Multilocus analysis of genetic divergence between outcrossing *Arabidopsis* species: evidence of genome-wide admixture

Wei-Kuang Wang[1], Chuan-Wen Ho[1], Kuo-Hsiang Hung[2], Kuo-Hsiung Wang[1], Chi-Chun Huang[1], Hitoshi Araki[3], Chi-Chuan Hwang[4], Tsai-Wen Hsu[1], Naoki Osada[5] and Tzen-Yuh Chiang[1]

[1]Department of Life Sciences, National Cheng-Kung University, Tainan 701, Taiwan; [2]Graduate Institute of Bioresources, Pingtung University of Science and Technology, Pintung 912, Taiwan; [3]Department of Fish Ecology and Evolution, Center for Ecology, Evolution and Biogeochemistry, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Kastanienbaum CH–6047, Switzerland; [4]Department of Engineering Science, National Cheng-Kung University, Tainan 701, Taiwan; [5]Department of Population Genetics, National Institute of Genetics, Yata, Mishima, Shizuoka 411-8540, Japan

Authors for correspondence:
Naoki Osada
Tel: +81 55 981 5820
Email: nosada@lab.nig.ac.jp

Tzen-Yuh Chiang
Tel: +886 6 2742583
Email: tychiang@mail.ncku.edu.tw

Received: *25 April 2010*
Accepted: *7 June 2010*

New Phytologist (2010) **188**: 488–500
doi: 10.1111/j.1469-8137.2010.03383.x

Key words: *Arabidopsis*, coalescent model, genomic divergence, introgression, multilocus analysis, polyploidy, reticulate evolution, speciation.

Summary

- Outcrossing *Arabidopsis* species that diverged from their inbreeding relative *Arabidopsis thaliana* 5 million yr ago and display a biogeographical pattern of interspecific sympatry vs intraspecific allopatry provides an ideal model for studying impacts of gene introgression and polyploidization on species diversification.
- Flow cytometry analyses detected ploidy polymorphisms of 2× and 4× in *Arabidopsis lyrata* ssp. *kamchatica* of Taiwan. Genomic divergence between species/subspecies was estimated based on 98 randomly chosen nuclear genes. Multilocus analyses revealed a mosaic genome in diploid *A. l. kamchatica* composed of *Arabidopsis halleri*-like and *A. lyrata*-like alleles.
- Coalescent analyses suggest that the segregation of ancestral polymorphisms alone cannot explain the high inconsistency between gene trees across loci, and that gene introgression via diploid *A. l. kamchatica* likely distorts the molecular phylogenies of *Arabidopsis* species. However, not all genes migrated across species freely. Gene ontology analyses suggested that some nonmigrating genes were constrained by natural selection.
- High levels of estimated ancestral polymorphisms between *A. halleri* and *A. lyrata* suggest that gene flow between these species has not completely ceased since their initial isolation. Polymorphism data of extant populations also imply recent gene flow between the species. Our study reveals that interspecific gene flow affects the genome evolution in *Arabidopsis.*

Introduction

Hybridization and polyploidization usually lead to rapid genomic changes and are considered key mechanisms of speciation and diversification in plants (Baack & Rieseberg, 2007; Mallet, 2007). Despite the occurrence of gene exchange with congeners, most hybridizing plant species remain morphologically discrete (Yatabe *et al.*, 2007). Recent advances in molecular genetic techniques enable botanists to analyse the diversification process via a comparative genomics approach (Mitchell-Olds & Clauss, 2002; Bomblies & Weigel, 2007) and this has led to several classic diploid species being shown to exhibit genome-wide duplications and/or other genomic signatures of past hybridization events (Wendel, 2000). Nevertheless, empirical data on species delimitation and interspecific hybridization are often difficult to interpret because both shared ancestral polymorphisms and gene introgression after speciation can result in species paraphyly and admixture in genetic composition (Muir & Schlötterer, 2005; Lexer *et al.*, 2006; Chiang *et al.*, 2009).

Arabidopsis thaliana is the first higher plant whose complete genome has been sequenced (AGI, 2000; Clauss & Koch, 2006). As sisters to *A. thaliana*, nonmodel *Arabidopsis* species have been recently used to elucidate the genetics of speciation and molecular evolution of diagnosable traits (Clauss & Koch, 2006). *Arabidopsis lyrata* and *A. halleri* split from *A. thaliana c.* 5 million yr ago (Al-Shehbaz & O'Kane, 2002), and are believed to have diverged from each other *c.* 2 Ma (Koch & Matschinger, 2007). Morphologically, these *Arabidopsis* species are well differentiated. *Arabidopsis lyrata* is a rosette-like herb, with no orbicular lobing of basal leaves, whereas, *A. halleri* is stoloniferous with orbicular to suborbicular terminal lobes of basal leaves (Al-Shehbaz & O'Kane, 2002; Clauss & Koch, 2006). Nevertheless, both of these perennial, outcrossing species retain high genetic similarities with annual, selfing *A. thaliana*, while differing in some biological characteristics, such as self incompatibility and heavy metal tolerance (Mitchell-Olds, 2001; Hall *et al.*, 2002).

Geographically, *A. halleri* ssp. *gemmifera* is distributed in northeast Asia, while its conspecific sister *A. halleri* ssp. *halleri* is mainly distributed in Europe. Geographical barriers of the Tienshan Mountain Range isolate these two intraspecific taxa from each other (Hegi, 1986; Miyashita *et al.*, 1998; Savolainen *et al.*, 2000). The phylogenetic relationship of the *A. lyrata* species complex is more complicated than that of *A. halleri* owing to polymorphisms in ploidy and possible hybridization between species. Al-Shehbaz & O'Kane (2002) consider that it is composed of three subspecies, ssp. *lyrata*, *petraea* and *kamchatica*. *Arabidopsis l. lyrata* is found only in North America, while *A. l. petraea* occurs in northern Eurasia and central Europe (Al-Shehbaz & O'Kane, 2002; Mitchell-Olds & Clauss, 2002). Both *A.l. lyrata* and *A. l. petraea* have diploid and tetraploid forms. *Arabidopsis l. kamchatica* occurs in Russia, Japan, Taiwan and northwest North America, and there is good evidence that it is an allotetraploid of *A. halleri* and *A. lyrata* (Ramos-Onsins *et al.*, 2004; Shimizu *et al.*, 2005; Shimizu-Inatsugi *et al.*, 2009) which has originated independently on several different occasions (Shimizu-Inatsugi *et al.*, 2009; Schmickl *et al.*, 2010). Because of the nature of allotetraploidy, Shimizu *et al.* (2005) elevated *A. l. kamchatica* to the species level (i.e. *A. kamchatica*), which includes an additional taxon of ssp. *kawasakiana*. Elven *et al.* (2007) recognized *A. petraea* and three intraspecific taxa: ssp. *petraea*, *septentrionalis* and *umbrosa*. Interestingly, in Taiwan, where *A. l. kamchatica* and *A. h. gemmifera* are occasionally sympatric, both diploid and tetraploid forms of *A. l. kamchatica* are recognized. Hayata (1911) discovered an *Arabidopsis* accession at Yushan Mountain (Taiwan) and named it as *Arabis morrisonensis*, which is characterized by long stolons, but this was later synonymized to *A. l. kamchatica* (Inoue, 1971; O'Kane & Al-Shehbaz, 1997) and was recognized as an elongated form with a prostrate habit. Based on the morphology of the type specimen (Taiwan, Mt Morrison, Nov. 1905, S. Nagasawa 680, Holotype: TI) and the evidence of flow cytometry analyses, which showed this 'morrisonensis' morph to be diploid (see later), we refer to this plant hereafter as a diploid form of *A. l. kamchatica*.

