right) / \hat{\Pi}_{total} \qquad (2)$$

where K is the number of groups. R_{ST} was computed according to Slatkin (1995) for SSLPs.

Correlations between the genetic and the geographic distance were used to infer isolation-by-distance. The genetic pairwise distances were computed and combined to one pairwise distance matrix for the accessions in question as described above. The pairwise geographic distance between each pair of accessions was computed from the geographical locations of the different accessions (Tables 2 and 3). The correlation between the genetic distance and the geographical distance was calculated for each pairwise comparison of accessions, using the Spearman rank correlation coefficient (Sokal & Rohlf, 1995). The Mantel test was used to determine the level of significance of the correlations (Mantel, 1967).

Estimation of T_{MRCA}

A probabilistic analysis employing a Markov Chain Monte Carlo algorithm (MCMC, Wilson & Balding, 1998) was conducted using the computer software BATWING (Wilson *et al.*, 2003). The models used in the analysis were the standard coalescent, which assumes a constant population size (cpc), a coalescent model with constant exponential population growth (cgc) and a coalescent model with late exponential population growth (lgc). Several values of the growth parameter N_c/N_a (the ratio of the current to the ancestral population size) were tested for both population growth models and for the lgc-model different values of the start-of-growth were tested. The Stepwise Mutation Model (SMM) was used for SSLPs and the Infinite Sites Model for SNPs. The indels were not used in this analysis as homoplasy was found among them (see the results). Only perfect repeats with an average length >7 (44 loci) were considered as being SSLPs, as SSLPs of this type are more likely to be similar in their mutation rates and mutation processes. Two MCMC chains were run for 10 000 samples, with an initial burn-in of 2000 samples. A total of 40 branch-swapping steps between successive attempts to update Θ were proposed, 100 attempts to update Θ being made between successive samples. A uniform prior distribution (0,100) was used for Θ. For each chain, the posterior distributions of Θ, T_{MRCA} and the total tree length G were checked for nonconvergence. Finally, the samples from the two chains were combined to give a total of 20 000 samples (corresponding to 8×10^7 branch-swapping attempts). The posterior distributions of the combined data were then used to calculate the mean, the median and the 95% credibility interval for Θ and T_{MRCA}.

This MCMC approach is not feasible for estimating the T_{MRCA} of the *A. suecica* cp as there were only two variable SSLPs and no SNPs among the 8 *A. suecica* accessions. The linear relationship between the average pairwise difference of repeat lengths of SSLPs, S, and T_{MRCA} (Slatkin, 1995) was instead used to estimate the T_{MRCA} of the *A. suecica* cp from estimates of the T_{MRCA} of the *A. thaliana* cp. To be able to do that, we have to make two assumptions. First we assume that the mutation rate is the same in both *A. suecica* and *A. thaliana*. This may not be true, but all thoroughly investigated species in the *Arabidopsis* genera seem to have similar mutation rates (Koch *et al.*, 2000). Second, we assume that the SSLPs evolve under the SMM, which is also assumed by BATWING (Wilson & Balding, 1998). Other mutational models of microsatellites are available, but we choose to use the same assumption as in BATWING for consistency. Unless the cpSSLPs truly evolves under the SMM, estimates of the divergence time will be biased when the SMM is used for inference and the bias will depend on which mutational model that actually governs the cpSSLPs.

In order to estimate the T_{MRCA} we use (Slatkin, 1995; eqn 11b):

$$E(S) = 2\mu\bar{t}, \qquad (3)$$

where $\bar{t}$ is the average coalescence time and $E(S)$ is the expected value of S (eqn 3 holds for a SMM with a mean of 0 and a variance of 1). By substituting $E(S)$ and $\bar{t}$ in eqn 3 by estimates of S and T_{MRCA}, dividing eqn 3 for an *A. thaliana* sample by eqn 3 for an *A. suecica* sample, the following expression is obtained:

$$\hat{T}_{MRCA,s} = \hat{T}_{MRCA,t} \times \hat{S}_s / \hat{S}_t, \qquad (4)$$

where $\hat{T}_{MRCA,s}$ is an estimate of T_{MRCA} for an *A. suecica* sample, $\hat{T}_{MRCA,t}$ is an estimate of T_{MRCA} for an *A. thaliana* sample, $\hat{S}_s$, is an estimate of S for an *A. suecica* sample and $\hat{S}_t$, is an estimate of S for an *A. thaliana* sample. Equation 4 can be used to estimate the T_{MRCA} of the

A. suecica cp. Note that $\hat{T}_{\mathrm{MRCA},t}$ and $\hat{T}_{\mathrm{MRCA},s}$ are scaled in the population sizes of *A. thaliana* and *A. suecica*, respectively. To get an estimate of the T_{MRCA} scaled in generations or years (both species appears to be annuals), we have to know, or rather make assumptions about the population sizes of the two species.

A second approach to estimating the T_{MRCA} of the *A. suecica* cp stems from the observation that there were no SNPs found in the *A. suecica* sample. Assume that the mutation rate μ is the same for the *A. suecica* cp as for the *A. thaliana* cp. For a particular site, the probability p of not observing a mutation in T generations (=years in *A. suecica* and in *A. thaliana*) for a sample that have the genealogy G (scaled in units of T_{MRCA}) is:

$$p = (1-\mu)^{TG}.$$

Then for L bp, the probability of not observing any mutations is:

$$p = (1-\mu)^{TGL}.$$

By setting $p = 0.05$ and solving for T, we have:

$$T = \ln 0.05/[\ln(1-\mu) \times G \times L]. \quad (5)$$

This can be considered as the upper limit of a 95% probability interval of not observing any mutations (SNPs) in a sample with the genealogy G. The lower extreme of G is 2 (when excluding the case of one branch only), which would be the result in a tree with only two branches. Such a tree would appear if all external nodes (accessions) fell into two groups, and the within-group-distances are zero. For eight accessions, the upper extreme is $G = 8$, which would be the result of a star-like tree, and for a neutral tree is $\mathrm{E}(G) = 5.44$ (e.g. Li, 1997).

