Molecular Biology and Evolution

A Unique Recent Origin of the Allotetraploid Species *Arabidopsis suecica*: Evidence from Nuclear DNA Markers

Mattias Jakobsson,† Jenny Hagenblad,‡ Simon Tavaré,§|| Torbjörn Säll,† Christer Halldén,¶ Christina Lind-Halldén,# and Magnus Nordborg§*

*Bioinformatics Program, Department of Human Genetics, University of Michigan; †Department of Cell and Organism Biology, Genetics, Lund University, Lund, Sweden; ‡Department of Biology, Linköping University, Linköping, Sweden; §Molecular and Computational Biology, University of Southern California; ||Department of Oncology, University of Cambridge, Cambridge, United Kingdom; ¶Department of Clinical Chemistry, Malmö University Hospital; #Department of Mathematics and Natural Sciences, Kristianstad University, Kristianstad, Sweden

A coalescent-based method was used to investigate the origins of the allotetraploid *Arabidopsis suecica*, using 52 nuclear microsatellite loci typed in eight individuals of *A. suecica* and 14 individuals of its maternal parent *Arabidopsis thaliana,* and four short fragments of genomic DNA sequenced in a sample of four individuals of *A. suecica* and in both its parental species *A. thaliana* and *Arabidopsis arenosa*. All loci were variable in *A. thaliana* but only 24 of the 52 microsatellite loci and none of the four sequence fragments were variable in *A. suecica*. We explore a number of possible evolutionary scenarios for *A. suecica* and conclude that it is likely that *A. suecica* has a recent, unique origin between 12,000 and 300,000 years ago. The time estimates depend strongly on what is assumed about population growth and rates of mutation. When combined with what is known about the history of glaciations, our results suggest that *A. suecica* originated south of its present distribution in Sweden and Finland and then migrated north, perhaps in the wake of the retreating ice.

Introduction

It has long been recognized that polyploidy has played an important role in speciation, especially in plants (Stebbins 1971). In contrast with other forms of speciation, polyploidization is instantaneous: a bona fide "speciation event." new polyploid is isolated from its parental species because backcrossing to either parent is likely to produce sterile offspring. The new species will therefore evolve independently of its parental species, and if the polyploidization occurs only once there will be no additional gene flow into the species once it is formed. Indeed, even if it occurs several times, different polyploid lineages may well be isolated because of the rapid chromosome rearrangements (Song *et al*. 1995; Pontes *et al*. 2004) and changes in gene expression (Chen, Comai, and Pikaard 1998; Comai *et al*. 2000; Osborn *et al*. 2003) that may accompany polyploidization. The fraction of plant species that are of polyploid origin has been debated; however, it increasingly seems as if this is not a meaningful question. If polyploid species become "diploidized" over time, then it may well be that all plant species have polyploid ancestors (Ku *et al*. 2000; Wendel 2000). Although comparative genomics can detect more ancient events than classical cytology, we will never be able to detect all events.

Arabidopsis suecica (Fr.) was first described as an allotetraploid ($2n = 26$) by Hylander (1957), who suggested *Arabidpsis thaliana* and *Arabidopsis* (*Cardaminopsis*) *arenosa* as its parental species, a suggestion since supported by molecular phylogenetics (Kamm *et al*. 1995; O'kane, Schaal, and Al-Shehbaz 1996). *A. suecica* is found in Sweden and southern Finland (Hultén 1971) and it is highly, but not completely, selfing (Säll *et al*. 2004). *A. thaliana* is a highly selfing, cosmopolitan weed, familiar as the leading model plant. It is diploid ($2n = 10$), although tetraploid *A. thaliana* individuals are occasionally found (L. Comai and M. Koornneef, personal communication). *A. arenosa* is a European species which is tetraploid ($2n = 4x = 32$) in most of its range and reportedly diploid ($2n = 2x = 16$) in a small area in eastern Europe (Mesicek 1970). Our observations in the greenhouse indicate that *A. arenosa* is self-incompatible and thus an obligate outbreeder (Säll *et al*. 2004). Several lines of evidence, including most recently cpDNA sequencing, have identified *A. thaliana* as the maternal parent of *A. suecica* (Mummenhoff and Hurka 1994; Price, Al-Shebaz, and Palmer 1994; Comai *et al*. 2000; Säll *et al*. 2003). Furthermore, when producing "artificial" *A. suecica* through crosses, Comai *et al*. (2000) only succeeded in producing viable offspring when *A. thaliana* was the maternal parent. Given this, there are still several possible routes through which *A. suecica* could have originated (table 1). The most likely scenario would appear to involve a normal male gamete from a tetraploid *A. arenosa* fertilizing either a normal female gamete from a tetraploid *A. thaliana* (which are rare, but do exist), or an unreduced female gamete from a diploid *A. thaliana*.

An important aspect of the formation of a polyploid species is the number of independent origins. Many allopolyploid species are believed to have multiple origins (D. Soltis and P. Soltis 1993). The traditional way to investigate this question is to produce a phylogenetic tree based on data from different accessions of the allopolyploid and its potential parental species. If all the polyploids form a single branch within the tree, there is evidence of a single origin, whereas if the allopolyploids appear at different positions relative to the parent, there is evidence of multiple origins. This method was used to analyze cpDNA from *A. suecica,* and the results indicated a single origin (Säll *et al*. 2003). Important drawbacks of this approach include that it is limited to non-recombining data, such as cpDNA in plants and mtDNA in animals, and that it provides no measure of statistical confidence in the conclusions. The genealogy of a single locus does not necessarily reflect the history of the species (Rosenberg and Nordborg 2002).

Key words: *Arabidopsis suecica*, *Arabidopsis thaliana*, polyploidy, polymorphism, speciation, founders.

E-mail: mjakob@umich.edu.

Mol. Biol. Evol. 23(6):1217–1231. 2006
doi:10.1093/molbev/msk006
Advance Access publication March 20, 2006

Table 1
Possible Scenarios for the Origin of *Arabidopsis suecica* (the number of chromosomes is indicated in parenthesis)

Arabidopsis thaliana—maternal	*Arabidopsis arenosa*—paternal	
	Diploid—exists? ($2n = 16$)	Tetraploid ($4x = 32$)
Diploid ($2n = 10$)	Genome duplication following fertilization	Unreduced ovule
Tetraploid—rare ($4x = 20$)	Unreduced pollen	Normal meiosis and fertilization

In this paper we introduce a Bayesian multi-locus method based on coalescent theory that can be used to infer the number of origins and the time of the origin of a polyploid. The method is based on a rejection algorithm where sets of demographic parameters are simulated under a demographic model of two species (one "parent" and one allopolyploid "offspring" species). We utilize our method to analyze two different data sets: the first is 52 nuclear microsatellite markers from 8 *A. suecica* and 14 *A. thaliana* individuals; the second is a smaller data set of four (~0.5 kb each) sequenced fragments of nuclear DNA from each of four individuals of *A. suecica*, as well as from the two parental species *A. thaliana* and *A. arenosa*. We conclude that there has been a single origin of *A. suecica* between 12,000 and 300,000 years ago, although the time-of-origin estimates are sensitive to modeling assumptions.

Materials and Methods

Sequencing

DNA was extracted from fresh leaves using Qiagen miniprep kits. Four *A. suecica* individuals from four geographically distinct locations covering much of the species range were sampled (table 2). Four short fragments located on chromosome 4 in *A. thaliana* were polymerase chain reaction (PCR)-amplified in each of the four *A. suecica* individuals, using primers designed from the *A. thaliana* genome sequence (Table S1, Supplementary Material online). The primers had previously been used to amplify and sequence four fragments in 20 accessions of *A. thaliana* (see Nordborg and Bergelson (1999) and Hagenblad and Nordborg (2002) for details on these accessions) and in one *A. arenosa* accession from Piteå, northern Sweden (6518N, 2131E). Attempts were made to amplify the fragments from one individual of each of the outgroup species *Capsella bursa-pastoris* and *Cardamine hirsuta*, both of which were sampled in Lund, southern Sweden (5542N, 1311E). PCRs were run with standard Taq supplied by Roche Diagnostics (Basel, Switzerland) and according to the supplier's recommendations. Sequencing was done on a Beckman Coulter CEQ 2000 XL sequencer, using reagents supplied by the manufacturer. A fourfold dilution of the reagents was used for sequencing; apart from this, standard protocols were used. All fragments were sequenced in both directions and sequences were aligned and evaluated in Sequencher 4.1 (GeneCode, Ann Arbor, Mich.). Errors in basecalling were corrected manually after the chromatograms were examined.

Table 2
Origins of the Four *Arabidopsis suecica* Individuals for which Four Fragments were Sequenced

Name	Location	Region	Coordinates
As1	Tjörnarp	S. Sweden	5559N, 1338E
As2	Ytterhogdal	C. Sweden	6210N, 1456E
As3	Stocktjärn	C. Sweden	6346N, 2011E
As4	Hanko	S. Finland	5953N, 2307E

For the allotetraploid *A. suecica* individuals, the PCR products were cloned in order to allow individual alleles to be sequenced. Cloning was done with the pGEMT Easy Vector System using standard protocols. Clones were then selected randomly and used for an additional round of PCR amplification to provide template for direct sequencing of ten clones from each individual and fragment. Assuming no bias in cloning or PCR amplification, this gives a 99.9% chance of amplifying at least one copy of each of the two ancestral allelic types (*A. arenosa* and *A. thaliana*).

Once sequences from the parental genomes in *A. suecica* had been obtained by sequencing cloned fragments, we designed allele-specific primers to directly amplify and sequence each sequence type. This allowed us to distinguish variation introduced by cloning and PCR from natural variation.

Sequencing of cloned PCR fragments from *A. suecica* genomic DNA resulted in sequences that could easily be designated as thaliana-like or arenosa-like by comparison with sequences from the ancestral species. For each primer pair, there were always more clones with thaliana-like than with arenosa-like sequences, and for one fragment, no arenosa-like sequences at all were obtained even though an additional 10 clones were sequenced. The bias toward thaliana-like sequences was not surprising, as the primers used for PCR amplification were designed using the *A. thaliana* genome sequence.

Recombinant fragments were found in 6 out of the original 40 clones sequenced. Recombination between chromosomes of different parental origin is possible but seems unlikely, and allele-specific primers revealed that the recombinant fragments were indeed cloning artifacts.

In summary we successfully sequenced both the arenosa-like and the thaliana-like alleles for three out of the four fragments. For the fourth fragment, only a thaliana-like sequence was obtained (table 3). There was no variation within allelic classes, either within or between individuals.