Stable allotetraploid and diploid forms derived from identical progenitors may represent different lineages, although might be given the same binome (Mahelka *et al.*, 2007; Tateishi *et al.*, 2007). An allotetraploid tends to be reproductively isolated from its parental species as their triploid offspring are often sterile, whereas diploid hybrids have higher probabilities of backcrossing leading possibly to hybrid swarm formation. However, gene introgression via hybrid bridges is unlikely to occur between highly divergent species.

A multilocus analysis provides genealogical information of a species history and the power to discern various evolutionary forces that have acted in the past (Wright *et al.*, 2002; Städler *et al.*, 2005). The advent of the genome sequence of *A. thaliana* allows one to design such multilocus analyses of the relatives of this species. This approach has been extensively applied when examining divergence and relationships in primate and *Drosophila* species (Kliman *et al.*, 2000; Hey & Nielsen, 2004), as well as to plants, such as tomatoes (Städler *et al.*, 2005), sunflowers (Strasburg & Rieseberg, 2008), silverswords (Lawton-Rauh *et al.*, 2007), palms (Trenel *et al.*, 2008), sedges (King & Roalson, 2009) and maize (Tiffin & Gaut, 2001).

Using rigorous statistical tests based on the coalescent theory, we have focused on examining divergence between four species/subspecies related to *A. thaliana*, namely *A. l. lyrata*, *A. l. kamchatica*, *A. h. halleri* and *A. h. gemmifera* (after O'Kane & Al-Shehbaz, 2003). In doing so, we have addressed the following questions: Given long divergence time between species, are gene genealogies consistent among different loci? Does the diploid *A. l. kamchatica* that shares high morphological similarities with the allotetraploid form also possess a mosaic genome derived from the putative parental species? Has gene flow between *A. lyrata* and *A. halleri* completely ceased since their divergence? (That is, did the species derived from regular branching processes maintain high similarity in the genome?)

Materials and Methods

Sampling

Four relatives of *A. thaliana*, *A. h.* ssp. *gemmifera*, *A. h.* ssp. *halleri*, *A. l.* ssp. *kamchatica* (both diploid and tetraploid forms) and *A. l.* ssp. *lyrata*, were examined. For population genetic analysis, ten random samples were collected from each of three populations of *A. l. kamchatica* (two from Taiwan and one from Japan), two populations of *A. l. lyrata* (from the USA), and three populations of *A. h. gemmifera* (two from Taiwan and one from China) (Table 1). Nucleotide sequences of *A. h. halleri* were obtained from Ruggiero *et al.* (2008) (DDBJ/EMBL/Genbank acc. nos. EU273946–EU273966, EU274257–EU274267,

Table 1 Species of *Arabidopsis* used for population genetic analysis

A. lyrata ssp. *kamchatica* (2*x* or 4*x*)				
Populations	Country	Location	Sample size	Ploidy
K1	Taiwan	Mt Sheishan	10	All 4*x*
K2	Taiwan	Mt Yushan	10	4*x*: #2, #10, 2*x*: #1, #3, # 4, #5, #8, #9 Unknown: #6, #7
K3	Japan	Shikoku	10	All 4*x*
A. lyrata ssp. *lyrata* (All 2*x*)				
Population	Country	Location	Sample size	
L1	USA	Indiana	10	
L2	USA	Illinois	10	
A. halleri ssp. *gemmifera* (All 2*x*)				
Population	Country	Location	Sample size	
G1	Taiwan	Mt Sheishan	10	
G2	Taiwan	Mt Nahutashan	10	
G3	China	Chilin Province	10	

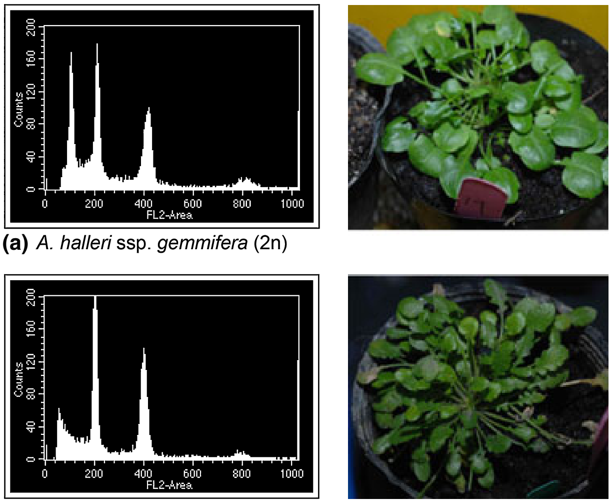
EU274188–EU274201). For a multilocus analysis, a diploid sample of each taxon, confirmed by flow cytometry (details given later), including *A. l. lyrata* (L2-2, Indiana, North America), *A. l. kamchatica* (K2-2, Mt Yushan, Taiwan), *A. h. halleri* (Germany) and *A. h. gemmifera* (G3-6, Chilin Province, China) was chosen (Al-Shehbaz & O'Kane, 2002; Clauss & Koch, 2006; Beck *et al.*, 2007). Young, healthy leaves were collected and dried in silica gel. Leaf tissue was ground to powder in liquid nitrogen and stored in a −70°C freezer. Total genomic DNA was extracted from the powdered tissue following a cetyltrimethylammonium bromide (CTAB) procedure (Murray & Thompson, 1980).