A third approach estimating the T_{MRCA} of the *A. suecica* cp is based on the idea that it will take a certain amount of time for variation to accumulate in a species when starting from a very low level. The growth in diversity can be calculated by use of the following recursive formula (see e.g. Gillespie, 1998):

$$H_t = H_{t-1} - H_{t-1}/2N_e + 2\mu(1 - H_{t-1}), \quad (6)$$

where H_t is the gene diversity at time t, N_e is the effective population size and μ is the total mutation rate. If N_e increases to a stable level rapidly, this process can be described in approximate terms by the following expression (Malécot, 1948; Nei *et al.*, 1975):

$$H_t = 2N_e\mu\{1 - \exp[-t(2\mu + 1/N_e)]\}/(2N_e\mu + 1). \quad (7)$$

The probability of no variation being observed in a series of eight accessions of *A. suecica* is:

$$p = \sum_i q_i^8, \quad (8)$$

where q_i is the haplotype-frequency of haplotype i, assuming that the entire sequence acts as a single nonrecombining unit.

Results

Levels and pattern of variation in *A. thaliana* and *A. suecica*

For 15 *A. thaliana* and eight *A. suecica* accessions (dataset TS23), a total of 9699 bp were sequenced (Table 3), representing 6.3% of the *A. thaliana* and *A. suecica* cp genome. Among the 15 *A. thaliana* accessions, a total of 35 SNPs were found, of which 20 were singletons (note that we only expect approximately 2×10^{-3} mutations to be caused by sequencing errors in this dataset). Five of the SNPs were located in coding regions, one of them (position 83880; Table 3) changes the aminoacid sequence of the proteins and one (position 42590; Table 3) changes an aminoacid to a stop-codon (and thereby terminating the protein two amino acids earlier than the other allele). The average number of nucleotide differences Π was 7.24 (0.75×10^{-3} per bp) and Θ_W was 10.20 (1.05×10^{-3} per bp). Tajima's D was −1.20 for the 15 *A. thaliana* accessions, which was nonsignificant ($p > 0.05$). Thus, we found no evidence for an effect of selection on the cp genome. No evidence of homoplasy was detected among 105 comparisons, i.e. the SNPs behaved strictly as unique events polymorphisms (UEP).

A total of 11 indels were found among the 15 *A. thaliana* accessions, ranging from 1 to 23 bp in length (with an average of 6.9 bp). Only one of them was a singleton (Table 3). All of these indels were located in noncoding regions. The average gene diversity (H_e; Nei, 1987) on the basis of the indels was 0.37. Two homoplasies were found among the indels (out of 45 comparisons).

In a study of the mutational dynamics of microsatellites in the cp genome of *A. thaliana* (M. Jakobsson, T. Säll, C. Lind-halldén & C. Halldén, unpublished data), the level of variation was found to be dependent upon the number of repeat units. In particular, we found very low levels of variation for loci with <7 repeat units. Accordingly, only SSLP loci with a mean number of repeat units >7 were considered in the following analysis (Table 3). Of such loci, 72% (47 of 65) were variable, with an average H_e of 0.30 and an average of 2.4 alleles per SSLP locus. When only loci with a mean number of repeat units >10, a commonly used cut-off point, were considered, 91% (29 of 32) of the loci were found to be variable. The average H_e was 0.43 and the average number of alleles 3.1. All but one SSLP were located in noncoding regions. The SSLP located in a coding region was a compound of two SSLPs (position 115009; Table 3) which changes the aminoacid of a protein, but not the reading frame as the two alleles found have the same length.

A total of three variable sites were found among the eight *A. suecica* accessions in dataset TS23, two of them were SSLPs and one was an indel. Thus, no SNP was found. The indel locus, as described by Säll *et al.* (2003), involved a 5 bp difference between the two alleles. The

two SSLP loci each had two alleles among the *A. suecica* accessions. One of these SSLP loci was a mononucleotide repeat in which the two alleles were 19 and 20 repeat units long. The other SSLP locus was a dinucleotide repeat in which the alleles were 9 and 12 repeat units long. When only loci with a mean number of repeat units >10 were considered, 6% (two of 32) of the loci were found to be variable, the average H_e was 0.02, and the average number of alleles was 1.06.

Phylogeny of the *A. thaliana* and *A. suecica* cp chromosome

Two NJ-trees based on dataset TS23 were constructed, one for SNPs and indels and the other for SSLPs (Fig. 1a and b). The first tree was based on 35 SNPs and 12 indels, (eight and four, respectively, overlapping with those reported by Säll *et al.*, 2003) and the second tree was based on 51 loci (none of which overlapped with those reported by Säll *et al.*, 2003). Both trees showed that the *A. suecica* accessions clustered together; the bootstrap support was, however, relatively weak (63% in both cases).

To search more widely for the *A. thaliana* accessions that are most genetically similar to *A. suecica*, an additional 90 *A. thaliana* accessions were typed for nine fragments, resulting in 27 polymorphic sites in a total of 105 *A. thaliana* and eight *A. suecica* accessions (dataset TS113). Note that we chose here to type two of the variable sites found in *A. suecica*, which leads to an overestimation of the genetic differences among *A. suecica* accessions. Then NJ-trees were obtained based on SNPs and indels, SSLPs and combining all of them. We here present the result of the combined data as there were relatively few polymorphisms of each individual type of polymorphism. The NJ-tree based on the combined data showed all the *A. suecica* accessions to cluster in one part of the tree together with 17 *A. thaliana* accessions (Fig. 2). These *A. thaliana* accessions (An-1, Blh-1, Bu-0, EDEN, EDSÄTER, Enkheim-1, Gr-1, Le-0, Nok-0, Rdl-1, Rsch-0, Sei-0, T120, T190, T20, T340 and T50) were mainly from central Europe (southern Sweden, Germany, the Netherlands, the Czech Republic, Austria, Belgium and northern Italy), but there were also two accessions from northern Sweden and one from western Russia. However, there was no strong bootstrap support of this cluster and central European *A. thaliana* accessions were also found outside this cluster. The trees based on either SNPs and indels or SSLPs showed similar patterns as the combined data, but the resolution was much lower for these trees.