Multiple alignments were done in Sequencher, with additional adjustments by hand. Genealogical trees were drawn using the phylogeny program PAUP 4.0.0b10 (Sinauer Associates, Sunderland, USA) or by hand (as the sequences were on the whole very similar). The trees were rooted using outgroup sequences when these were available. NCBI annotation for *A. thaliana* BAC clone F6N23 (accession number AF058919) was used to determine the positions of exons and introns in the fragments.

For the outgroup species, we used whichever species that was successfully amplified and sequenced. This turned out to be *Cardamine hirsuta* for two fragments and *Capsella bursa-pastoris* for a third. No outgroup was obtained for the fourth fragment (see table 3, below). The outgroup

Table 3
Sequences Obtained as Part of This Study

Fragment	Location on chromosome 4	Length	*Arabidopsis suecica*		Outgroup
			Thaliana-like	Arenosa-like	
1	233801–234371	668	As1–As4	As1–As4	*Cardamine hirsuta*
2	264692–265062	379	As1–As4	As1–As4	*Capsella bursa-pastoris*
3	269509–269959	450	As1–As4	Failed	*C. hirsuta*
4	301616–301966	352	As1–As4	As1–As4	Failed

NOTE.—All fragments are located on chromosome 4 in *Arabidopsis thaliana* and the coordinates are based on the available *Col* sequence. The lengths given for the fragments are the length of the aligned sequences excluding the outgroup. In addition to the sequences listed, one individual of *Arabidopsis arenosa* was successfully sequenced for all fragments.

sequence was used to assign mutations to either the branch of *A. thaliana* or *A. arenosa*. Mutations that could not be assigned to any branch were divided evenly between the two branches. The mutation rates were then calculated by dividing the number of mutations on each branch by the product of the length of the sequence and the time since divergence of the two species.

Microsatellite Analysis

A total of eight *A. suecica* individuals from geographically distinct locations covering most of the species range were sampled in addition to 14 *A. thaliana* individuals (table 4). DNA was isolated from young greenhouse-grown offspring of the sampled plants using a Plant DNeasy kit from Qiagen according to the manufacturer's instructions. DNA integrity was checked by agarose gel electrophoresis and DNA concentration was determined using fluorometry with Pico Green (Molecular Probes, Carlsbad, Calif.). A total of 52 microsatellite loci were identified from the complete sequence of *A. thaliana* (version 2.0) using Tandem Repeat Finder (Benson 1999) and appropriate primers (C. Lind-Halldén, C. Halldén, M. Jakobsson, and T. Säll, unpublished data) were designed with Oligo (v. 6.3). Primer sequences were subsequently analyzed for specificity using Blast. One primer for each pair was labeled with either HEX or 6-FAM. PCR assays were optimized with regard to annealing temperature and primer concentrations. The exact PCR conditions for each primer pair are available from the authors on request. PCR reactions contained 5 ng template DNA, 2 mM $MgCl_2$, 100 μM dNTP (Amersham/Pharmacia, Uppsala, Sweden), 0.75 units AmpliTaq Gold polymerase (Applied Biosystems, Foster City, Calif.), 1 × PCR reaction buffer (Applied Biosystems), and 50–900 nM of each primer in a final volume of 25 μl. PCR products were resolved using capillary electrophoresis run on ABI 310 or 3730 sequencers employing GeneScan and GeneMapper v.3.0 software (Applied Biosystems). The allele sizes of the

Table 4
The 14 *Arabidopsis thaliana* and 8 *Arabidopsis suecica* Accessions Typed with 52 Microsatellite Markers

Name	Species	Location	Region	Coordinates
T160	*A. thaliana*	Västervik[a]	SC. Sweden	5745N, 1640E
T81	*A. thaliana*	Karhumäki[b]	W. Russia	3255N, 3425E
T93	*A. thaliana*	Tvärminne[b]	S. Finland	5951N, 2319E
T104	*A. thaliana*	Nurmes[b]	C. Finland	6332N, 2910E
Oy-0	*A. thaliana*	Oystese	Norway	6023N, 0612E
T10	*A. thaliana*	Lilla Edet[a]	SC. Sweden	5806N, 1209E
T700	*A. thaliana*	Anten[a]	SC. Sweden	5759N, 1225E
T340	*A. thaliana*	Höör[a]	S. Sweden	5556N, 1332E
T350	*A. thaliana*	Klevshult[a]	SC. Sweden	5721N, 1405E
Kas-1	*A. thaliana*	Kashmir	India	3436N, 7448E
Lip-1	*A. thaliana*	Lipowiec	Poland	5005N, 1927E
Sv-0	*A. thaliana*	Svebolle	Denmark	5538N, 1116E
Wil-1	*A. thaliana*	Wilma	Lithuenia	5500N, 2500E
Bu-0	*A. thaliana*	Burghaun	Germany	5042N, 0943E
S60	*A. suecica*	Vännas[a]	C. Sweden	6355N, 1946E
S90	*A. suecica*	Västanbäck[a]	C. Sweden	6347N, 1705E
S130	*A. suecica*	Strömsbruk[a]	C. Sweden	6153N, 1719E
As2 (S150)	*A. suecica*	Ytterhogdal[a]	C. Sweden	6210N, 1456E
S261	*A. suecica*	Hammarstrand[c]	C. Sweden	6307N, 1622E
S300	*A. suecica*	Sörfjärda[d]	C. Sweden	6202N, 1727E
S354	*A. suecica*	Iisalmi[b]	C. Finland	6343N, 2712E
As4 (S361)	*A. suecica*	Hanko[b]	S. Finland	5953N, 2307E

NOTE.—Except in the cases mentioned in the designated footnotes, the accessions were acquired from the Nottingham Seed Stock Center and The SENDAI *Arabidopsis* Seed Stock Center.

[a] Collected by the authors.

[b] Outi Savolainen, Oulo University.

[c] Håkan Lindström, Tjälarne.

[d] Svante Holm, Mitthögskolan.

microsatellite markers were determined in relation to a size marker, GeneScan-500 LIZ (Applied Biosystems). The 52 microsatellite loci were recruited from the whole nuclear *A. thaliana* genome with an overrepresentation of loci from chromosome 2.

Models and Rejection Algorithms

Because the sequence data set was much smaller, did not include a good sample of Swedish *A. thaliana* accessions, and no variation in *A. suecica*, we used slightly different models for the two data sets. For the sequence data we only consider polymorphism within *A. suecica*, and restrict ourselves to inferring the time-of-origin under the assumption of a single origin. Because no variation was detected in *A. suecica* the maximum likelihood estimate of the number of origins would be a single origin under any model. For the microsatellite data we model polymorphism within both species, and infer the number of founders as well.

Before we describe the explicit models used in this article let us introduce some notation. For a population with constant size N, the coalescent model states that the time W_j during which a sample of size n has j ancestors ($2 \leq j \leq n$) is exponentially distributed with parameter $j(j-1)/2$. The times for different j are independent and we assume in the following that time is counted in units of N generations. The time to the most recent common ancestor T is given by

$$T = \sum_{j=2}^{n} W_j \tag{1}$$

and the total length of the tree L is

$$L = \sum_{j=2}^{n} jW_j. \tag{2}$$

When the population size is variable, the times W_n, W_{n-1}, …, W_2 are no longer independent. In the case when the population size grows exponentially (forward in time), the population size at time t (backwards in time) is $N(t)=N(0)\exp(-\rho t)$, for some constant value of ρ. The conditional distribution of the time W_j for which there are exactly j ancestors, given that the time in which there are more than j ancestors is s, is (Tavaré *et al.* 1997; eq. 21):

$$\mathbf{P}(W_j > t \mid W_n + \cdots + W_{j+1} = s) = \exp\left(-\binom{j}{2}\int_s^{s+t} \lambda(u)\mathrm{d}u\right), \tag{3}$$

where $\lambda(t)=\exp(\rho t)$ in the case of exponential growth.

For a set of parameters Ψ, the conditional probability density function of Ψ given the data D is denoted $f(\Psi \mid D)$,

$$f(\Psi \mid D) = \frac{\pi(\Psi)\mathbf{P}(D \mid \Psi)}{\mathbf{P}(D)} = \frac{\pi(\Psi)}{\mathbf{P}(D)}\int_G \mathbf{P}(D \mid \Psi, G)\mathbf{P}(G \mid \Psi)\mathrm{d}G, \tag{4}$$

where $\pi(\Psi)$ is the product of the probability density functions (pdf) of the prior distributions and $\mathbf{P}(G)$ is the probability of the genealogy that is given by the joint pdf (3). We will also assume that mutations occur independently in each branch with constant rate μ. Then, if a branch is of length w, the number of mutations, k, will be Poisson distributed with mean $wN\mu$, $\mathrm{Po}(k, wN\mu)$.

A Model of a Species Founded by One Individual, Model 1

For the sequence data we considered a model (*Model 1*) where *A. suecica* grew exponentially from a single individual τ years (or generations: *A. suecica* appears to be annual) ago to a present effective size of N individuals (we tried different scenarios for growth, but they all gave similar results). Gene genealogies follow the standard coalescent distribution for a growing population (eq. 3). We assume that *A. suecica* is sufficiently highly selfing that the rate of coalescence is twice that of an equivalent outbreeding population (Nordborg and Donnelly 1997). Neutral mutations were assumed to occur with probability μ_s per bp per generation. For $\Psi = \{\tau, N, \mu_s\}$ and by substituting the data D with the number of segregating sites S, the conditional pdf (4) can be expressed as (Tavaré *et al.* 1997):

$$f(\tau, N, \mu_s \mid S = k) \propto \int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty f(l^*)\pi_N(u)\pi_{\mu_s}(v)\pi_\tau(x) \times \mathrm{Po}(k, luv)\mathrm{d}l^*\mathrm{d}u\mathrm{d}v\mathrm{d}x, \tag{5}$$

where $f(l^*)$ denotes the pdf under the coalescent model of the total length of the genealogy up to $\min(\tau, T)$. The pdf $\pi_{\mu_s}(v)$, $\pi_N(u)$ and $\pi_\tau(x)$ are the prior distribution of N, μ_s, and τ.

Because the sequenced fragments were located within a 70-kb region, the degree of selfing influences how we should treat the seven fragments (four thaliana-like and three arenosa-like). If *A. suecica* was completely selfing, then all seven fragments must have the same genealogy. If *A. suecica* was completely outcrossing, then the seven fragments would probably have almost independent genealogical histories depending on the recombination fraction between the fragments. We considered both extremes, complete selfing and complete outcrossing, but the results turned out to be almost identical so we only present the results from the former model here.