Flow cytometry analysis

To examine the ploidy level of *A. l. kamchatica* samples, flow cytometry was conducted using a protocol modified from Dart *et al.* (2004). In total, 30 individuals of *A. l. kamchatica* from one Japanese population and two Taiwan populations (Table 1) were examined, using diploid *A. h. gemmifera* as a reference. Young leaves (50–100 mg) of *A. l. kamchatica* and *A. h. gemmifera* collected from different locations were chopped using a single-edged razor blade in buffer consisting of 30 mM sodium citrate, 45 mM magnesium chloride, 20 mM 3-Morpholinopropanesulfonic acid (MOPS) buffer and 0.01% v : v Triton X-100 (Galbraith *et al.*, 1983, 1998). Samples were purified by passage through a 30-μm mesh filter and the volume was adjusted to 2 ml. Samples were stained with 1 mg ml^{-1} propidium iodide and incubated with 10 mg ml^{-1} DNase-free RNaseA for 5 min at room temperature to eliminate RNA. Propidium iodide was added to a final concentration of 50 mg ml^{-1} and samples were incubated in the dark for 15 min on ice before analysis. Samples were analysed using a Cytomation MoFLo cytometer (Cytomation, Fort Collins, CO, USA) equipped with a 488 nm laser excitation source operated at an output of 300 mW. Fluorescence emission was collected using a 630/40 bandpass filter. Histograms were processed using SUMMIT software (Cytomation). Multiple peaks in each sample were caused by a high degree of endopolyploidy, a common phenomenon in the Brassicaceae (Barow & Meister, 2003). Our analyses revealed that all *A. h. gemmifera* individuals showed consistent flow cytometry spectra patterns (Fig. 1a). We determined the ploidy level of *A. l. kamchatica* using *A. h. gemmifera* as a reference.

Nucleotide sequencing and sequence analysis

Using the *A. thaliana* genome sequence as a reference, we randomly selected 98 nuclear genes, each spanning at least one coding exon. These loci are distributed over all chromosomes of *A. thaliana*, and physically tightly linked genes were avoided (see the Supporting Information, Table S1). The annotated *A. thaliana* genome was used to design primers to amplify these targeted regions from the selected diploids from each taxon. Sequences of target regions in *A. thaliana* were obtained from the TAIR Database (http://www.arabidopsis.org/index.jsp) (TAIR7; Swarbreck *et al.*, 2008). To avoid super gene families or duplicated genes, all forward and reverse primers were designated to be located at 5′-UTR and 3′-UTR sites. For all 98 loci, orthologous sequences were successfully obtained from the five *Arabidopsis* taxa. Polymerase chain reaction for the amplification of each locus



(a) *A. halleri* ssp. *gemmifera* (2n)

(b) *A. lyrata* ssp. *kamchatica* (2n)

Fig. 1 Flow cytometry histogram of DNA contents of *Arabidopsis halleri* ssp. *gemmifera* (a) and *Arabidopsis lyrata* ssp. *kamchatica* (b).

was performed in a reaction volume of 50 µl using 20 ng template DNA. The PCR cycling scheme consisted of one cycle of denaturation at 95°C for 2 min, 30 cycles of 45 s denaturation at 95°C, 1 min annealing at 51°C and 1 min 45 s extension at 72°C followed by 10 min extension at 72°C. The PCR products were agarose-gel purified with the PCR product purification kit (Viogene, Taipei, Taiwan) and cloned into a pGEM-T easy cloning vector (Promega, Madison, WI, USA). Products of the cycle sequencing reactions were run on an ABI 377XL automated sequencer (Life Technologies, Carlsbad, CA, USA). Cloned PCR products were sequenced using universal primers located on pGEM-T easy vector termination sites. A single clone was chosen for nucleotide sequencing. Nucleotide sequences of *A. thaliana* and its relatives were aligned with the CLUSTAL W program (Thompson *et al.*, 1994) and visually corrected. Kimura's (1980) two-parameter model was used to estimate genetic divergence. Numbers of synonymous and nonsynonymous substitutions were estimated by the method of Li (1993). Neighbor-joining (NJ) trees (Saitou & Nei, 1987) of individual genes and of the combined sequence data (total evidence) were generated using the Kimura two-parameter model (Kimura, 1980) with MEGA 4 (Tamura *et al.*, 2007). Neighbor-joining tree nodes were statistically tested using bootstrapping with 1000 replicates (Felsenstein, 1985).

Coalescence analysis of multilocus data

A multilocus analysis of the 98 nuclear loci was conducted. To test whether inconsistency in gene tree topologies among loci was caused by the retention of ancestral polymorphisms, we derived theoretical probabilities for each type of gene tree under a null model without interspecies gene flow. A case of four lineages in two species (A and B in one species and C and D in the other) is schematically represented in Fig. 2(a). Suppose that AB is the common ancestors of A and B, and CD is the common ancestors of C and D. Let t_1 be the time between the speciation of AB and CD and the divergence of A and B. Similarly, t_2 is the time between the speciation of AB and CD and the divergence of C and D. The two species have diverged at t_3 before the present. During the times t_1 and t_2, the ancestral population sizes are assumed to be constant (N_1 and N_2, respectively). Genes from the same major group, such as lineages A and B, coalesce within t_1 with a probability of $1 - e^{-T_1}$, where $T_1 = t_1/2N_1$ (Takahata & Nei, 1985; Rosenberg & Tao, 2008). Similarly, genes from C and D coalesce within t_2 with a probability of $1 - e^{-T_2}$ ($T_2 = t_2/2N_2$). For genes from A and B, there is a chance to coalesce before the time of speciation (t_3) with a probability of $1 - \left(1 - e^{-T_1}\right)$, and a probability of $1 - \left(1 - e^{-T_2}\right)$ for genes from C and D. In this case, the ancestral population is a single panmictic population and we assume that all possible tree topologies are observed at the same probability. Based on these

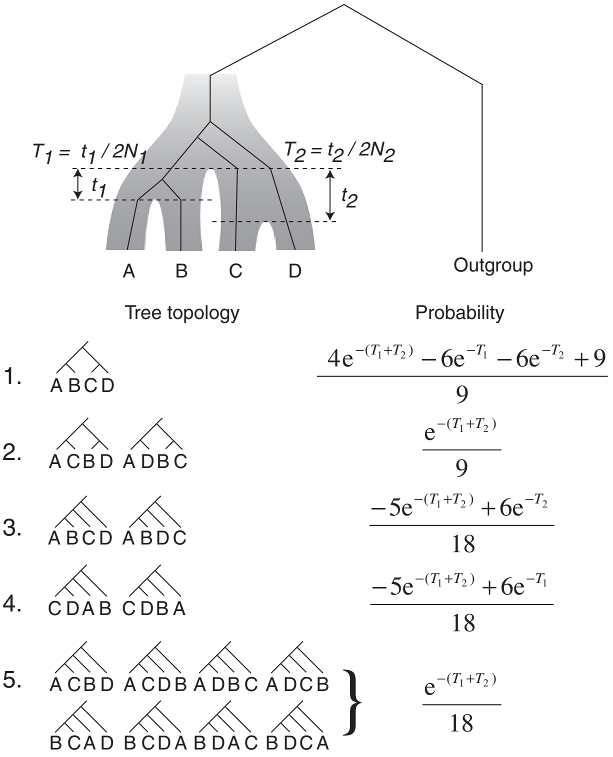

$T_1 = t_1 / 2N_1$
t_1
$T_2 = t_2 / 2N_2$
t_2
A
B
C
D
Outgroup
Tree topology
Probability
1.
ABCD
$\dfrac{4e^{-(T_1+T_2)} - 6e^{-T_1} - 6e^{-T_2} + 9}{9}$
2.
ACBD ADBC
$\dfrac{e^{-(T_1+T_2)}}{9}$
3.
ABCD ABDC
$\dfrac{-5e^{-(T_1+T_2)} + 6e^{-T_2}}{18}$
4.
CDAB CDBA
$\dfrac{-5e^{-(T_1+T_2)} + 6e^{-T_1}}{18}$
5.
ACBD ACDB ADBC ADCB
BCAD BCDA BDAC BDCA
$\dfrac{e^{-(T_1+T_2)}}{18}$