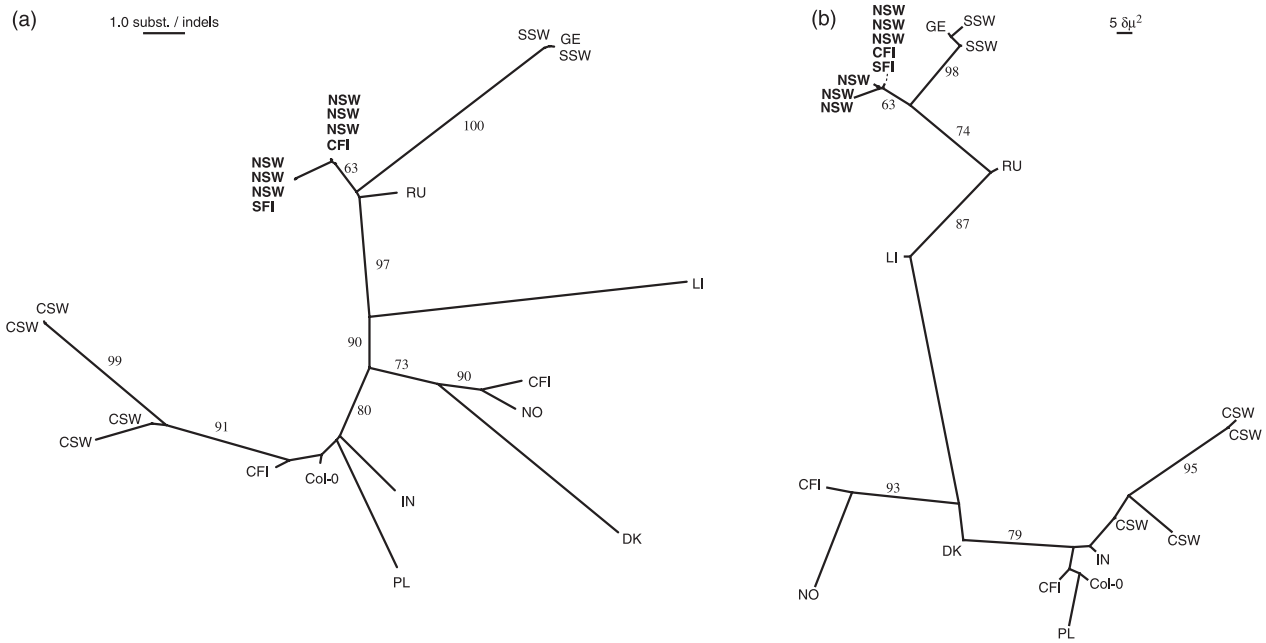

Fig. 1 Neighbour-Joining trees of 16 *Arabidopsis thaliana* and eight *Arabidopsis suecica* (bold face) accessions. The geographical origin of the accessions is indicated in the figure by the following notation: CFI, central Finland; CSW, central Sweden; DK, Denmark; GE, Germany; IN, India; LI, Lithuania; NSW, north Sweden; NO, Norway; PL, Poland; RU, Russia; SFI, south Finland; SSW, south Sweden (see also Table 1). Bootstrap values (%) with a support >70% are indicated beside the branches. The bootstrap values for the branches leading to the *A. suecica* accessions are also indicated. The trees were constructed from (a) 35 SNPs and 12 indels and (b) 51 SSLPs. Accession Col-0 is the sequence in GenBank (accession number NC_000932).

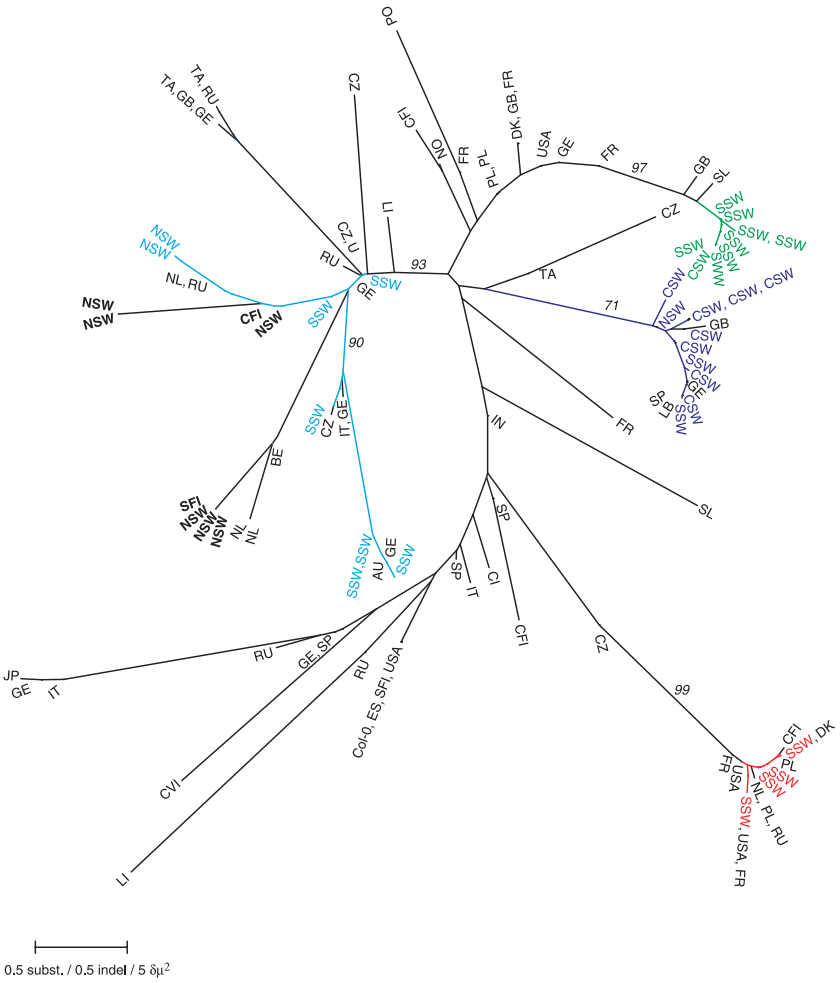

Fig. 2 Neighbour-Joining trees of 105 *Arabidopsis thaliana* and eight *Arabidopsis suecica* (boldface) accessions as based on the combination of three different types of polymorphic sites (nine SNPs, five indels and 13 SSLPs; see Methods). The geographical origin of the accessions is indicated in the figure by the following notation: AU, Austria; BE, Belgium; CA, Canada; CFI, central Finland; CI, Canary Islands; CSW, central Sweden; CVI, Cape Verde Islands; CZ, Czech Republic; DK, Denmark; ES, Estonia; FR, France; GB, Great Britain; GE, Germany; IN, India; IT, Italy; JP, Japan; LB, Libya; LI, Lithuania; NSW, north Sweden; NL, Netherlands; NO, Norway; PL, Poland; PO, Portugal; RU, Russia; SFI, south Finland; SL, Switzerland; SP, Spain, SSW, south Sweden; TA, Tajikistan; USA, United States of America and U, Unknown. (see also Table 1). Bootstrap values (%) with a support >70% are indicated beside the branches. Accession Col-0 is the sequence in GenBank (accession number NC_000932). The 32 Swedish accessions cluster into four groups (light-blue, green, dark-blue and red; see also Figs 3 and 4).