Estimating the Time-of-origin from Model 1

We assumed uniform prior distributions for the parameters. The posterior distribution given the data (eq. 5) was evaluated using a simple rejection algorithm (cf. Tavaré *et al.* 1997):

Algorithm 1. Rejection algorithm for $f(\tau, N, \mu_s \mid S = k)$.

1. Simulate a parameter set $\Psi = \{\tau, N, \mu_s\}$ from the prior distributions.
2. Simulate a standard haploid (to reflect selfing) coalescent for a population that has undergone τ generations of exponential growth from a single individual to reach a current size of N (eq. 3), and record the total length L^* (eq. 2) of the branches up to the minimum of τ and T (eq. 1) of the sample.
3. Accept the parameter set with probability $\exp(-NL^*\mu_s b)$, where b is the number of sequenced bases (thus only

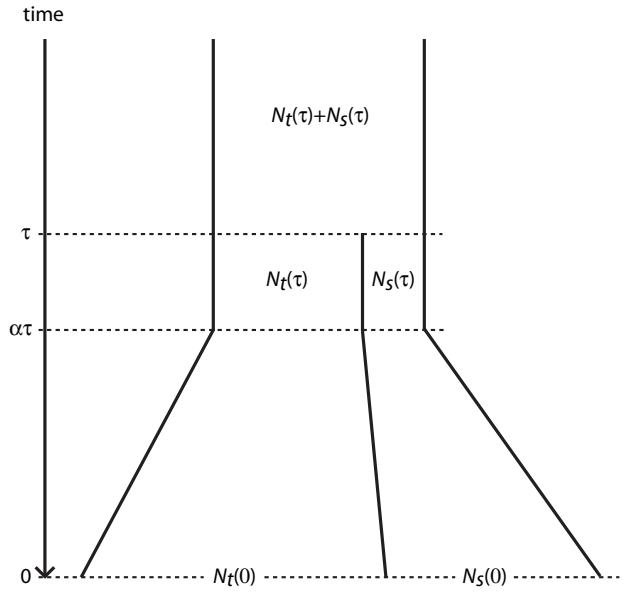

FIG. 1.—Overview of *Model 2* that is used in *Algorithm 2* to infer the number of origins of *A. suecica* and the time-of-origin. Time 0 is the present and τ is the time-of-origin. $N_t(0)$, $N_s(0)$, $N_t(\tau)$, and $N_s(\tau)$ are the population sizes for species *t* and *s* at time 0 and τ.

parameter sets without mutations are accepted), otherwise reject it.

The parameter values that result from accepted observations in this algorithm are then samples from the posterior distribution.

A Model of a Parent Species and an Offspring Species, Model 2

For the microsatellite data we considered a model (*Model 2*; fig. 1) of two species, *t* and *s* (*A. thaliana* and *A. suecica*), which have been completely separated from each other for τ years (or generations). Both species have grown exponentially from ατ years ago ($0 < \alpha \leq 1$) to the present. The population size of species *t* was $N_t(\tau)$ at time τ (as well as at time ατ) and has grown to a population size of $N_t(0)$ at present. We define $N_s(\tau)$ as the population size of species *s* a negligibly short time τ_s before (backwards in time) the actual founding event. This means that the *n* founders are allowed to increase to $N_s(\tau)$ during τ_s, and the effect of drift during τ_s will be considered below. Before (backwards in time) this short initial phase drift is modeled using the coalescent. The population size of species *s* grows exponentially from $N_s(\tau)$ to a population size of $N_s(0)$ at present. After τ years (backwards in time) the two populations were considered to be one population with a constant population size of $N_t(\tau)+N_s(\tau)$. Gene genealogies from 0 to ατ years ago follow the standard coalescent distribution for a growing population (eq. 3) and the gene genealogies after ατ years follow the standard coalescent distribution with constant population size (see, e.g., Nordborg 2001). We again assume that both species are sufficiently highly selfing that the rate of coalescence is twice that of an equivalent outbreeding population (Nordborg and Donnelly 1997). Microsatellite loci were modeled according to a stepwise mutation model (Ohta and Kimura 1973), that is, mutations were assumed to be Poisson distributed on each branch and a mutation is an addition or a subtraction of one repeat unit at a particular microsatellite locus. Mutations were assumed to occur with probability μ_m per generation and locus. Here we use another summary statistic, the average pairwise difference (S'), to simplify the data *D*. S' was computed using the following expression:

$$S' = \sum_{i=1}^{I}\left[2\sum_{j=1}^{n}\sum_{k>j}|l_{i,j}-l_{i,k}|/[n_i(n_i-1)]\right], \quad (6)$$

where *I* is the number of loci, $l_{i,j}$ and $l_{i,k}$ are the lengths of locus *i*, in individual *j* and *k* in repeat units, and n_i is the number of sampled individuals at locus *i*. The statistic used to summarize the data was then

$$DS = |S'_{t,sim} - S'_{t,obs}| + |S'_{s,sim} - S'_{s,obs}|, \quad (7)$$

where the statistics $S'_{t,sim}$ and $S'_{s,sim}$ are computed from eq. (6) for simulated data sets for species *t* and *s*, and the statistics $S'_{t,obs}$ and $S'_{s,obs}$ are computed from eq. (6) for observed data sets for species *t* and *s*. The posterior distributions given the data

$$f(\tau, N_t(0), N_s(0), N_t(\tau), N_s(\tau), \mu_m \mid DS<\varepsilon) \quad (8)$$

under *Model 2* was evaluated using a rejection algorithm similar to *Algorithm 1*. The value of α was fixed to a predefined value between 0 and 1. For each of the parameters a uniform prior distribution was assumed.

Algorithm 2. Rejection algorithm for $f(\tau, N_t(0), N_s(0), N_t(\tau), N_s(\tau), \mu_m|DS < \varepsilon)$.

1. Simulate a parameter set $\Psi = \{\tau, N_t(0), N_s(0), N_t(\tau), N_s(\tau), \mu_m\}$ from the prior distributions.
2. Repeat steps a and b for the total number of modeled loci, *I*:
   a. Simulate two standard haploid (to reflect selfing) coalescent for the two populations (species *t* and species *s*) until τ years ago, where the population sizes were $N_t(\tau)$ and $N_s(\tau)$ (note that the population sizes were constant from ατ years to τ years) that have undergone ατ generations of exponential growth (eq. 3) to reach a current sizes of $N_t(0)$ and $N_s(0)$.
   b. Simulate a standard haploid coalescent from τ years with constant population size of $N_t(\tau)+N_s(\tau)$ for all remaining lines at τ.
3. Accept the parameter set if $DS < \varepsilon$, where ε is a value > 0, otherwise reject it.

We chose ε to equal $\beta(S'_{t,obs}+S'_{s,obs})$ where β was set to some small value (β was chosen to be as small as possible and still accept reasonable number of samples from the algorithm in a reasonable amount of time. We used $\beta = 0.05$, unless otherwise indicated). The parameter values that result from accepted observations in *Algorithm 2* are then samples from the posterior distributions (8) of the parameters. Our model of an ancestral population that splits into two populations is relatively similar to the model of Hey (2005). The method described above is derived from Tavaré *et al.* (1997) and Pritchard *et al.* (1999). Our method uses multi-locus data and infers the number of origins

of a particular species, which are the major differences between our method and the method of Prichard *et al.* (1999). For additional similar approaches see Approximate Bayesian Computation methods (Beaumont, Zhang, and Balding 2002).

A Model of n Founders, Model 2a

To estimate the number of founders we modified *Model 2* to allow modeling a fixed number of founders, n slightly. The modified model, *Model 2a*, includes an extra parameter n and founders denoted $j = 1, \ldots, n$. All lines from species s that remain at time τ are randomly assigned to the n founders with equal probability $(1/n)$ for each locus i. We thus do not model drift during the establishment of the species during the first τ_s generations since the appearance of *A. suecica* (or, rather, the first *A. suecica* that could have been the ancestor of modern *A. suecica*: it is possible that earlier polyploidization events occurred that subsequently went extinct) explicitly. Our results can be viewed as being conditional on the composition of the species after it has become established. The maximum number of founders of species s is then n, but fewer founders are possible in this model. This is the only difference between *Model 2* and *Model 2a*. The accepted parameter sets are then samples from the posterior distribution of n. We also used *Model 2a* to compare scenarios of different number of founders n, given the data, for fixed values of τ, $N_t(0)$, $N_s(0)$, $N_t(\tau)$, $N_s(\tau)$, and α. Likelihoods were approximated by the acceptance rate of *Algorithm 2*.

An Unequal Contribution Model, Model 2b

Model 2a assumes that each founder j contributed equally to the founding of the species. A slightly more realistic model incorporates the possibility of unequal contributions of the n founders. We therefore made a modification of *Model 2a* that allows unequal contributions of different founders to the whole founding population. This was done by assigning the lineages of species s that remain at τ to either of the j founders with probability p_j, and $\sum_{j=1}^{n} p_j = 1$. The probability p_j can be considered to be the proportion that founder j contributed to the whole founding population. The founding population is then a common gene pool [of population size $N_s(\tau)$] at time τ to which each founder j contributed p_j. This founding population is then modeled in the same way as in *Model 2*. This model is henceforth known as *Model 2b*. We approximated the posterior distribution of p_j given the data in the same way as for *Model 2a*.

Results

Sequence and Microsatellite Variation in *A. thaliana*, *A. arenosa*, and *A. suecica*

Sequence divergence between *A. thaliana* and *A. arenosa* is summarized in table 5. Fragment 1 contained a number of indels and could not be aligned unambiguously, so the alignment requiring the least number of indels was used. The exact number of indels and substitutions will depend on the alignment, but an alternative alignment seems unlikely to affect the general conclusions drawn. We also note that the substitution rates for this fragment are broadly in agreement with those for other fragments, suggesting that the alignment is reasonable. As expected, non-coding regions differed more than exons, in particular with respect to indels, which were completely absent from the coding regions. In the coding regions, the rate of synonymous substitution was higher than the non-synonymous rate ($K_S = 0.18 > K_A = 0.04$).