Fig. 2 Theoretical probabilities of gene trees under the null model of no interspecies gene flow. (a) Schematic representation of a gene tree (line) and species tree (background) with four lineages in two species. In this case, two of these (lineages A and B; lineages C and D) are lineages within species and hence closely related to each other. T_1 and T_2 are durations of two successive divergence events, scaled by the ancestral population sizes. (b) Probabilities of tree topologies. For example, the probability of obtaining the gene tree shown in Fig. 1a is $(-5e^{-(T_1+T_2)} + 6hboxe^{-T_2})/18$.

probabilities, we were able to derive the probability for each gene tree topology (Fig. 2b).

We suppose that lineage A is *A. l. lyrata*, B is *A. l. kamchatica*, C is *A. h. halleri* and D is *A. h. gemmifera*. The taxonomic treatment is based on the classification of Al-Shehbaz & O'Kane (2002) and is supported by the majority of the loci we investigated (see the Results section). For the tree topologies shown in Table 3, the observed number of genealogies among 98 loci and theoretical expectations were compared. Six rarely observed patterns and three unobserved patterns were summed into one category. Using the contingency table of observation and expectation, we derived the likelihood of observing data with given e^{-T_1} and e^{-T_2}, assuming that the observed number of gene trees follows a binomial distribution. The procedure is equivalent to the G-test (Sokal & Rohlf, 1994). Although e^{-T_1} and e^{-T_2} are unknown parameters, they range from 0 to 1. Therefore, we could efficiently search the parameter space by numerical iteration. The likelihood values were numerically evaluated by changing e^{-T_1} and e^{-T_2} with an interval of 0.001 ($0 \leq e^{-T_1}, e^{-T_2} \leq 1$).

Estimating ancestral polymorphisms

The MCMCcoal program that implements the Markov Chain Monte Carlo (MCMC) algorithm was used to estimate levels of ancestral polymorphisms and species divergence times. The method using DNA sequence data from multiple loci extracts the information of conflicts among gene trees and of coalescent times to estimate ancestral polymorphisms (Rannala & Yang, 2003). The model assumes no genetic recombination within a locus, free recombination between loci, no gene flow between species, and neutral evolution. The total-evidence tree agreeing with the taxonomic treatment was taken as the species tree.

The MCMCcoal program accommodates parameters of the divergence time ($\tau = t\mu$) and the level of genetic polymorphism ($\theta = 4N_e\mu$) for ancestral species, where t, μ and N_e denote the divergence time, mutation rate, and effective population size, respectively. The divergence time (τ) is measured by the expected number of mutations per site from the ancestral node in the species tree to the present time. The number of synonymous changes per synonymous site (K_s) was used to estimate the parameters.

Estimating patterns of extant polymorphism

We examined levels of genetic polymorphisms and population structure in the *Arabidopsis* taxa by looking at sequence variation at three nuclear genes, *scADH*, *CAUL*, and *Aly9*, all of which have been sequenced for a population study of *A. h. halleri* (Ruggiero *et al.*, 2008). The PCR amplification took place using the primers of Ruggiero *et al.* (2008). The PCR products were inserted into cloning vectors. For each individual, three to five clones were randomly chosen and sequenced. When more than one allele was found in a sample, the allele number was shown at the end of sample name.

Genetic diversities within and between populations were examined in a hierarchical manner. The nucleotide diversity per site at synonymous sites (π) was used to measure the levels of extant polymorphisms (Nei & Li, 1979). For the closely related *Arabidopsis* taxa, samples were collected from hierarchically structured populations. Following the assumptions of Zhou *et al.* (2007), we calculated π_R and π_T for each locus, where π_R is the nucleotide diversity averaged across regional samples and π_T is the nucleotide diversity calculated using all sampled sequences. To measure the level of genetic differentiation, the following definitions are used: $F_{RT} = 1-\pi_R^*/\pi_T^*$, where π_R^* and π_T^* are mean values of π_R and π_T averaged across the three loci, respectively (Hudson *et al.*, 1992).

Estimating rates of gene flow

Rates of gene flow between *Arabidopsis* species were estimated with a coalescent model implemented in the software MIGRATE ver. 2.3 using the three-locus polymorphism data (Beerli & Felsenstein, 2001). MIGRATE can estimate the extent of gene flow in both directions between populations. The software calculates maximum likelihood estimates of the level of polymorphisms θ (defined as $4N_e\mu$, where μ represents mutation rate) and migration rate M (defined as m/μ, where m represents the fraction of migrants), and allows one to detect asymmetries in migration over the species/populations histories (Beerli & Felsenstein, 1999). Because MIGRATE assumes no recombination within a locus, all of the sequence data were trimmed using the IMgc program so as not to contain possible recombined DNA fragments with four-gamete violations (Woerner *et al.*, 2007). The ML mode was used with 10 chains; for each chain, 50 000 genealogies were sampled with a sampling increment of 20 genealogies. We estimated the population migration rate for genes moving into a population per generation ($2N_em$), with the equation $2N_em = (4N_e\mu) \times (m/\mu)/2 = \theta M/2$.

Gene ontology analysis

To access the association between the gene function and pattern of gene introgression, gene ontology annotation of the 98 genes was obtained from the TAIR database (Swarbreck *et al.*, 2008). We chose the gene ontology categories that contained more than three genes. For each gene ontology category, Fisher's exact test was conducted to examine whether the numbers of the congruent and introgressed genes were significantly skewed. A test of multiple comparisons was also performed using the method of Benjamini & Hochberg (1995) with a false discovery rate of 0.05.