Geographic distribution of variation

To estimate the level of population structure in *A. thaliana*, the worldwide sample of *A. thaliana* accessions (dataset TS113) was divided into geographical groups. Nordborg *et al.* (2005) found strong evidence for population structure in *A. thaliana* on a worldwide scale as well as within Sweden. We therefore assigned the accessions to nine geographical groups which were defined by longitudes 54N and 43N and latitudes 20E and 40E. After removing the accessions from North America and Japan as *A. thaliana* was presumably introduced there very recently (Jørgensen & Mauricio, 2004; Nordborg *et al.*, 2005), six groups contained more than three accessions. For these six groups were $F_{ST} = 0.22$ (based on SNPs) and $R_{ST} = 0.23$ (based on SSLPs). We also found a small, yet clearly significant, correlation between geographic and genetic distance for *A. thaliana* ($r_S = 0.079$, $p = 0.0010$), when the accessions from North America and Japan were excluded.

The subset including the 32 Swedish *A. thaliana* accessions from dataset TS113 (the Finnish *A. thaliana* accessions in dataset TS113 were excluded as there were only four of them) has a geographical distribution similar to that of the *A. suecica* sample in S45. These *A. thaliana* accessions, which represent most of the species range in Sweden, were divided into three groups on the basis of longitude: North Sweden (NSW) > 61N, 61N > Central Sweden (CSW) > 57N, South Sweden (SSW) < 57N. F_{ST} was 0.50 and R_{ST} was 0.52 for these three groups which means that about half of the variation is between these three groups that are oriented along a north-south axis. This fact indicates that *A. thaliana* is structured geographically in a north-south direction in Sweden. Note also that these F_{ST}- and R_{ST}-values are much higher than F_{ST} and R_{ST} of the worldwide sample.

If we construct a NJ-tree (and collapse all braches with bootstrap support <65%) of the same 32 accessions together with all the *A. suecica* accessions, a clustered pattern appears that, to some degree, corresponds to geography. Except for eight accessions (light blue in Fig. 3), all the *A. thaliana* accessions from Sweden appeared in three distinct clusters. The first cluster (green in Figs 3 and 4) included eight accessions from SSW

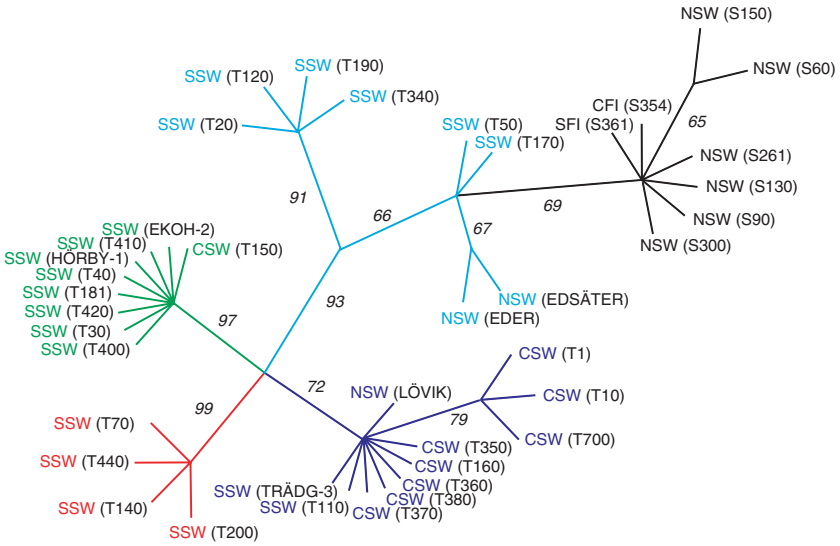

Fig. 3 Schematic Neighbour-Joining tree of 32 Swedish *Arabidopsis thaliana* and eight Swedish and Finnish *Arabidopsis suecica* accessions (black) as based on the combination of different types of polymorphic sites (nine SNPs, five indels and 13 SSLPs; see methods). Branches with a bootstrap support <65% have been collapsed. Branches with a bootstrap support ≥65% are indicated beside the branch. The geographical origin of the accessions is indicated in the figure by the following notation: CFI, central Finland; CSW, central Sweden; NSW, north Sweden; SFI, south Finland; SSW, south Sweden (see also Table 1). The *A. thaliana* accessions clustered into four groups (light-blue, green, dark-blue and red; see also Fig. 4).

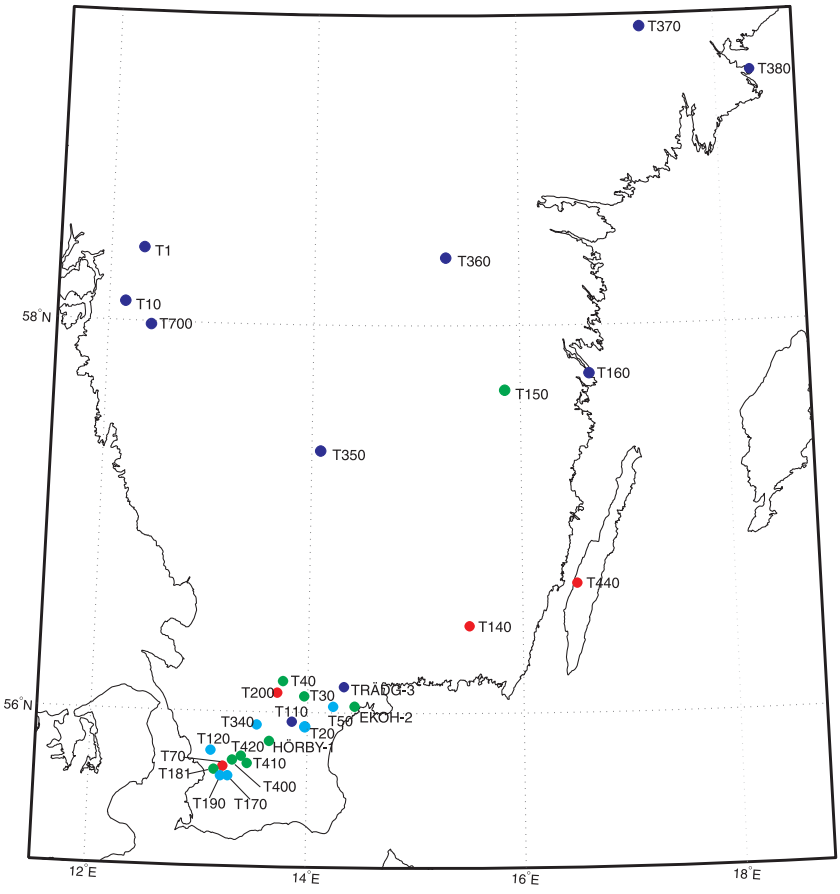

Fig. 4 The Swedish *Arabidopsis thaliana* accessions cluster into four main groups. The coloured circles correspond to the coloured accessions in Figs 2 and 3. Note that the three northern Swedish accessions (two light-blue and one dark-blue) are located outside this map.