Table 5
Divergence Between *Arabidopsis arenosa* and *Arabidopsis thaliana*

	Fragment				
	1	2	3	4	Total
Coding sequence					
No. of sites	0	135	339	286	759
No. fixed difference	—	2	28	16	46
Per bp	—	0.015	0.088	0.058	0.061
Synonymous	—	0	16	8	22
K_S	—	0	0.254	0.185	0.176
Nonsynonymous	—	2	12	8	20
K_A	—	0.019	0.048	0.031	0.035
Non-coding sequence					
No. of sites	540	234	111	63	950
No. fixed differences	62	15	9	10	96
Per bp	0.125	0.067	0.086	0.178	0.109

NOTE.—For non-coding sequence, only sites shared by both species have been included and sites where there has been an indel event between the two species has been removed from the analysis. K_A is the number of nonsynonymous differences per nonsynonymous site and K_S is the number of synonymous differences per synonymous site. The values for number of fixed differences per bp, K_A and K_S are corrected for recurrent mutations using Jukes-Cantor's one parameter model.

The divergence time between *A. thaliana* and *A. lyrata* has been estimated to be between 3.8 and 5.8 MYA by Kuittinen and Aguadé (2000) and to be between 5.1 and 5.4 MYA by Koch, Haubold, and Mitchell-Olds (2000). As *A. arenosa* diverged from *A. lyrata* only after the two species diverged from *A. thaliana* (Yang *et al.* 1999; Heenan, Mitchell, and Koch 2002), 5 Myr is a reasonable estimate of the divergence time between *A. thaliana* and *A. arenosa*. Using this divergence time (and outgroup information), the synonymous substitution rate was found to be 6.0×10^{-9} substitutions per bp per year in the *A. thaliana* lineage and 6.2×10^{-9} in the *A. arenosa* lineage. This is about half of the estimate of 1.5×10^{-8} for several *Arabidopsis* and *Arabis* species obtained by Koch, Haubold, and Mitchell-Olds (2000).

The ancestral state of indels could be determined by comparison with the outgroup sequence, with the exception of a single bp indel on fragment 4 for which no outgroup was available, and several complex rearrangements on fragment 1. *A. arenosa* had only experienced a slight reduction in size, but the *A. thaliana* sequences appeared to be 4% shorter than the ancestral state, which is similar to the 5% reduction in the *A. thaliana* intron size compared to *A. lyrata* previously found by Wright, Lauga, and Charlesworth (2002).

A. suecica contains two nuclear genomes, one from *A. thaliana* and one from *A. arenosa*. Thus, when sequencing, we expected to find two quite different fragments. When we sequenced the four fragments in *A. suecica*, four thaliana-like sequences and three arenosa-like sequences

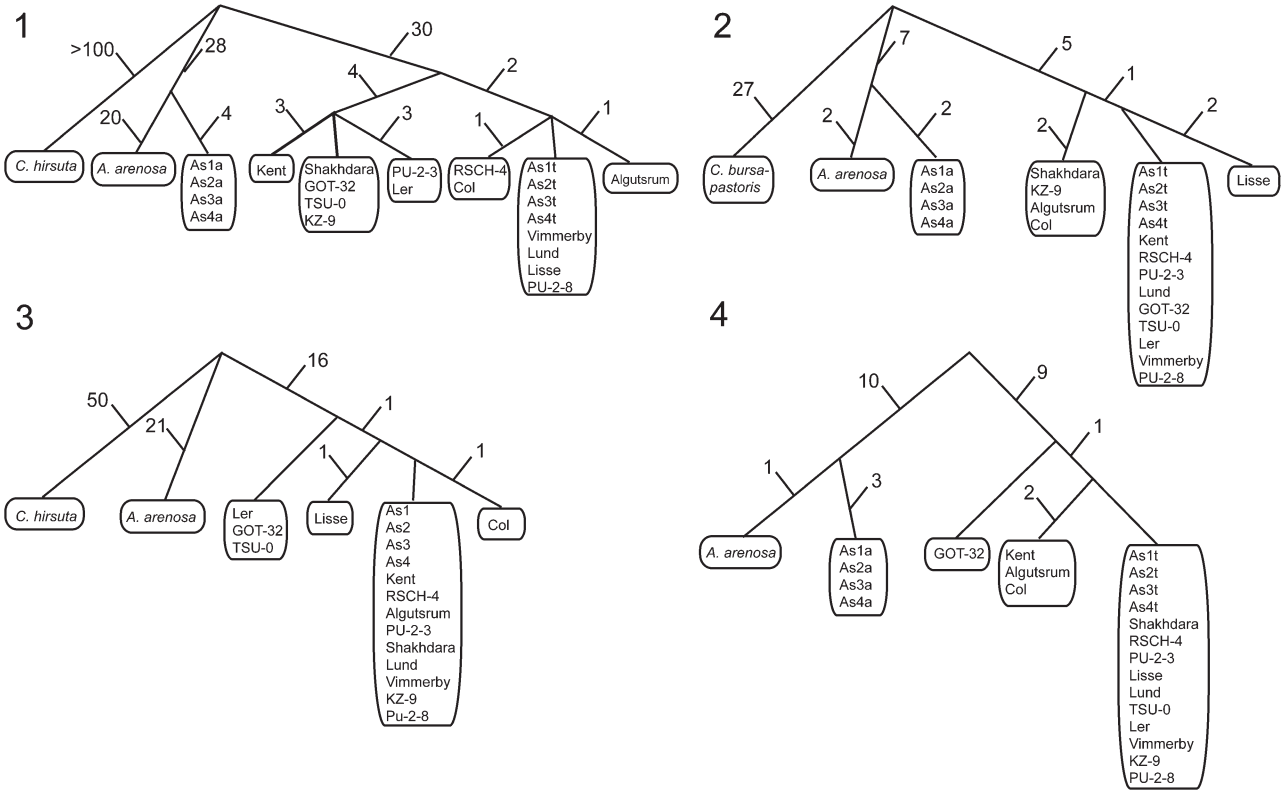

FIG. 2.—Inferred gene genealogies for each of the four sequenced fragments. As*i*a refers to the arenosa-like allele in *Arabidopsis suecica* individual *i*, As*i*t analogously for the thaliana-like allele (cf. table 3). Individuals not identified by scientific names are *Arabidopsis thaliana* accessions. The accessions Mt-0, Dem-4, NC-6, and Köln are identical to Col in all fragments; Tamm-46 and Kondara are identical to Shakhdara in all fragments. These six accessions have therefore been left out of the trees above. The number of mutations separating each allele is indicated on the branches.

were obtained (in total seven sequences; table 3). Figure 2 summarizes the observed pattern of variation as gene genealogies. The relationship between the *A. suecica* sequences and the ancestral species is evident. For each fragment, the four identical thaliana-like sequences fall within the cluster of actual *A. thaliana* sequences, and there are always *A. thaliana* sequences identical to the thaliana-like sequences. The arenosa-like sequences always cluster with *A. arenosa*, but they are different from the single *A. arenosa* allele we sequenced. It seems likely, however, that sequences identical to the four arenosa-like sequences would have been encountered, had we sequenced a larger sample of *A. arenosa*. Overall, *A. suecica* appears to have diverged little from its ancestral species, at least where homologous sequences exist. The fact that an arenosa-like allele was not successfully amplified from one of the four fragments could imply either that this allele does not amplify with our primers (although at least one *A. arenosa* allele does) or that the allele inherited from *A. arenosa* has been lost entirely for this fragment.

A total of 52 microsatellite loci were also typed for 14 *A. thaliana* and 8 *A. suecica* accessions. All 52 loci were variable within *A. thaliana* and 24 loci were variable in *A. suecica*. Because the microsatellites were chosen without any prior knowledge of variability in either of the two species, we can compare the variability of these two species using all microsatellite loci. *A. thaliana* is thus more variable than *A. suecica* considering all 52 loci (table 6). *A. thaliana* was still more variable, even if we only considered the 24 microsatellite loci that are variable within *A. suecica*. There were a total of 384 alleles among the 14 *A. thaliana* accessions and 106 among the eight *A. suecica* accessions. The comparable number of alleles for eight *A. thaliana* accessions would be 259.3 using a rarefaction correction (Kalinowski 2004) of the 14 *A. thaliana* accessions.

Number of Founders of *A. suecica*

We computed the likelihood of the *Model 2a* for 1–10 founders, using the microsatellite data. We also tested a wide range of values of the remaining parameters (μ_m, $N_s(0), N_s(\tau), N_t(0), N_t(\tau)$, τ, and α). However, because only two parameters, τ and $N_s(\tau)$, affected the likelihood for

Table 6
Levels of Variation for 8 *Arabidopsis suecica* and 14 *Arabidopsis thaliana* Individuals

Statistic	*A. thaliana*	*A. suecica*
Gene diversity (H_e)	0.796	0.273
Average pairwise difference (S')	238.3	39.8
Number of alleles	259.3 (384)	106

NOTE.—The gene diversity, average pairwise difference, and number of alleles of *A. thaliana* and *A. suecica* from 52 microsatellite loci. The number of alleles for *A. thaliana* has been rarefaction corrected to a sample size of eight and the total number of alleles for the 14 *A. thaliana* accessions is given in parenthesis.

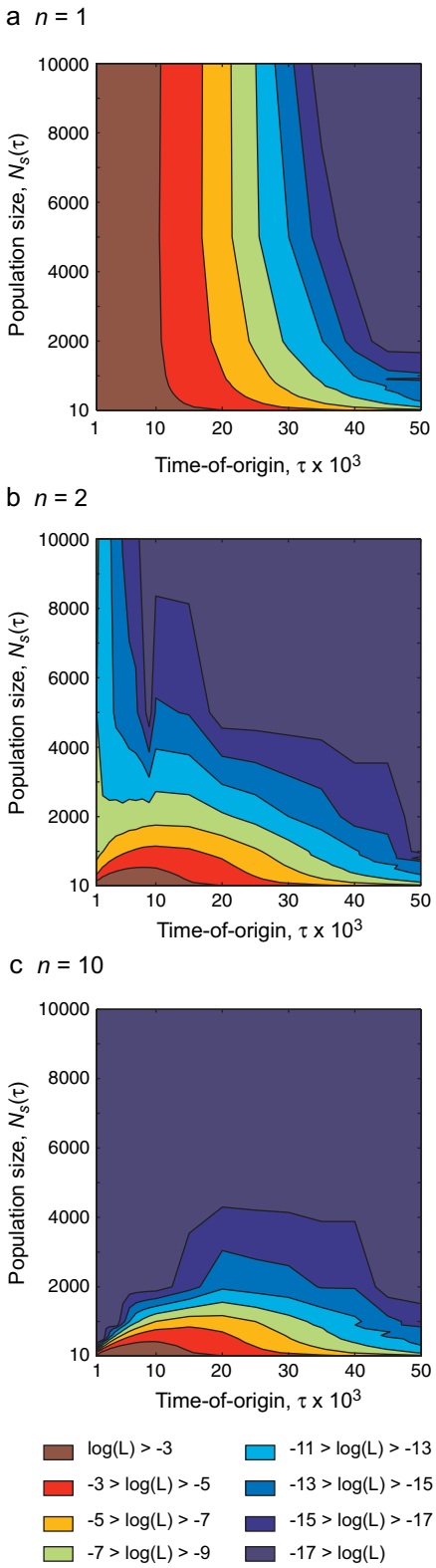

FIG. 3.—The likelihood surfaces for an n-founder model as functions of the time-of-origin, τ, and the population size of *Arabidopsis suecica* at the time-of-origin, $N_s(\tau)$. *Model 2a* and *Algorithm 2* was used to compute the likelihood surfaces (see text for details) with the following parameter values: $\mu_m \sim$ uniform(10^{-5}, 10^{-3}), $N_s(0) = 10^6$, $N_t(0) = 10^8$, $N_t(\tau) = 10^5$, and $\alpha = 1$. (a) $n = 1$. (b) $n = 2$. (c) $n = 10$.