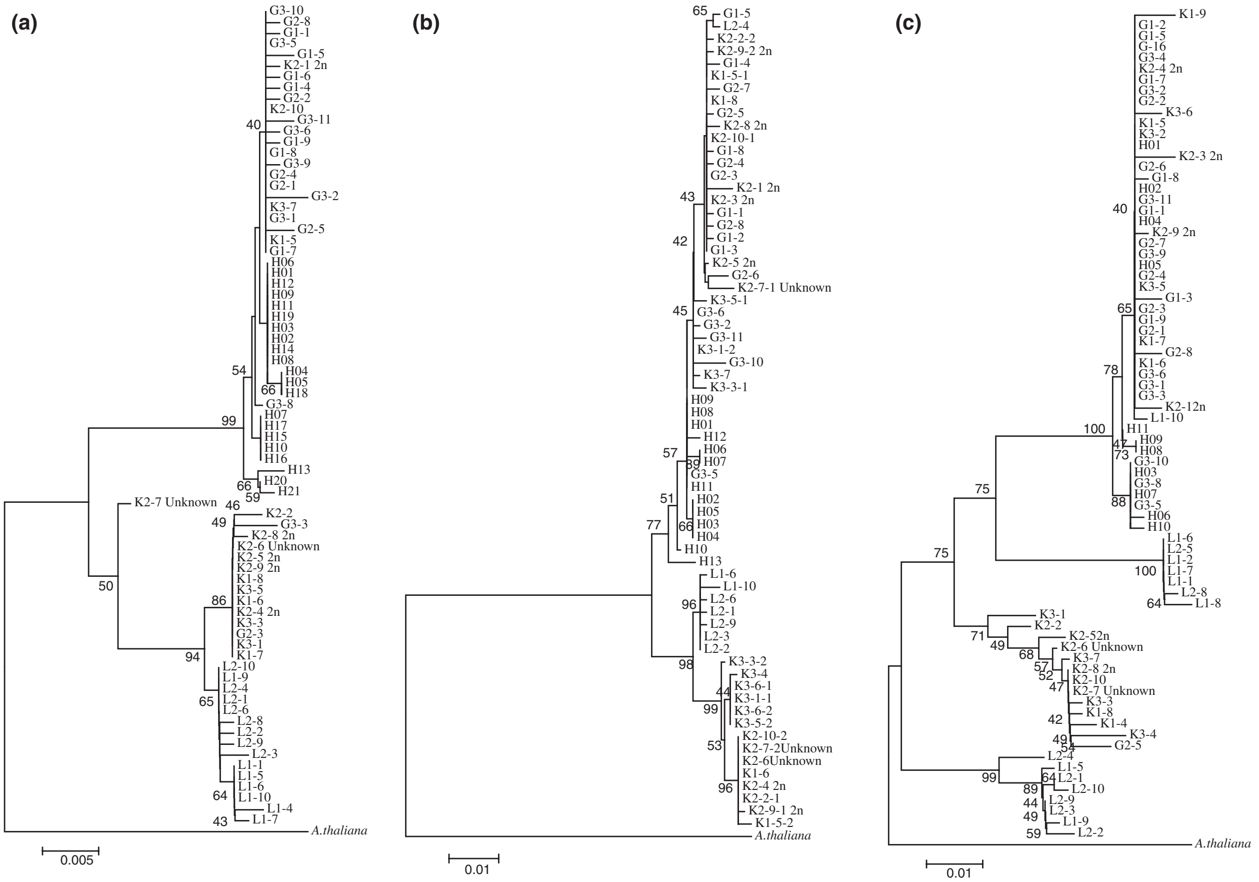


Fig. 3 Neighbor-joining trees of nucleotide sequences of *Aly9* (a), *CAUL* (b), and *scADH* (c) in *A. lyrata* ssp. *kamchatica*, *A. lyrata* ssp. *lyrata*, and *A. halleri* ssp. *gemmifera*. Numbers at nodes indicate bootstrap values.

Results

Ploidy levels and gene genealogies

Ploidy levels of *Arabidopsis* species were determined using flow cytometry. Both diploid and tetraploid individuals of *A. l. kamchatica* were identified using *A. h. gemmifera*, a diploid, as reference (Figs 1 and S1). Diploid *A. l. kamchatica* occurred only in the Yushan population of Taiwan (Table 1) where it was present along with the tetraploid form of the species. Population genetic analysis was conducted based on the genetic variation at *CAUL*, *Aly9* and *scADH* loci. Neighbor-joining trees of the genes are shown in Fig. 3. Monophyly only occurred in *A. l. lyrata* at *Aly9* and *CAUL* loci; while all other taxa were paraphyletic. It is noticeable that almost all *A. l. kamchatica* individuals, both diploids and tetraploids, possessed one or two *halleri*-like alleles. The pattern was essentially same when we removed all singletons from the tree reconstruction (data not shown). The mosaic genetic composition supported hybridity of the diploid *A. l. kamchatica*. Several *A. h. gemmifera* individuals carried *lyrata*-like alleles and some *A. l. lyrata* individuals carried *halleri*-like alleles. For example, at *Aly9* and *scADH* loci, some *A. h. gemmifera* individuals from Taiwan and China were clustered with *A. l. lyrata* (Fig. 3a,c).

Multilocus genealogical analyses

The phylogeny of *A. l. lyrata, A. l. kamchatica, A. h. halleri* and *A. h. gemmifera*, was first reconstructed, rooted at *A. thaliana*, using the concatenated sequences of 98 nuclear loci. Hereafter, we designate the five taxa as L (*lyrata*), K (*kamchatica*), H (*halleri*), G (*gemmifera*) and T (*thaliana*). All samples were diploids. This 'total-evidence' tree displayed a topology agreeing with the phylogenetic treatment of O'Kane & Al-Shehbaz (2003): {[(H, G), (L, K)], T}. Gene trees of 98 loci were then reconstructed separately using all sites. Topologies of these gene trees are summarized in Table 3. After removing four genes that had tree topologies with *A. thaliana* nested within the ingroups, and another four genes displaying star-like phylogenies, the remaining 90 loci were subjected to subsequent analyses. Topological consistency between gene trees and the total-evidence tree occurred only for 28 genes, while most other genes yielded variant phylogenies.Nearly the same number of genes (23 loci) displayed topologies with

Table 2 Pairwise synonymous substitution rate (K_s,below diagonal) and nonsynonymous substitution rate (K_a, above diagonal) values between five *Arabidopsis* taxa

	kamchatica	*lyrata*	*gemmifera*	*halleri*	*thaliana*
kamchatica		0.0076	0.0052	0.0068	0.0171
lyrata	0.0367		0.0081	0.0088	0.0167
gemmifera	0.0298	0.0538		0.0050	0.0160
halleri	0.0453	0.0546	0.0245		0.0168
thaliana	0.1233	0.1234	0.1225	0.1206	

A. l. kamchatica and *A. h. gemmifera* genes coalescing first. Furthermore, the systematic inconsistency across nuclear genes remained even when *A. l. kamchatica* samples were removed from the genealogical analysis. Following the removal of *kamchatica*, out of 98 genealogies of L, H, G and T, 68 (69.4%) were consistent with the total-evidence tree {[(G, H), L], T}, while 15 (15.3%) and 12 (11.4%) displayed topologies of {[(L, G), H], T} and {[(L, H), G], T}, respectively. Another three (3.1%) failed in rooting at *A. thaliana* (Table S3).