(EKOH-2, HÖRBY-1, T181, T30, T40, T400, T410 and T420) and one CSW accession (T150). The second cluster (dark blue in Figs 3 and 4) included eight accessions from CSW (T1, T10, T160, T350, T360, T370, T380, T700), two from SSW (T110 and TRÄDG-3) and one from NSW (LÖVIK), whereas the third cluster (red in Figs 3 and 4) contained four accessions from SSW (T70, T200, T140 and T440) where the latter two accessions were both

from the south-east of Sweden (Fig. 4). The fourth cluster (light blue in Figs 3 and 4) contained six accessions from SSW (T120, T170, T190, T20, T340 and T50) and two from NSW (EDEN and EDSÄTER), yet this cluster was not as distinct as the other ones, and within the cluster there were two branches with bootstrap support of 91% and 66%, respectively (Fig. 3). Note that all *A. suecica* accessions fall into this cluster. To further investigate the pattern of geographical structure indicated in Figs 3 and 4, we calculated the correlation between geographic and genetic distance (the combined distance of all three types of polymorphisms as described in the methods) for the 32 *A. thaliana* accessions from Sweden. A moderate, but clearly significant, correlation was observed ($r_S = 0.141$, $p = 0.0020$). Thus, on this fairly small geographical scale *A. thaliana* was clearly structured into partly overlapping populations that each occupied one or two regions.

Two of the three fragments known to be variable in *A. suecica* were sequenced for an additional 37 *A. suecica* accessions (dataset S45; see Material and methods) resulting in two polymorphic sites. The set of 45 studied accessions were chosen to effectively cover the entire range of the species (Hultén, 1971; T. Säll, unpublished data). The two variable loci were both biallelic and there are thus four possible haplotypes for these two loci. However, only three haplotypes were found among the 45 *A. suecica* accessions and these three haplotypes were represented in both the Swedish and the Finnish accessions with no obvious geographical pattern discernable.

T_{MRCA} of *A. suecica* and *A. thaliana*

It is worth pointing out that the goal of this analysis is to date the time-of-origin of *A. suecica*. However, we can only hope to infer the lower bound of the time-of-origin because the T_{MRCA} of the *A. suecica* cp may occur anytime between the present and the time-of-origin. To estimate the T_{MRCA} of *A. thaliana* we used the MCMC method of Wilson & Balding (1998). To infer a T_{MRCA} the *A. suecica* cp we used the linear relationship between time and the average length difference among alleles at microsatellites evolving under a strict SMM (Slatkin, 1995). This approach was chosen because there was almost no variation in *A. suecica* and using the approach of Wilson & Balding (1998) would therefore not be feasible. The data were analysed in three different model settings: a standard coalescent model (constant population size), a coalescent model with constant exponential population growth, and a coalescent model with late exponential population growth. Using the SNPs, the results for *A. thaliana* were very similar for the different models and almost completely unaffected model settings. The mean of T_{MRCA} equals 0.97 for the standard coalescent model, with a 95%-credibility region (0.50, 1.76), scaled in N_e (Table 4). The MCMC analysis also provides estimates of Θ (Table 4). The mean Θ was 12.8, with a 95% credibility region (6.9, 21.8) for the standard coalescent model. Note that Π and Θ_W presented above (7.24 and 10.20) were clearly within this credibility interval.

Table 4 Estimates of the time to the most recent common ancestor (T_{MRCA}) and Θ of *Arabidopsis thaliana* based on SNPs or microsatellites (SSLPs) using the computer software BATWING (Wilson & Balding, 1998). The parameters were estimated under three different scenarios, constant population size (cpc), constant exponential population growth (cgc) and late exponential population growth (lgc). For the cgc-model and the lgc-model was the growth parameter $N_c/N_a = 10\,000$ (the ratio of the current to the ancestral population size) and for the lgc-model, the growth was limited to the last tenth of the T_{MRCA}. The values given are means and in parenthesis, the 2.5% quantiles and the 97.5% quantiles.

Model	Polymorphism	T_{MRCA}	Θ
cpc	SNPs	0.973 (0.499, 1.763)	12.76 (6.85, 21.82)
cgc	SNPs	0.979 (0.507, 1.765)	12.67 (6.87, 21.58)
lgc	SNPs	1.036 (0.546, 1.860)	11.86 (6.55, 19.71)
cpc	SSLPs	0.613 (0.304, 1.071)	3.74 (2.05, 6.70)
cgc	SSLPs	0.293 (0.206, 0.383)	6.89 (5.02, 9.68)
lgc	SSLPs	0.718 (0.387, 1.189)	3.09 (1.84, 5.13)

By contrast, estimates based on the 44 variable SSLP loci differed for the three models and the results were affected by the model settings (Table 4). For a growing population the estimate of T_{MRCA} was shorter than for a constant population size model, whereas the estimate of Θ was higher than for a constant population size model. When the growth was initiated at a late stage (close to the present) the estimates of T_{MRCA} and Θ approached the estimates from the constant population size model. The credibility regions of T_{MRCA} from SNPs and SSLPs overlap, to a large extent, the SSLP data indicating the time to the T_{MRCA} to be shorter.