different number of founders significantly (results not shown), we limit our discussion to these parameters. The parameter α affected the likelihoods moderately. A small value of α ($\alpha < 1$) resulted in a more evenly distributed likelihood as a function of τ (a more detailed analysis of the effect of α follows below). However, the effect on the likelihood distributions was almost identical for any tested number of founders. Figure 3 shows the likelihood surfaces for 1, 2, and 10 founders using *Model 2a* with remaining parameters held constant. From fig. 3 it is clear that if $N_s(\tau)$ is larger than about 500, we can be fairly certain that the species was founded by one event. However, there are clearly some parts of the [τ, $N_s(\tau)$] parameter space in which the likelihoods are very similar for 1, 2, and 10 founders. If $N_s(\tau)$ is less than about 500, the likelihoods are similar, except for small values of τ. The number of founders cannot be determined in this part of the parameter space. The reason that we cannot infer the number of founders for small values of N_s (τ) is that all *A. suecica* lines have an MRCA before (going backwards in time) τ, regardless of the number of founders. Figure 4 shows the posterior distribution of the number of founders ($n = 1, \ldots, 10$) for fixed parameter values of the population sizes and a wide range of values of the time-of-origin τ. Unless $N_s(\tau) < 1000$ there is strong support for one origin of *A. suecica*. There also appears to be little support for $\tau > 40{,}000$ for either value of n (see below for a more detailed estimation of the time-of-origin).

If there were more than one founder, these founders may have contributed unequal fractions to the founding population of *A. suecica*. We evaluated this scenario using *Model 2b* and *Algorithm 2*. Figure 5a shows the posterior distribution of the contributing fraction p_2 from a second founder, given the model, for four different values of $N_s(0)$ and $N_s(\tau)$. The second founding event was arbitrarily chosen to be the founder that contributed less than 0.5 to

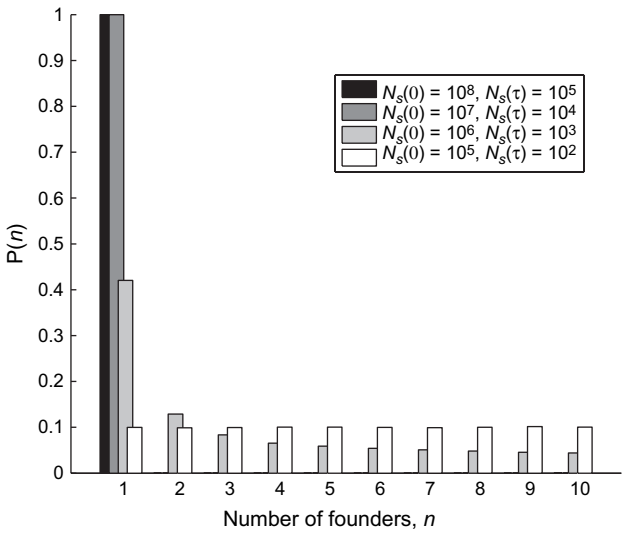

FIG. 4.—The posterior distribution of the number of founders n for different values of the population size of *Arabidopsis suecica*. *Model 2a* and *Algorithm 2* was used to compute the distribution for the following parameter values: $\tau \sim$ uniform(10^4, 10^5), $N_t(0) = 10^8$, $N_t(\tau) = 10^5$, $\alpha = 1$, and $\mu_m \sim$ uniform(10^{-5}, 10^{-3}).

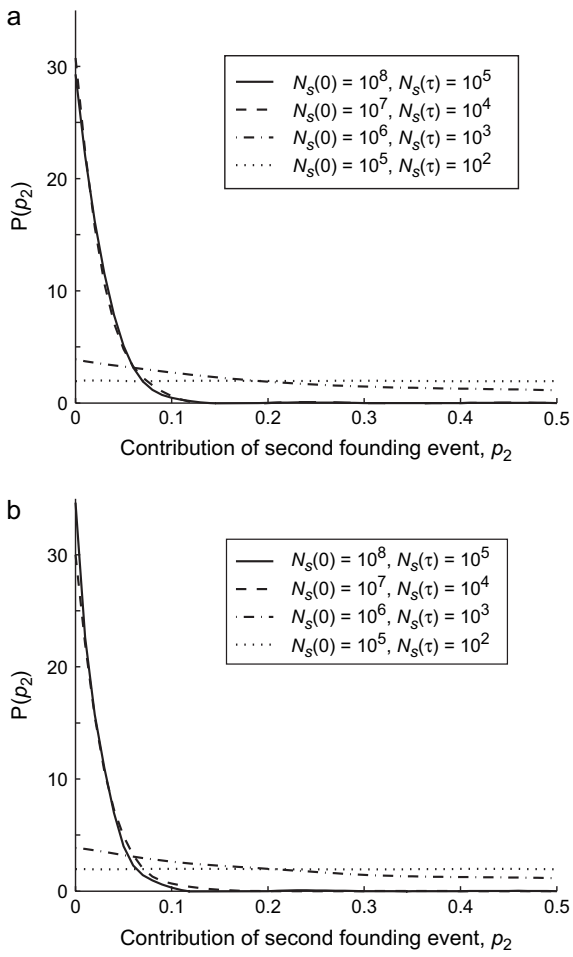

FIG. 5.—The posterior distribution of the fraction p_2 contributed by the second founder to the present population of *Arabidopsis suecica* (*Model 2b*). The posterior distributions are presented for four different values of the population size of *A. suecica* [$N_s(\tau)$ and $N_s(0)$] and the number of founders n were 2. The remaining parameters were $\tau \sim$ uniform(10^{-5}, 10^{-3}), $N_t(0) = 10^8$, $N_t(\tau) = 10^5$, $\alpha = 1$, in (a) $\mu_m \sim$ uniform(10^{-5}, 10^{-3}) and in (b) $\mu_m \sim$ uniform(10^{-4}, 10^{-3}).

the founding population. Note that this model is identical to *Model 2a* with $n = 2$ when $p_1 = p_2 = 0.5$, and it is identical to *Model 2a* with $n = 1$ when $p_2 = 0$ (or $p_1 = 0$). From fig. 5a we see that if $N_s(\tau) \geq 10^4$, most of the weight is on small values of p_2. It is thus unlikely that a second founder contributed more than 0.05–0.1 to the founding population of *A. suecica* given that $N_s(\tau) \geq 10^4$. If $N_s(\tau) = 10^3$ we see the same tendency of smaller values of p_2 being more likely than larger values of p_2, whereas if $N_s(\tau) = 10^2$ we cannot distinguish small contributions to the founding population from a second founder. Figure 5b shows the posterior distribution of the contributing fraction p_2 from a second founder given the data, and when the mutation rate is assumed to be between 10^{-4} and 10^{-3}. From fig. 5b it is clear that the general conclusions about a second founding event are robust with regard to the mutation rate.

The complete lack of variation in the *A. suecica* sequence data (four thaliana-like fragments and three arenosa-like fragments) also suggests that our geographically diverse sample of *A. suecica* stem from a single polyploidization event. Given the level of variation in *A. thaliana* (the average pairwise difference per bp was 0.00719, 0.0736, 0.00157 and 0.00301 in the four fragments respectively; see also fig. 2), it is unlikely that two or more polyploidization events would have involved *A. thaliana* parents identical at all four fragments (although we note that some *A. thaliana* accessions studied are indeed identical in all four fragments).

Time-of-origin

The time-of-origin of the nuclear genome of *A. suecica* was estimated using *Model 2*, *Algorithm 2*, and the microsatellite data. We used uniform distributions for the six priors [μ_m, τ, $N_t(0)$, $N_s(0)$, $N_t(\tau)$, $N_s(\tau)$] and tested different types of growth scenarios ($\alpha = 1.0$, 0.5, 0.25, or 0.1). The prior distribution of μ_m was uniform in the interval 10^{-5} to 10^{-3}. These bounds were based on microsatellite mutation rate estimates from a variety of organisms (Schug, MacKay, and Aquadro 1997; Vazquez *et al.* 2000; Xu *et al.* 2000; Vigouroux *et al.* 2002). The origin of *A. suecica* is unlikely to be more recent than a thousand years ago, given its present geographical distribution and level of variation. It is also unlikely that the origin is more than a million years ago given a divergence between its parents of about 5 Myr. Thus, we used a uniform distribution between 10^3 and 10^6 as a prior for τ. Strong evidence of population growth in *A. thaliana* has recently been reported (Nordborg *et al.* 2005) and we used uniform priors for the past and present population size of *A. thaliana* to reflect this: $N_t(0) \sim$ uniform(10^7, 10^9) and $N_t(\tau) \sim$ uniform(10^5, 10^7). For the population sizes of *A. suecica* we used uniform priors, $N_s(0) \sim$ uniform(10^5, 10^6) and $N_s(\tau) \sim$ uniform(1, 10^4). In addition to these "reasonable" priors (runs 1–4 in table 7), several more or less unrealistic priors were tested (runs 5–11, table 7) to check whether the priors chosen above were justified or not as control.