Various evolutionary forces can result in a gene tree deviating from the species phylogeny. Lack of a molecular clock owing to branch length heterogeneity is one of them. To evaluate this possibility, we estimated nonsynonymous (K_a) and synonymous (K_s) substitution rates across nuclear genes between *Arabidopsis* taxa (Table 2). As expected, *A. thaliana* was most distant from all other species. Based on the combined data of 98 loci, average K_s values from *thaliana* are 12.3% to *A. l. kamchatica*, *A. l. lyrata* and *A. h. gemmifera*, and 12.1% to *A. h. halleri*. Approximating levels of genetic divergence suggest that *A. thaliana* is equally distant from other *Arabidopsis* species and an appropriate outgroup for estimating divergence. Genomic divergence between *A. lyrata* and *A. halleri* was also estimated. The K_s values were 5.38% between *A. h. gemmifera* and *A. l. lyrata*, 5.46% between *A. h. halleri* and *A. l. lyrata*, 4.53% between *A. h. halleri* and *A. l. kamchatica*, and 2.98% between *A. h. gemmifera* and *A. l. kamchatica*. Genetic divergence at synonymous sites between *A. h. gemmifera* and *A. l. kamchatica* was significantly smaller than that between *A. h. gemmifera* and *A. l. lyrata* ($P = 8.64 \times 10^{-5}$), between *A. h. halleri* and *A. l. lyrata* ($P = 5.96 \times 10^{-6}$), and between *A. h. halleri* and *A. l. kamchatica* ($P = 2.71 \times 10^{-7}$; Steel-Dwass' test). The K_s value between *A. l. lyrata* and *A. l. kamchatica* was 3.67%, which was higher than that between *A. h. gemmifera* and *A. l. kamchatica*.

Low synonymous divergence between *A. l. kamchatica* and *A. h. gemmifera* might result from strong negative selection on synonymous sites and/or low substitution rates in *A. h. gemmifera* and *A. l. kamchatica*. To test if these two lineages evolved at a slower pace than others, a molecular clock between sister lineages of *A. h. gemmifera-halleri* and *A. l. kamchatica-lyrata* was tested with the Tajima's (1993) relative rate test using *A. thaliana* as an outgroup. The null hypothesis of rate constancy was not rejected.

In order to test if the pattern of aberrant gene trees deviates from the isolation model without interspecies gene flow, we derived theoretical probabilities of observing gene tree topologies (see the Materials and Methods section). Fig. 2(a) shows the phylogeny of five hypothetical taxa, in which two of the four taxa are closely related to each other. Previous studies have shown that the probability of obtaining incongruent gene trees depends on the time interval between two successive speciation events scaled by the ancestral population size (T_1 and T_2 in Fig. 2(a); Nei, 1987). The probability for each topology is summarized in Fig. 3(b). The likelihood of the given data was estimated using the theoretical expectations (see the Materials and Methods section). The likelihood surface was highly smooth (Fig. S2). The maximum likelihood value of −17.751 was obtained when $T_1 = 0.49$ and $T_2 = 1.03$. However, even with these parameters, a goodness-of-fit test showed that the statistics were significantly incompatible with the expected numbers of trees under the null hypothesis of no gene introgression ($P < 10^{-22}$; G-test, Table 3). Consistent results were obtained even when we restricted the analysis to trees of high statistical reliability with bootstrap values > 70 ($P < 10^{-22}$; G-test, Table 3). The results imply that the inconsistency between gene trees and species phylogeny is unlikely to be caused by incomplete lineage sorting. Besides, inconsistent gene trees are highly asymmetric towards {[(K, G), H], L} (23 genes) compared with {[(K, G), L], H} (two genes), suggesting directional gene flow from *A. h. gemmifera* to *A. l. kamchatica*, being largely responsible for the systematic inconsistency across genes.

Table 3 Observed numbers of genealogies vs the expected under the model of no gene flow

Tree topology	Probability	Expected[b]	Observed[c]
(KL)(HG)	$\frac{4e^{-(T_1+T_2)}-6e^{-T_1}-6e^{-T_2}+9}{9}$	40.7	28 (15)
{[(KG)H]L}	$\frac{e^{-(T_1+T_2)}}{18}$	1.1	23 (16)
{[(GH)K]L}	$\frac{-5e^{-(T_1+T_2)}+6e^{-T_1}}{18}$	12.9	11 (7)
{[(GL)K]H}	$\frac{e^{-(T_1+T_2)}}{18}$	1.1	7 (4)
(KG)(HL)	$\frac{e^{-(T_1+T_2)}}{9}$	2.2	6 (4)
{[(KL)H]G}	$\frac{-5e^{-(T_1+T_2)}+6e^{-T_2}}{18}$	5.3	5 (3)
Other patterns[a]	$\frac{-e^{-(T_1+T_2)}+3e^{-T_1}+3e^{-T_2}}{9}$	26.8	10 (5)

[a] Rare observed patterns include 3 {[(KL)G]H}, 2 {[(GH)L]K}, 2 {[(KG)L]H}, and one of each {[(LH)K]G}, {[(KH)G]L}, (KH)(LG).
[b] Expectation under the model of no interspecies gene flow (see the Materials and Methods section).
[c] Numbers of the observed gene tree with high bootstrap values (> 70) are shown in the parentheses.

Association of gene functions and phylogenetic inconsistency

Whether different functional gene groups were involved in the systematic inconsistency was tested based on the gene ontology annotation obtained from the TAIR database (Wang & Zhang, 2009). Nuclear genes were classified into congruent genes that recovered the taxonomic treatment and introgressed genes that supported the clustering of *A. l. kamchatica* and *A. h. gemmifera*. Statistical analyses revealed that three functional categories were over-represented in the congruent genes (i.e. under-represented in the incongruent genes): nucleus localization (GO:0005634; $P = 0.036$), DNA-dependent regulation of transcription (GO:0006355; $P = 0.045$), and DNA-binding function (GO:0003677; $P = 0.045$; Fisher's exact test) (Table S2). These genes tended to recover the taxonomic treatment and, therefore, are unlikely to have migrated freely between species. Interestingly, all these functional groups were related to DNA-binding transcription factors, indicating functional importance of these factors to species uniqueness. After correction for multiple testing, however, only nucleus localization remained significant. The list of candidate genes for causing species uniqueness should, therefore, be treated with caution.