For *A. suecica* the T_{MRCA} was estimated using the *A. thaliana* estimates in expression (4). For the eight *A. suecica* accessions $\hat{S}_s = 0.091$ and for the 15 *A. thaliana* accessions $\hat{S}_t = 3.34$. Then to get an estimate of $\hat{T}_{MRCA,s}$ we just have to multiply the estimate of $\hat{T}_{MRCA,t}$ for a particular model by $\hat{S}_s/\hat{S}_t \approx 0.02725$. For example, given the lgc-model, the estimate of $\hat{T}_{MRCA,s}$ of the cp is 0.020 (0.011, 0.032), which is about 1/37 as long as $\hat{T}_{MRCA,t}$. To get an idea of what these estimates mean in terms of years, we can assume that the population sizes of the two species are similar. Fig. 5 shows $\hat{T}_{MRCA,t}$ and $\hat{T}_{MRCA,s}$ scaled in years for a wide range of values of N_e. If N_e is $>10^6$ then $\hat{T}_{MRCA,s}$ is >11 000 years. However, if N_e is allowed to be as low as 10^5 then $\hat{T}_{MRCA,s}$ would only be >1100 years. Based on the distribution range of the two species, we may instead assume that the population size of *A. suecica* is smaller than the population size of *A. thaliana* and the estimates of $\hat{T}_{MRCA,s}$ would then be proportionally smaller.

A second approach to dating *A. suecica* was based on the fact that no SNP variation was observed. Using expression (5) we found the 95% upper limits of T_{MRCA} to be 53 000, 20 000 and 13 000 years ago, respectively, using

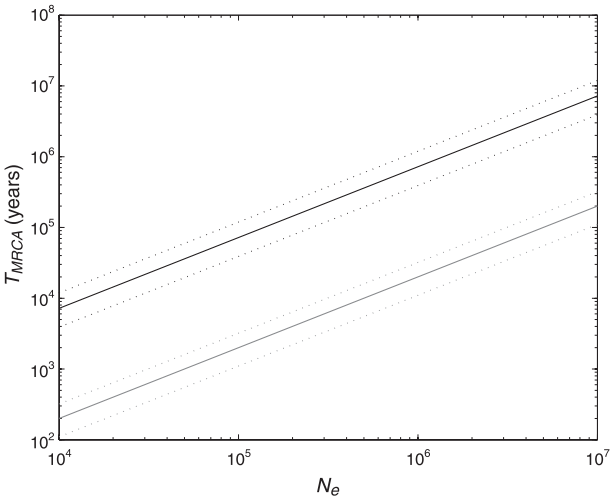

Fig. 5 The time to the most recent common ancestor of the cp genome for a sample of 15 *Arabidopsis thaliana* (black) and eight *Arabidopsis suecica* (grey) accessions. N_e is the scaling factor, the effective population size of both species. The dotted lines show the 95%-credibility interval.

the lower extreme (two branches), the expectation of a neutral tree and the upper extreme (star phylogeny) of the topology G. These upper bounds suggest that if the T_{MRCA} was beyond approximately 50 000, we would have found at least one SNP.

Given a single origin of the *A. suecica* cp, all the variation present today must have developed as the species originated. The upper limit of a 95% probability interval of not observing any variation (only SNPs being considered) in 9699 bp sequenced in eight *A. suecica* accessions corresponds to a H_e of 0.43 (in the case of only two possible haplotypes). Using a mutation rate of 2.9×10^{-9} per bp (Säll *et al.*, 2003), which means a mutation rate of 2.8×10^{-5} for a sequence of 9699 bp, the corresponding values for T_{MRCA} were computed for increasing values of N_e of *A. suecica* (Fig. 6). Thus, for this

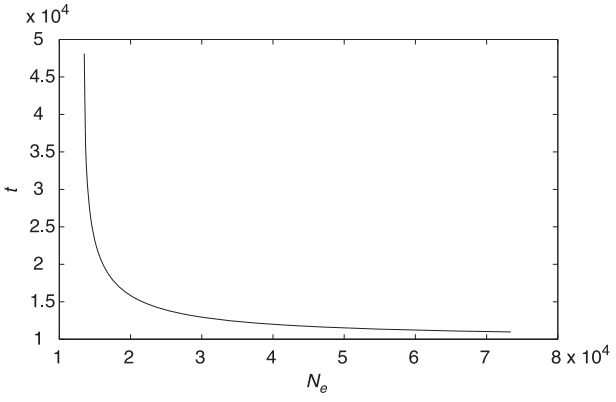

Fig 6 Time (t) to the upper 95% limit of not observing any SNPs among eight accessions of *Arabidopsis suecica* in 9699 bp as a function of the effective population size of *A. suecica* (N_e; see text for details).

simple scenario, the T_{MRCA} of the *A. suecica* sample was consistent with <20 000 years ago if $N_e > 16\,200$ and with <12 000 years ago if $N_e > 40\,000$. On the other hand, if $N_e < 14\,000$, the time-of-origin could be >30 000 years ago.

Discussion

Levels and pattern of variation

The cp genome of *A. thaliana* is variable with respect to SNPs, SSLPs as well as indels. Thus the general observation of *A. thaliana* as a highly variable species with respect to the nucleus (Nordborg *et al.*, 2005) is true for the cp as well. The fact that the SNPs behaved as UEP was expected, but more interesting is the fact that the indels indicated a recurrent occurrence of mutations. The cause of this is not known but one can speculate that it is because of the mutational process. With respect to *A. suecica*, Säll *et al.* (2003) found only one polymorphism in the *A. suecica* when sequencing roughly 4 kb of from eight *A. suecica* accessions. One of the objectives of the present investigation was to extend the number of polymorphisms in *A. suecica* in order to make detailed analyses of population structure in the species. In spite of an extensive effort, we only managed to find two additional polymorphisms, both of which were SSLPs. The mutations appeared to have occurred in loci with relatively large numbers of repeat units, a finding consistent with the observation that the mutation rate increases with the number of repeat units of microsatellites. Our results show that *A. suecica* is a species with a genuinely low level of molecular variation in its cpDNA.

No evidence of selection acting on the cp genome of *A. thaliana* was found. As the cp genome codes for several proteins that are important for its functioning, selection is expected to act on the cp genome. Moreover, selection pertaining to a particular gene affects the entire cp genome as recombination within it is expected to be rare or nonexistent (Palmer, 1987; Palmer *et al.*, 1988). Although Tajima's test is a common test of selection, the demography of the tested sample affects the outcome of the test (see e.g. Hein *et al.*, 2004). The size of the *A. thaliana* population probably expanded dramatically since the last ice age (Nordborg *et al.*, 2005). This population expansion is particularly apparent in northern Europe, as this region was completely covered by ice during the last ice age (Andersen & Borns, 1994). Although both purifying selection and population expansion are expected to shift Tajima's D towards negative values, the value obtained here was not statistically significant.