Figure 6a shows the posterior distribution of τ for four different growth scenarios and with a uniform prior distribution of μ_m between 10^{-5} and 10^{-3} using "reasonable" priors for all parameters (runs 1–4 in table 7). It is obvious from fig. 6a that the growth scenario affects the conclusions for the time-of-origin of *A. suecica*. If growth started close to the origin ($\alpha = 1$) the posterior probability distribution of τ is relatively narrow with a median of 45,000 years. If the growth started long after the origin ($\alpha = 0.25$ or 0.1) the posterior distributions of τ are quite wide with medians of 100,000 and 240,000. The probability that $\tau <$ 12,000 is about 5% for $\alpha = 1$ (where τ is larger at the 5% limit for all $\alpha < 1$, table 8). Figure 6b shows one minus the cumulative probability distribution for the same four growth scenarios. This figure illustrates that the upper limit of τ also depends heavily on α. When $\alpha \geq 0.25$ we find that the probability that $\tau > 290{,}000$ is about 5%. Figure 7 shows the heavy negative correlation of τ and μ_m. For example, if $\mu_m > 5 \times 10^{-5}$, then τ is <100,000. Clearly, inference of τ is heavily dependent on what is assumed about μ_m and the population growth, whereas the other parameter had much smaller effects on τ.

Estimates of microsatellite mutation rates from plants are in the range of 10^{-4} to 10^{-3} mutations per generation

Table 7
Prior Distributions of the Parameters That were Tested for Inferring the Time-of-origin Using *Algorithm 2*

Run	Prior distribution[a]	α	μ_m	τ	$N_t(0)$	$N_t(\tau)$	$N_s(0)$	$N_s(\tau)$
1	Uniform(a,b)	1	10^{-3}, 10^{-5}	10^3, 10^6	10^7, 10^9	10^5, 10^6	10^5, 10^7	1, 10^4
2	Uniform(a,b)	0.5	10^{-3}, 10^{-5}	10^3, 10^6	10^7, 10^9	10^5, 10^6	10^5, 10^7	1, 10^4
3	Uniform(a,b)	0.25	10^{-3}, 10^{-5}	10^3, 10^6	10^7, 10^9	10^5, 10^6	10^5, 10^7	1, 10^4
4	Uniform(a,b)	0.1	10^{-3}, 10^{-5}	10^3, 10^6	10^7, 10^9	10^5, 10^6	10^5, 10^7	1, 10^4
5	Uniform(a,b)	1	10^{-3}, 10^{-4}	10^3, 10^6	10^7, 10^9	10^5, 10^6	10^5, 10^7	1, 10^4
6	Uniform(a,b)	0.5	10^{-3}, 10^{-4}	10^3, 10^6	10^7, 10^9	10^5, 10^6	10^5, 10^7	1, 10^4
7	Uniform(a,b)	0.25	10^{-3}, 10^{-4}	10^3, 10^6	10^7, 10^9	10^5, 10^6	10^5, 10^7	1, 10^4
8	Uniform(a,b)	0.1	10^{-3}, 10^{-4}	10^3, 10^6	10^7, 10^9	10^5, 10^6	10^5, 10^7	1, 10^4
9[b]	Uniform(a,b)	0.1	10^{-4}, 10^{-5}	10^2, 10^6	1, 10^9	1, 10^6	1, 10^9	1, 10^6
10[b]	Uniform(a,b)	0.1	10^{-3}, 10^{-5}	10^2, 10^6	1, 10^9	1, 10^6	1, 10^9	1, 10^6
11[b]	Uniform(a,b)	0.1	10^{-6}, 10^{-4}	10^2, 10^6	1, 10^9	1, 10^6	1, 10^9	1, 10^4

[a] This is the prior distribution used for μ_m, τ, $N_t(0)$, $N_t(\tau)$, $N_s(0)$, and $N_s(\tau)$ between a and b, which are presented under each parameter in the table.

[b] For these runs were the acceptance criteria $0.25\times(S'_{t,\text{obs}}+S'_{s,\text{obs}})$ instead of $0.05\times(S'_{t,\text{obs}}+S'_{s,\text{obs}})$.

(Vigouroux *et al*. 2002). If we assume that μ_m is between 10^{-4} and 10^{-3}, the upper bound on τ would be much smaller; $\tau > 44{,}000$ is about 5% for $\alpha \geq 0.25$ (fig. 8). Under this assumption on μ_m the probability distributions of τ are shifted toward the lower end of the prior of τ. For example, the median of τ when $\alpha = 1$, 0.5, and 0.25 is about 9,000, 13,000, and 22,000, respectively.

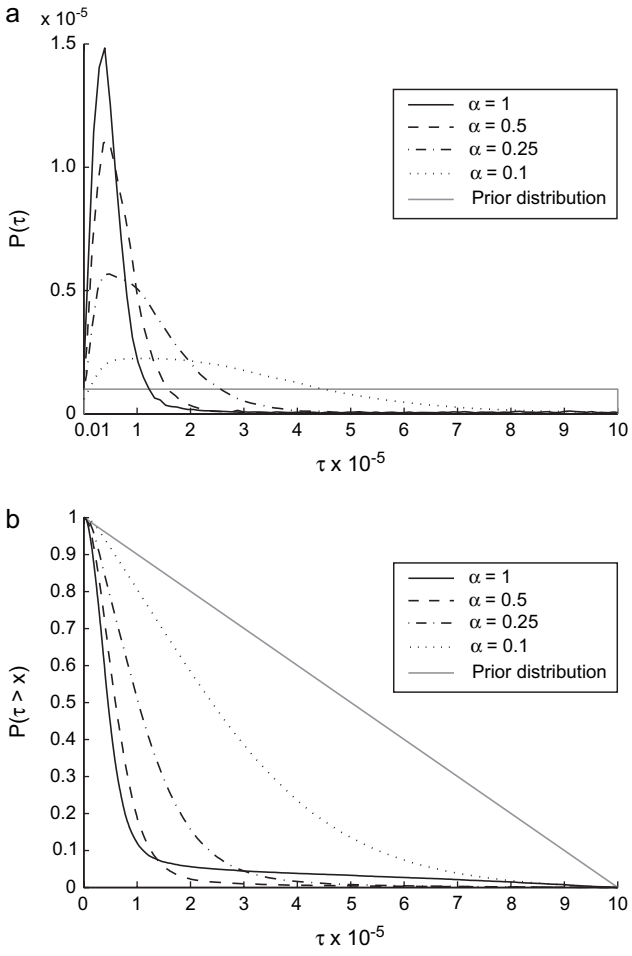

FIG. 6.—Posterior distributions of τ for four different growth scenarios, $\alpha = 1.0$ (solid), 0.5 (dashed), 0.25 (dotted-dashed), and 0.1 (dotted), with a prior mutation rate $\mu_m \sim$ uniform(10^{-5}, 10^{-3}) under *Model 2*. The prior of τ is shown as a solid gray line. (a) Probability distribution of τ; (b) one minus the cumulative probability distribution τ.

Given *Model 1* described above, what does the fact that we observed no variation in the four sequenced fragments tell us about τ, N, and μ_s? It is clear that the information in these data is limited: the maximum likelihood estimate of all three parameters is zero, and the best we can hope for is to rule out certain parameter combinations. A similar problem has been studied previously in the context of a human Y chromosome data set without variation (Tavaré *et al*. 1997). Figure 9 shows the posterior distributions using *Algorithm 1* for each of the three parameters τ, N, and μ_s. To emphasize the fact that we can really only hope to rule out large values, one minus the cumulative distribution is shown. For example, the probability that $\tau >$ 200,000 years is about 20% and the probability that $\tau >$ 650,000 years is about 5%. It is clear that a wide range of values are plausible for all three parameters. It is also clear that there is almost no information about N: the cumulative posterior distribution is a straight line, which is the same as the prior. The only exception is for very small values of N. The reason for this is that, with the exception of very small N, observing no variation is extremely unlikely unless either τ or μ_s is small enough. A small effective population size is not a good explanation for the data.

It is much more likely that the lack of variation reflects the relatively recent origin of *A. suecica*. It may seem

Table 8
The 90% Credibility Regions of μ_m and τ of the Posterior Distributions Obtained from *Algorithm 2* for Different Growth Scenarios, α, and Mutation Rates, μ_m

		Prior	Posterior (90% interval)	
Run	α	$\mu_m \times 10^5$	$\mu_m \times 10^5$	$\tau \times 10^{-3}$
1	1	1, 100	1.6, 7.0	11.9, 248
2	0.5	1, 100	1.8, 8.2	15.1, 157
3	0.25	1, 100	1.9, 9.4	20.7, 291
4	0.1	1, 100	1.9, 10.3	35.6, 663
5	1	10, 100	10.2, 18.7	2.8, 14.9
6	0.5	10, 100	10.2, 19.6	4.8, 25.1
7	0.25	10, 100	10.2, 19.4	7.8, 44.2
8	0.1	10, 100	10.2, 19.7	13.7, 107

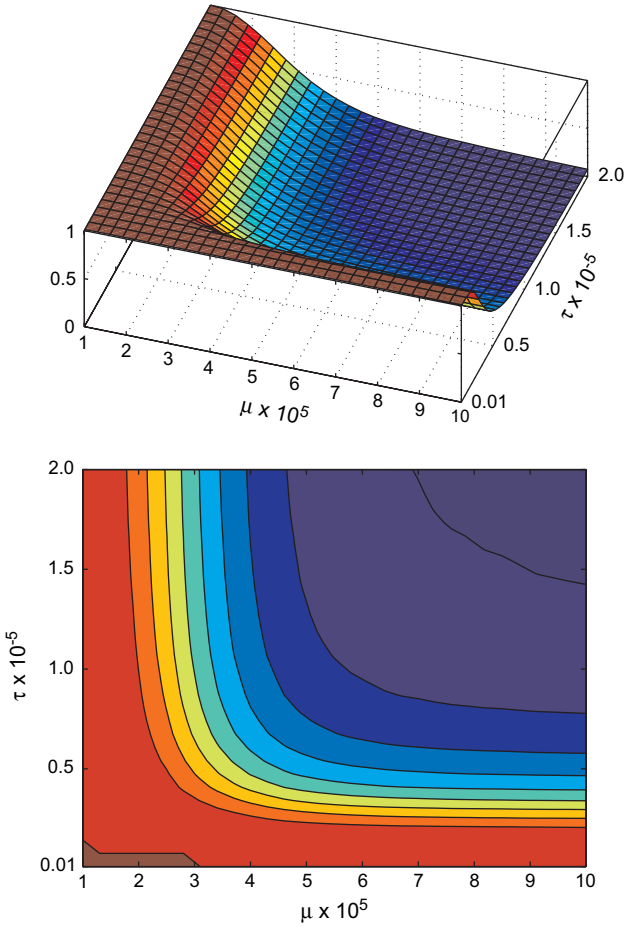

FIG. 7.—Joint cumulative posterior density of τ and μ_m under *Model 2* (run 1, table 7), where the prior of μ_m was between 10^{-5} and 10^{-3} and $\alpha = 1.0$.

disappointing that we cannot pin down this origin more narrowly than indicated in fig. 9. The reason for this is the uncertainty in μ_s. Similar to what was observed for the microsatellite model (fig. 7), the joint posterior density

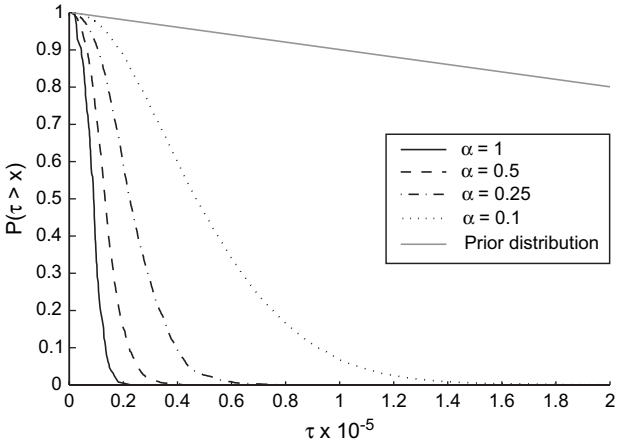

FIG. 8.—Cumulative probability distribution τ when the mutation rate, μ_m, is between 10^{-4} and 10^{-3} under *Model 2*. Four different growth scenarios are shown, $\alpha = 1.0$ (solid), 0.5 (dashed), 0.25 (dotted-dashed), and 0.1 (dotted). The prior of τ is shown as a solid gray line.