Interspecific gene flow between *A. lyrata* and *A. halleri*

Interspecific gene flow usually results in a mixed genomic composition in hybrids, but often may have low impact on parental species genomes. In other words, it may be very difficult to detect introgression in so-called 'good species'. Here, the genomic composition analysis revealed that not only tetraploid *A. l. kamchatica*, but also the diploid form of this species was a hybrid between *A. l. lyrata* and *A. h. gemmifera*. Whether a significant amount of gene flow was detected between the putative parental species, *A. l. lyrata* and *A. h. gemmifera*, was further examined. Population genetic analyses using coalescence-based methodology implemented in MIGRATE revealed a notable amount of gene flow between *A. l. lyrata* and *A. h. gemmifera*, which have diverged for some 2 Myr. The migration rates ($2N_em$) were estimated as 0.24 from *A. l. lyrata* to *A. h. gemmifera*, and 0.31 from *A. h. gemmifera* to *A. l. lyrata* (Table S4).

In the multilocus analysis, levels of ancestral polymorphism were also estimated in all extant taxa. Divergence time estimates (τ) were measured with the expected number of mutations per site from the ancestral node in the species tree to the present time. Using all taxa, the divergence time estimates of τ_{KL}, τ_{HG} and τ_{KLHG} were obtained (Fig. 4). Similarly, polymorphism parameters, π_{KL}, π_{HG} and π_{KLHG} for a set of three ancestral species were also estimated.

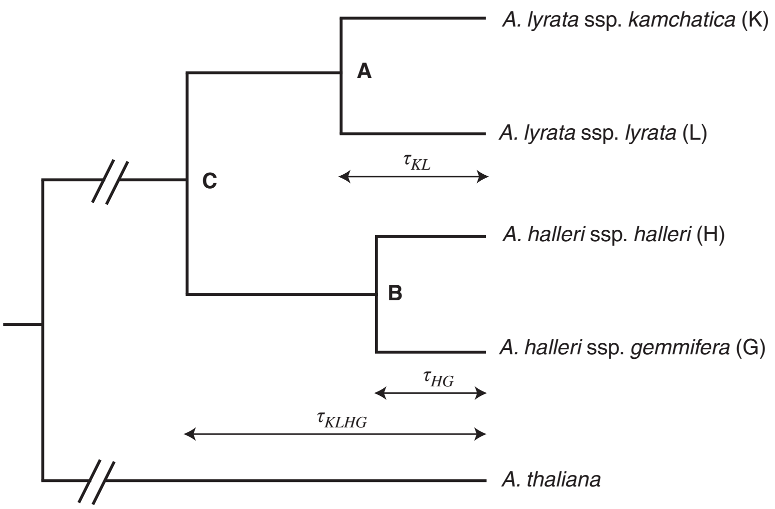
Population data for three nuclear loci were used to estimate the level of extant polymorphisms (Table 4). As wild *Arabidopsis* species have fragmented distributions, π at two levels, i.e. nucleotide diversities of region (π_{R}) and the entire species (π_{T}), was estimated based on variation at synonymous and noncoding sites. For all four taxa, the level of diversities decreased from the entire species to regions (Table 4). For example, the mean level of entire species diversity of *A. h. gemmifera* ($\pi_{\mathrm{T}} = 0.0185$) was 41% higher than that of the regional level ($\pi_{\mathrm{R}} = 0.0131$). The population structure was represented by F_{RT}, one of the *F*-indices, as shown in Table 4. Apparently, in all four *Arabidopsis* taxa, there existed a hierarchical population structure (Bakker *et al.*, 2006; Ross-Ibarra *et al.*, 2008). Each species could be divided into geographical populations that were isolated from each other, a pattern also observed in other plant species in East Asia (Dodd *et al.*, 2002; Chiang & Schaal, 2006).

As the model used in MCMCcoal assumes strict allopatry, gene flow between *A. h. gemmifera* and *A. l. kamchatica* would inflate the estimated ancestral polymorphisms between *A. halleri* and *A. lyrata*. The ancestral polymorphism for the most recent common ancestors of *A. lyrata* and *A. halleri* (π_{KLHG}) was estimated to be 0.0466 (Fig. 4). The estimated value for ancestral polymorphisms considerably exceeded that for extant polymorphisms. To test whether this inflation is caused solely by interspecific gene flow between *A. h. gemmifera* and *A. l. kamchatica*, we repeated the analysis by excluding these two taxa. As expected, the ancestral polymorphisms decreased from 0.0466 to 0.0316 (Fig. 4), but still exceeded the value for observed extant polymorphisms, suggesting that gene flow between *A. halleri* and *A. lyrata* continued after their speciation.

Discussion

Phylogenetic inconsistency likely caused by introgression after speciation

Species phylogenies present bifurcate branching patterns and progenitor-descendant relationships of a group of taxa. Among species evolving via cladogenesis, genetic divergence reflects these branching processes from common ancestors. In the present study, gene genealogies were reconstructed individually based on 98 nuclear genes in *A. h. halleri*, *A. h. gemmifera*, *A. l. lyrata*, and the diploid form of *A. l. kamchatica* using *A. thaliana* as the outgroup. Correspondence between phylogenetic and genetic distances occurred in the comparisons between *A. thaliana* and its relatives with the greatest synonymous divergence (*c.* 12%) (Wright *et al.*, 2002; Barrier *et al.*, 2003; Ramos-Onsins *et al.*, 2004; Clauss & Koch, 2006), and between *A. halleri* and *A. lyrata* with 4–5% synonymous divergence. Lowest genetic divergence was unexpectedly detected between *A. l.*


A. lyrata ssp. *kamchatica* (K)
A
A. lyrata ssp. *lyrata* (L)
C
τ_{KL}
A. halleri ssp. *halleri* (H)
B
A. halleri ssp. *gemmifera* (G)
τ_{HG}
τ_{KLHG}
A. thaliana

	Ancestral polymorphisms (95% C.I.)	Speciation time (95% C.I.)
Using K, L, H, and G		
Node **A**	$\pi_{KL} =$ 0.00768 (0.00363, 0.01423)	$\tau_{KL} =$ 0.00657 (0.00551, 0.00805)
Node **B**	$\pi_{HG} =$ 0.00338 (0.00216, 0.01376)	$\tau_{HG} =$ 0.00648 (0.00542, 0.00784)
Node **C**	$\pi_{KLHG} =$ 0.04658 (0.03774, 0.05449)	$\tau_{KLHG} =$ 0.00684 (0.00547, 0.00834)
Using only L and H		
Node **C**	$\pi_{KLHG} =$ 0.03155 (0.02426, 0.04417)	$\tau_{KLHG} =$ 0.02231 (0.01729, 0.03054)

Fig. 4 Patterns of ancestral polymorphisms estimated for four taxa. Ancestral species in the phylogeny are denoted as nodes A, B, and C. Ancestral polymorphisms (π) and time after speciation (τ) under the null model of no interspecies gene flow are indicated.