Population structure in *A. thaliana*

We found indications of population structure within *A. thaliana* at global level. Sharbel *et al.* (2000) obtained

a correlation coefficient of 0.24 between geographic distance and genetic distance when studying 79 nuclear AFLP markers from a worldwide sample of 113 *A. thaliana* accessions. Nordborg *et al.* (2005) found that the worldwide *A. thaliana* population was structured using 876 sequenced nuclear fragments from 96 worldwide accessions. Although our study used a different set of samples and targeted a different genome than the Nordborg *et al.* (2005) study, a similar picture of a global population structuring of the species was found in our study.

When only *A. thaliana* accessions from Sweden were investigated we found stronger patterns of geographic structure than in the worldwide sample. The high F_{ST} and R_{ST} values for a north-south division of the Swedish accessions, as well as the relatively high correlation between the geographic and genetic distances, may be a result of the recolonization of Fennoscandia (Sweden, Norway and Finland) from two glacial refugia (Sharbel *et al.*, 2000). One recolonization path may have been from the northeast and the other from the south. A variety of other taxa show similar patterns of geographic structure in Fennoscandia, for example mammals (reviewed by Jaarola *et al.*, 1999) and plants (Berglund & Westerbergh, 2001; Malm & Prentice, 2002).

The limited level of variation in *A. suecica* provides very weak power for investigating the level of population structure in this species. It is still noteworthy that all haplotypes observed were found in both Sweden and Finland.

Phylogeny

The *A. suecica* accessions falls onto a single branch in both trees in Fig. 1. Although the support is low, the clustering of *A. suecica* in these two trees, in combination with its low level of variation, is consistent with the *A. suecica* cp genome having a single origin. As the cp genome represents only a single evolutionary data point, its utility for inferring the evolutionary history of a plant is limited. Each cp study of a species must therefore be viewed as a simple case history rather than providing a general description of how the species has evolved. Strictly speaking, the indication in our case about the single origin is only valid for the eight investigated accessions. The probability that the deepest split in a population would be represented in a sample of only eight is, however, as high as 0.78 under neutrality (see e.g. Nordborg, 2001).

If the species had more than one origin and the (maternal) parents of the different founding individuals of *A. suecica* happened to be genetically very similar, then it would be very difficult today to detect whether there were more than one origin. However, from a genetic point of view this case may be viewed as an instance of a species having *effectively* a single origin. It is of course also impossible to show that every individual of species such as *A. suecica* originates from one origin, as there is always a chance that a local minority exists with a different origin from the rest of the species. Finally, note that the results presented here concern the cp chromosome alone. It is still possible that the cp genome is of single origin although the nuclear genome of the species has a more complex background.

Arabidopsis thaliana was found to have a clear but weak level of population structure on the global scale. It is therefore not surprising that the *A. thaliana* accessions that show the greatest similarity to *A. suecica* are geographically well dispersed. We still find the lack of *A. thaliana* accessions from Finland and Central Sweden in this group conspicuous, however. Thus, the earlier observation by Säll *et al.* (2003) that *A. suecica* is associated with southern Scandinavian/central European accessions of *A. thaliana*, rather than those from central Sweden and Finland, was also suggested by the present, much larger dataset.

T_{MRCA} of *A. thaliana* and *A. suecica*

Regardless of the number of origins of *A. suecica*, a single MRCA of the entire cp genome of *A. suecica* exists, given that no recombination in the cp genome has occurred. We have attempted to date this MRCA of the cp genome which, in the case of a single origin, is also the lower bound of the time-of-origin of the species.

Our results indicate that the branch containing the *A. suecica* accessions originated during the last several tens of thousands of years. We could not exclude the possibility that *A. suecica* originated less than 10 000 years ago behind the retreating ice of the last ice age. Neither could we exclude the possibility that *A. suecica* originated somewhere south of Fennoscandia and then migrated north into its present range in Fennoscandia. However, taking the results of the different analyses into consideration, there appeared to be an upper limit to the T_{MRCA} of the *A. suecica* cp genome of about 50 000 years ago and a lower limit of about 10 000 years ago. Thus, the origin of *A. suecica* may have coincided with the ending of the last glaciation, which was a gradual process occurring over a period of several thousand years, from about 15 000 to 9 000 years ago (Andersen & Borns, 1994). It is possible that the survival of this new species was aided by the dramatic ecological changes occurring during that period.

Acknowledgments

We thank S. Holm, A. Kurrto, H. Lindström, M. Nordborg, O. Savolainen and P. Stål for their help with the material, M. Sterner and L. Hall for technical assistance. We thank B. O. Bengtsson, A. Ceplitis and two anonymous reviewers for their comments on the manuscript. The work was supported by the Swedish Research

Council, the Erik-Philip Sörensen Foundation and the Nilsson-Ehle Foundation.

References

Abbott, R.J. & Lowe, A.J. 2004. Origins, establishment and evolution of new polyploid species: *Senecio cambrensis* and *S. eboracensis* in the British Isles. *Biol. J. Linn. Soc.* **82**: 467–474.

Ackerfield, J. & Wen, J, 2003. Evolution of *Hedera* (the ivy genus, Araliaceae): Insights from chloroplast DNA data. *Int. J. Plant Sci.* **164**: 593–602.

Ainouche, M.L., Baumel, A. & Salmon, A. 2004. *Spartina anglica* C. E. Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. *Biol. J. Linn. Soc.* **82**: 475–484.

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

Bennett, M.D. 2004. Perspectives on polyploidy in plants – ancient and neo. *Biol. J. Linn. Soc.* **82**: 411–423.

Berglund, A.B.N. & Westerbergh, A. 2001. Two postglacial immigration lineages of the polyploid *Cerastium alpinum* (Caryophyllaceae). *Hereditas* **134**: 171–183.

Chen, Z.J., Wang, J., Tian, L., Lee, H.-S., Wang, J.J., Chen, M., Lee, J.J., Josefsson, C., Madlung, A., Watson, B., Lippman, Z., Vaughn, M., Pires, J.C., Colot, V., Doerge, R.W., Martienssen, R.A., Comai, L. & Osborne, T.C. 2004. The development of an *Arabidopsis* model system for genome-wide analysis of polyploidy effects. *Biol. J. Linn. Soc.* **82**: 689–700.

Clegg, M.T., Gaut, B.S., Learn, G.H. & Morton, B.R. 1994. Rates and patterns of chloroplast DNA evolution. *Proc. Natl. Acad. Sci. USA.* **91**: 6795–6801.