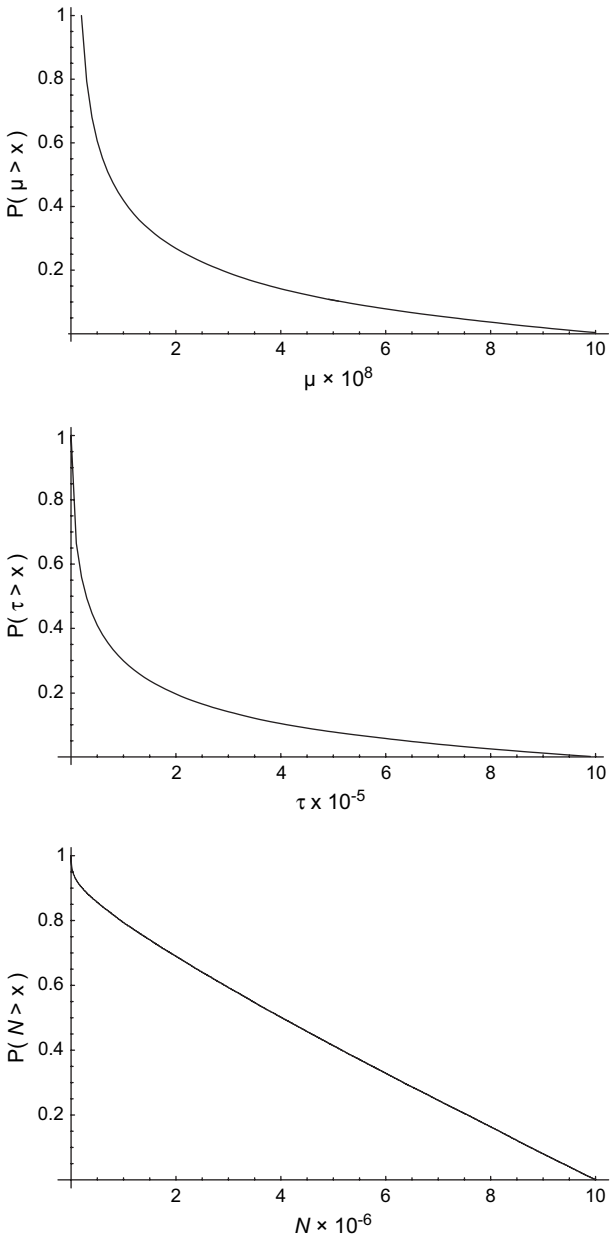

FIG. 9.—Cumulative posterior distributions of μ_s, τ, and N, estimated under *Model 1* (see text for details).

of τ and μ_s show a very strong negative correlation between the two parameters (fig. 10). We originally assumed a uniform prior between 10^{-9} and 10^{-7} for μ_s based on mutation rates estimated above and earlier reports of mutation rates of *Arabidopsis* and *Arabis* species (e.g., Koch, Haubold, and Mitchell-Olds 2000). If we were willing to assume that $\mu_s > 10^{-8}$, the marginal posterior distribution for τ would change dramatically. As shown in fig. 11, the probability that $\tau > 200{,}000$ years is now about 7% instead of the 20% we obtained earlier. Figure 11 illustrates how strongly these kinds of evolutionary inferences depend on what is assumed. It is clear that with a high μ_s the estimate of τ will be low and visa versa.

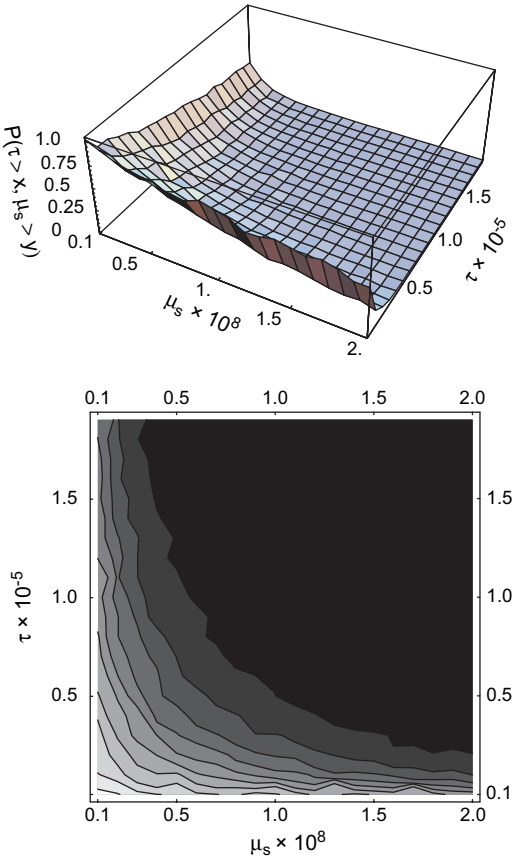

FIG. 10.—Joint cumulative posterior density of τ and μ_s under *Model 1* (see text for details).

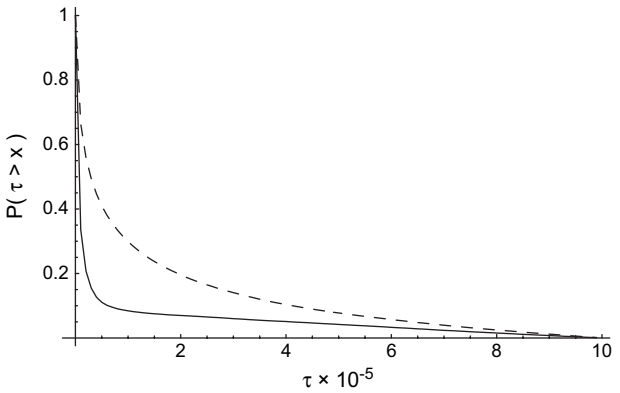

FIG. 11.—Cumulative posterior distribution for τ under different assumptions about μ_s under *Model 1*. The dashed line is the posterior distribution of τ when $10^{-9} < \mu_s < 10^{-7}$ and the solid line represents the posterior distribution when $10^{-8} < \mu_s < 10^{-7}$.

Discussion

Number of Origins

Unlike other forms of speciation, speciation by polyploidization is instantaneous. It is the result of a well-defined, discrete event (or events), and in contrast to, say, humans it is clearly meaningful to ask when and where the species arose. The new polyploid species is expected to be isolated from its parental species because crosses between a polyploid and its parental species are most likely infertile. However, repeated polyploidization events may result in a polyploid species with multiple origins. If alleles at a given locus of all existing individuals of a polyploid species have an MRCA before (going backwards in time) or at the time-of-origin, at least two scenarios are possible. These alleles either have an MRCA before the time-of-origin simply because of drift, or the species was founded by one polyploidization event. In addition to these two scenarios, the polyploid species may be the result of more that one polyploidization event of genetically very closely related parental individuals. This scenario would be very difficult if not impossible to distinguish genetically from a scenario of only one founding event. On the other hand, this case would lead to a polyploid species which, effectively, can be considered to be of a single origin because the founding individuals were genetically very similar. Distinguishing between single and multiple origins using polymorphism data is not trivial. In particular, the observation of monophyly at one or a small number of loci with respect to the parental species does not imply a single origin: alleles in the polyploid species could be monophyletic with respect to the parental species simply because they happen to coalesce more recently than the origins of the species (Rosenberg and Nordborg 2002).

We investigated the number of origins of *A. suecica* using an explicit coalescent model. We found that the likelihood of different founder numbers was determined by two parameters: the time-of-origin, τ, and the population size of *A. suecica* immediately after the founding, $N_s(\tau)$. We concluded that there is power to infer a single founder if $N_s(\tau) \geq 500$. It is worth clarifying what this means (cf. fig. 3). First, it is obvious that we cannot say anything about the number of founders if τ is sufficiently large because all lineages will have coalesced with *A. suecica*. However, the low level of polymorphism in *A. suecica* argues against a very large τ (unless the population size is tiny). Second, we are not interested in founder lines that went extinct in a small number of generations. To model multiple origins, we consider the composition of *A. suecica* immediately after the most recent founder event that resulted in an established population. We define the population size of *A. suecica* at this time $N_s(\tau)$. Thus, the statement that there is power to infer a single founder if $N_s(\tau) \geq 500$ means that we can reject multiple founders unless the total (effective) population size when multiple lineages last existed was less than 500. This seems unlikely given the high reproductive rate of *Arabidopsis*, and the fact polyploidization events are rare so that each founding event is likely to have been isolated (geographically) from other founding events for some time. An initial population size of 500 is quickly reached given the mode of reproduction of *A. suecica* (Säll *et al.* 2004).

For simplicity, we do not explicitly model multiple origins at different points in time (our method tests for multiple origins at one point in time). Nonetheless, our results bear on this (rather more plausible scenario) as well. In particular, our "unequal contribution" model (*Model 2b*) can be thought of as modeling a situation where two founding populations were established at different points in time, and have reached different sizes. We then consider the future of the merged population. We believe that the general

conclusions we reach about what can, and cannot be inferred about multiple founding events would be valid under more detailed models as well. However, an explicit model of multiple founding events at different points in time would require many assumptions about initial demography, and is beyond the scope of this paper.