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 *Arabidopsis allotetraploids*. *Plant. Cell.* **12**: 1551–1567.

Ennos, R.A., Sinclair, W.T., Hu, X.-S. & Langdon, A. 1999. Using organelle markers to elucidate the history, ecology and evolution of plant populations. In: *Molecular Systematics and Plant Evolution* (P.M. Hollingsworth, R.M. Bateman & R.J. Gornall eds), pp. 1–19, Taylor and Francis, London.

Fu, Y.X. & Li, W.-H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.

Gillespie, J.H. 1998. *Population Genetics a Concise Guide*. pp. 29. John Hopkins University Press, Baltimore and London.

Goldstein, B., Linares, A.R., Cavalli-Sforza, L.L. & Feldman, M.W. 1995. An evaluation of genetic distance for the use with microsatellite loci. *Genetics* **139**: 463–471.

Hein, J., Schierup, M.H. & Wiuf, C. 2004. *Gene Genealogies, Variation and Evolution*. Oxford University press, U.K.

Hill, F., Gemund, C., Benes, V., Ansorge, W. & Gibson, T.J. 2000. An estimate of large-scale sequencing accuracy. *EMBO Rep.* **1**: 29–31.

Hudson, R.R. & Kaplan, N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA-sequences. *Genetics* **111**: 147–164.

Hudson, R. R., Slatkin, M & Maddison, W.P. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.

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. *Symb. Bot. Ups.* **28**: 1–128.

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

Jørgensen, S. & Mauricio, R. 2004. Neutral genetic variation among wild North American populations of the weedy plant *Arabidopsis thaliana* is not geographically structured. *Mol. Ecol.* **13**: 3403–3413.

Jaarola, M., Tegelström, H. & Fredga, K. 1999. Colonization history in Fennoscandian rodents. *Biol. J. Linn. Soc.* **68**: 113–127.

Jakobsson, M., Hagenblad, J., Tavaré, S., Säll, T., Halldén, C., Lind-Halldén, C. & Nordborg, M. 2006. A recent unique origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. *Mol. Biol. Evol.* **23**: 1217–1231.

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.

Koch, M., 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.

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

Malécot, G. 1948. *Les Mathématiques de l'Hérédité*. Masson et Cie, Paris.

Malm, J.U. & Prentice, H.C. 2002. Immigration history and gene dispersal: allozyme variation in Nordic populations of the red campion, *Silene dioica* (Caryophyllaceae). *Biol. J. Linn. Soc.* **77**: 23–34.

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

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

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., Maruyama, T. & Chakraborty, R. 1975. The bottleneck effect and genetic variability in Populations. *Evolution* **29**: 1–10.

Nordborg, M. 2001. Coalescent theory. In: *Handbook of Statistical Genetics* (D. J. Balding, M. J. Bishop & C. Cannings, eds), pp. 179–212. John Wiley & Sons, Inc., Chichester, U.K.

Nordborg, M. & Donnelly, P. 1997. The coalescent process with selfing. *Genetics* **146**: 1185–1195.

Nordborg, M., Hu, T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Tomer, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N.A., Shah, C., Wall, J., Zhao, K., Kalbfleisch, T., Schultz, V., Kreitman, M. & Bergelson, J. 2005. The Pattern of Polymorphism in *Arabidopsis thaliana*. *PLoS Biology* **7**: 1289–1299.

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.

Olmstead, R.G. & Palmer, J.D. 1994. Chloroplast DNA systematics – a review of methods and data-analysis. *Am. J. Bot.* **81**: 1205–1224.

Palmer, J.D. 1987. Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. *Am. Nat.* **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. Mo. Bot. Gard.* **75**: 1180–1206.

Pontes, O., Lawrence, R.J., Neves, N., Silva, M., Lee, J.H., Chen, Z.J., Viegas, W. & Pikaard, C.S. 2003. Natural variation in nucleolar dominance reveals the relationship between nucleolus organizer chromatin topology and rRNA gene transcription in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* **100**: 11418–11423.

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, New York.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. & Feldman, M.W. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**: 1791–1798.

Provan, J., Powell, W. & Hollingsworth, M. 2001. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends Ecol. Evol.* **16**: 142–147

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

Rosenberg, N.A. & Nordborg, M. 2002. Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3**: 380–390.

Saitou, N. & Nei, M. 1987. The Neighbor-Joining Method – a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.

Säll, T., Jakobsson, M., Lind-Halldén, C. & Halldén, C. 2003. Chloroplast DNA indicates a single origin of the allotetraploid *Ararbidopsis suecica*. *J. Evol. Biol.* **16**: 1019–1029

Säll, T., Halldén, C., Jakobsson, M. & Lind-Halldén, C. 2004. Mode of reproduction and population structure in *Arabidopsis suecica*. *Hereditas.* **141**: 313–317.

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.

Segraves, K.A., Thompson, J.N., Soltis, P.S. & Soltis, D.E. 1999. Multiple origins of polyploidy of the geographic structure of *Heuchera grossularifolia*. *Mol. Ecol.* **8**: 253–262.

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.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457–462.

Sokal, R.R. & Rohlf, F.J. 1995. *Biometry*, 3rd edn. Freeman, 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. U. S. A.* **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.

Wang, J.L., Tian, L., Madlung, A., Lee, H.S., Chen, M., Lee, J.J., Watson, B., Kagochi, T., Comai, L., Chen, Z.J. 2004. Stochastic and epigenetic changes of gene expression in arabidopsis polyploids. *Genetics* **167**: 1961–1973.

Wendel, J. 2000. Genome evolution in polyploids. *Plant Mol. Biol.* **42**: 225–249.

Widmer, A. & Baltisberger, M. 1999. Molecular evidence for allopolyploid speciation and a single origin of the narrow endemic *Draba ladina* (Brassicaceae). *Am. J. Bot.* **86**: 1282–1289.

Wilson, I.J. & Balding, D.J. 1998. Genealogical inference from microsatellite data. *Genetics* **150**: 499–510.

Wilson, I.J., Weale, M.E. & Balding, D.J. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Stat. Soc. Ser. B* **166**: 155–188.

Supplementary Material

The following supplementary material is available for this article online:

Table S1 Primer sequences and 3′-positions used to amplify 50 DNA fragments in the choloroplast genome of *A. thaliana* and *A. suecica*.

This material is available as part of the online article from http://www.blackwell-synergy.com

Received 28 April 2006; revised 15 July 2006; accepted 24 July 2006