In summary, then, we believe that the current genome of *A. suecica* stems from a single founding event. While it is possible that some variation was inherited from *A. arenosa* (and survived the original bottleneck that must have taken place), it seems likely that the genome inherited from *A. thaliana* was originally that of an homozygous inbred parent.

Time-of-origin

The coalescent method was also used to estimate the time-of-origin of *A. suecica*. This was done separately using microsatellite data and sequence data. The results were more sensitive to assumptions than those for the number of founders. The microsatellite data show that the origin probably is >12,000 years ago, unless the microsatellite mutation rate is high (if $\mu_m > 10^{-4}$, then we can only conclude that the origin most likely is >3,000 years ago). The upper limit of the origin is almost certainly <290,000 years ago (based on the microsatellite data). Figures 6, 8, and 11 illustrate how strongly these kinds of evolutionary inferences depend on the assumptions. If we are willing to assume that the mutation rates are in the higher end of the intervals ($\mu_m > 10^{-4}$ and $\mu_s > 10^{-8}$), then the upper bounds on the time-of-origins will be much lower, $\tau < 44{,}000$ for the microsatellites and $\tau < 200{,}000$ for the sequence data. An advantage of the Bayesian approach taken here is that it provides a natural way of making uncertainty about parameters explicit. Our estimate of μ_s based on divergence between *A. thaliana* and *A. arenosa* is about 6×10^{-9}, but this estimate has considerable uncertainty. In our simulations, we assumed a uniform distribution for μ_s: we could have chosen some other distribution and obtained a different posterior distribution. This may seem arbitrary, but the correct conclusion is that there is not enough information in the sequence data. It should be remembered, however, that most estimates of divergence times in evolutionary biology treat the mutation rate as constant, completely ignoring the uncertainty.

There are strong genetic indications of growth for the *A. thaliana* population (Nordborg *et al.* 2005) and the *A. suecica* population must necessarily have grown since its origin. It is plausible that the growth has occurred in connection with the retreat of the ice cover of the last glaciation because both these species are presently found in areas that were completely or partly covered by ice. For scenarios when the growth started relatively late ($\alpha = 0.25$ or 0.1) we find that the time-of-origin (and thus the real starting time of growth) is further back in time than for scenarios when the growth started relatively early ($\alpha = 0.5$ or 1; figs. 6 and 8). Thus the real starting time of the growth for the different growth scenarios tends to be shifted toward a common point in time.

Estimates of the age of the most recent common ancestor of the *A. suecica* chloroplast genome (Säll *et al.* 2003; Jakobsson *et al.* 2005) are consistent with the estimates obtained here.

Place of Origin

Because the present range of *A. suecica* was covered by ice during the last glaciation (Andersen and Borns 1997), and because we estimate the origin to be >12,000 years ago, the species have probably originated somewhere south of Sweden and Finland and then migrated to its present range. If this is the case, then, because *A. suecica* is not found outside Sweden and Finland today, the species must have become extinct in the areas where it originated.

The fragment genealogies in fig. 2 reveal that the thaliana-like sequence from *A. suecica* always falls within clades of sequences from *A. thaliana*. As expected given the level of linkage disequilibrium in *A. thaliana* (Nordborg *et al.* 2002), the thaliana-like sequence clusters with different *A. thaliana* accessions in different fragments. However, it is always identical to the two accessions Lund and Vimmerby from southern Sweden as well as the accession Pu-2-8 from the Czech Republic. Studies of cpDNA have found that *A. suecica* clusters with *A. thaliana* originating from southern Sweden and central Europe (Säll *et al.* 2003; Jakobsson *et al.* 2005). In subsequent studies, we found many *A. thaliana* accessions from southern Sweden, as well as some from middle Sweden, and central and southern Europe, that carry a sequence identical to the thaliana-like sequences at fragments 1 and 4 (Hagenblad *et al.* 2004). Moreover, all *A. suecica* accessions cluster together when the accessions in the microsatellite data are analyzed using a Neighbor-Joining algorithm (Saitou and Nei 1987; data not presented). The *A. thaliana* accessions clustering closest to the *A. suecica* branch were from southern and central Sweden (T160, T340 and T350), Finland (T81), western Russia (T93), Poland (Lip-0), and Lithuania (Wil-0). These observations suggest that *A. suecica* migrated to Sweden and Finland in the wake of the retreating ice *together* with *A. thaliana* individuals related to its ancestor followed by extinction of the species outside Sweden and Finland.

Supplementary Material

The new sequences reported in this paper are available from GenBank with accession numbers AY319334–AY319368. *A. thaliana* sequences are available from GenBank with accession numbers AY092437–AY092456, AY092477–AY092496, AY092517–AY092536, and AY092637–AY092656. The above data and Table S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors wish to thank H. Ellegren for letting us use his sequencing equipment at an early stage of this project and N. Arnheim for suggesting the use of allele-specific PCR. We would also like to thank D. Charlesworth, M. Uyenoyama, and two anonymous reviewers for comments on the manuscript. S.T. is a Royal Society Wolfson Research Merit Award holder.

Literature Cited

Andersen, B., and H. Borns Jr. 1997. The ice age world. Scandinavian University Press, Oslo.

Benson, G. 1999. Tandem repeats finder: a program to analyse DNA sequences. Nucleic Acids Res. **27**:573–580.

Beaumont, M. A., W. Y. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics **162**:2025–2035.

Chen, Z. J., L. Comai, and C. S. Pikaard. 1998. Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in *Arabidopsis* allopolyploids. Proc. Natl. Acad. Sci. USA **95**:14891–14896.

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

Hagenblad, J., and M. Nordborg. 2002. Sequence variation and haplotype structure surrounding the flowering time locus FRI in *Arabidopsis thaliana*. Genetics **161**:289–298.

Hagenblad J., C. L. Tang, J. Molitor, J. Werner, K. Zhao, H. G. Zheng, P. Marjoram, D. Weigel, and M. Nordborg. 2004. Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. Genetics **168**:1627–1638.

Heenan, P. B., A. D. Mitchell, and M. Koch. 2002. Molecular systematics of the New Zealand Pachycladon (Brassicaceae) complex: generic circumscription and relationships to *Arabidopsis* sens. lat. and *Arabis* sens. lat. N. Z. J. Bot. **40**: 543–562.

Hey, J. 2005. On the number of new world founders: a population genetic portrait of the peopling of the Americas. PLoS Biol. **3**:965–975.

Hultgård, U. 1987. Parnassia palustris L. in Scandinavia. Symb. Bot. Ups. **28**:1–128.

Hultén, E. 1971. Atlas of the distribution of vascular plants in northwestern Europe. Generalstabens litografiska anstalts förlag, Stockholm.

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

Jakobsson, M., T. Säll, C. Lind-Halldén, and C. Halldén. 2005. The evolutionary history of the common chloroplast genome of *Arabidopsis thaliana* and *A. suecica*. in Genome Divergence in Progress - a population genetic analysis of the allopolyploid *Arabidopsis suecica* and its maternal parent *A. thaliana*. Doctoral dissertation (M. Jakobsson), Lund University, Lund, Sweden.

Kalinowski, S. T. 2004. Counting alleles with rarefaction: private alleles and hiearchical sampling design. Conserv. Genet. **5**:539–543.

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

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

Ku, H., T. Vision, J. Liu, and S. Tanksley. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: largescale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. USA **97**: 9121–9126.

Kuittinen, H., and M. Aguadé. 2000. Nucleotide variation at the Chalcone isomeras locus in *Arabidopsis thaliana*. Genetics **155**:863–872.

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

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

Nordborg, M. 1998. On the probability of Neanderthal ancestry. Am. J. Hum. Genet. **63**:1237–1240.

———. 2001. Coalescent theory. Pp. 179–212 *in* D. J. Balding, M. J. Bishop, and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, Inc., Chichester, UK.

Nordborg, M., and J. Bergelson. 1999. The effect of seed and rosette cold treatment on germination and flowering time in some *Arabidopsis thaliana* Brassicaceae ecotypes. Am. J. Bot. **86**:470–475.

Nordborg, M., J. Borevitz, J. Bergelson et al. (12 co-authors). 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. **30**:190–193.

Nordborg M., T. Hu, Y. Ishino et al. (23 co-authors). 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. **3**:1289–1299.

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

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

Osborn, T., J. Pires, J. Birchler et al. (11 co-authors). 2003. Understanding mechanisms of novel gene expression in polyploids. TIG **19**:141–147.

Ohta, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22**:201–204.

Pontes O., N. Neves, M. Silva, M. S. Lewis, A. Madlung, L. Comai, W. Viegas, and C. S. Pikaard. 2004. Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid *Arabidopsis suecica* genome. Proc. Natl. Acad. Sci. USA **101**:18240–18245.

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

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

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

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

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

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

Schug, M. D., T. F. C. MacKay, and C. F. Aquadro. 1997. Low mutation rates of microsatellite loci in *Drosophila melanogaster*. Nat. Genet. **15**:99–102.

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

Song, K., P. Lu, K. Tang, and T. C. Osborn. 1995. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc. Natl. Acad. Sci. USA **92**:7719–7723.

Stebbins, G. L. 1971. Chromosomal evolution in higher plants. E. Arnold, London.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly. 1997. Inferring coalescence times from DNA sequence data. Genetics **145**:505–518.

Vazquez, F., T. Perez, J. Albornoz, and A. Dominguez. 2000. Estimation of the mutation rates in Drosophila melanogaster. Genet. Res. **76**:232–326.

Vigouroux Y, J. S. Jaqueth, Y. Matsuoka, O. S. Smith, W. F. Beavis, J. S. C. Smith, and J. Doebley. 2002. Rate and pattern of mutation at microsatellite loci in maize. Mol. Biol. Evol. **19**:1251–1260.

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

Wright, S., B. Lauga, and D. Charlesworth. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol. Biol. Evol. **19**:1407–1420.

Xu, X., M. Peng, Z. Fang, and X. Xu. 2000. The direction of microsatellite mutations is dependent upon the allele length. Nat. Genet. **24**:396–399.

Yang, Y.-W., K.-N. Lai, P.-Y. Tai, D.-P. Ma, and W.-H. Li. 1999. Molecular phylogenetic studies of Brassica, Rorippa, *Arabidopsis* and allied genera based on the internal transcribed spacer region of 18S–25S rDNA. Mol. Phylogenet. Evol. **13**:455–462.

Marcy Uyenoyama, Associate Editor

Accepted March 16, 2006