Table 4 Extant polymorphisms of the *Arabidopsis* species

Species	π_R	π_T	F_{RT}	
A. lyrata ssp. *kamchatica*	0.0226	0.0345	0.344	This study
A. lyrata ssp. *lyrata*	0.0124	0.0207	0.401	This study
A. halleri ssp. *gemmifera*	0.0131	0.0185	0.292	This study
A. halleri ssp. *halleri*	0.0137	0.0216	0.365	Ruggiero *et al.* (2008)

kamchatica and *A. h. gemmifera*. Once the hypothesis of rate heterogeneity between lineages was rejected, other evolutionary forces have to be invoked to explain the unexpected pattern of low between-species (*A. l. kamchatica* and *A. h. gemmifera*) vs high within-species divergence (*A. l. kamchatica* and *A. l. lyrata*).

High phylogenetic inconsistency was detected across the 98 randomly selected loci examined, with genealogies for only 28 loci matching the phylogeny resolved in the total-evidence (species) tree (Table 3). Deviations of gene genealogies from the species phylogeny can be caused by: a wrong genealogy obtained owing to statistical errors caused by insufficient data and/or abundant recurrent mutations; a taxonomic treatment based on nonmolecular data being misjudged; lineage sorting owing to ancestral polymorphisms segregating at the time of speciation; and occurrence of hybridization after speciation leading to genealogy distortion. A series of genetic analyses suggest that the last two processes are the more likely causes of the deviations recorded.

Based on the probabilities for tree topologies under a complete isolation model, the chances of obtaining incongruent gene trees were determined by the speciation time scaled by the ancestral population size ($T = t/2N_e$, see Fig. 3). Here, at the entire parameter space of T, the observed numbers of incongruent gene trees were significantly larger than expected, suggesting that such high systematic inconsistency in *Arabidopsis* cannot be explained solely by the segregation of ancestral polymorphisms

(Table 3). In contrast, this result agrees with a scenario of hybridization and/or gene introgression between sympatric *Arabidopsis* species.

Mosaic genome of diploid *A. lyrata* ssp. *kamchatica*

We conducted a multilocus analysis to recover the evolution of what is considered to be a diploid form of *A. l. kamchatica*. Reticulate relationships between *Arabidopsis* species were revealed based on the reconstructed gene genealogies and it was evident from this that the genome of diploid *A. l. kamchatica* consists of both *lyrata*-like and *halleri*-like alleles. Of the nuclear loci selected, 28 genes supported the clustering of *A. l. lyrata* and diploid *A. l. kamchatica*, while another 23 genes revealed close affinity between *A. h. gemmifera* and *A. l. kamchatica*. These results indicate approximate contributions of the *lyrata* and *halleri* genomes to the putative hybrid. Population data also showed that both tetraploid and diploid forms of *A. l. kamchatica* possessed a mixed genome of *A. lyrata* and *A. halleri*, providing further evidence of their hybrid status at each ploidy level. The existence of a high proportion of *halleri*-like alleles in diploid *A. l. kamchatica*, but not vice versa, also indicated that *A. h. gemmifera* acted as a genome donor to the putative hybrid. Nonetheless, the occurrence at low frequency of *lyrata*-like alleles in the *A. h. gemmifera* genome, according to the population data, indicates the possibility of some backcrossing of *A. l. kamchatica* with its parental species (Beck *et al.*, 2007).

As *A. l. kamchatica* is the only form of *A. lyrata* that currently exists in Taiwan, it is difficult to determine whether the observed diploid *A. l. kamchatica* descended from an actual hybrid lineage or a parental lineage. If the latter is the case, the parental species might have been replaced by the hybrid lineage; while in the case of allotetraploid *A. l. kamchatica*, both parental species are likely to be extinct in Taiwan. Three possible evolutionary scenarios might explain the origins of diploid and tetraploid forms of *A. l. kamchatica* in Taiwan. First, the formation of allotetraploid *A. l. kamchatica* was followed by backcrossing with parental *A. lyrata*, leading to the formation of diploid *A. l. kamchatica* via a 'triploid bridge'. In contrast to general thinking, triploids are not always completely sterile and often can exchange genes across ploidy levels through the formation of 'balanced' haploid and diploid gametes (Henry *et al.*, 2007; Chapman & Abbott, 2010). Second, introgression from *A. h. gemmifera* to ancestral *A. lyrata* resulted in the origin of a stable diploid hybrid segregant, which subsequently gave rise through polyploidization to the allotetraploid form of *A. l. kamchatica*. Third, diploid and tetraploid *A. l. kamchatica* had independent origins, hence were genetically differentiated from each other. Although there is good evidence that tetraploid *A. l. kamchatica* is a hybrid that has originated via multiple polyploidization events (Shimizu-Inatsugi *et al.*, 2009; Schmickl *et al.*, 2010), the evolutionary interactions between diploid and tetraploid *A. l. kamchatica* remain to be explored. Population samples from a wide geographical range will be required to unravel their reticulate relationships.

Between genetically interconnected species, natural selection may determine the fates of genes and genomes of hybrids (Rieseberg *et al.*, 2003). In the present study, the ontology analysis revealed that genes classified as DNA-binding transcription factors were overrepresented among the genes showing genealogies that matched the species phylogeny, i.e. mostly from *A. lyrata*. Although statistical support was marginal, this high level of congruence between trees derived from genes encoding transcription factors implies that such genes have greater biological importance and evolutionary significance. These genes may exhibit species-specific mutations and/or encode species-specific transcriptional activities.

Gene flow between long diverged *Arabidopsis* species

To further investigate the possibility of gene flow between long-diverged *Arabidopsis* species, we examined if the diploid and tetraploid hybrid forms of *A. l. kamchatica* may have acted as effective barriers to genetic exchanges between *A. halleri* and *A. lyrata*, thus enforcing their reproductive isolation despite secondary contact. This was done by looking at the levels of ancestral polymorphism of *A. lyrata* and *A. halleri*. In the strictest case of an allopatric speciation model, an ancestral population splits into two geographically isolated regional populations and complete reproductive isolation is established during the geographically isolated phase (Mayr, 1954). At the moment of split, each of the regional populations tends to resemble the ancestral population at the level of genetic polymorphism. These regional populations eventually become reproductively isolated and evolve into new species. The expected nucleotide diversity pattern can be summarized as $\pi_A = \pi_R{}^a \sim \pi_R < \pi_T$ (cf. Zhou *et al.*, 2007), where the superscript 'a' denotes the level of polymorphism at the time of geographical isolation. However, if the strict allopatry model does not apply, the variance of gene divergence will be inflated by gene migration after the initial isolation. Because the level of ancestral polymorphisms (π_A) is estimated from the divergence variance, the level of ancestral polymorphisms is likely to be overestimated in the presence of gene flow (Osada & Wu, 2005).

Our results indicate an inflated estimate of ancestral polymorphisms between *A. lyrata* and *A. halleri,* which is likely attributable to interspecific gene flow. When *A. h. gemmifera* and *A. l. kamchatica* (i.e. taxa that caused systematic inconsistency owing to rampant gene introgression) were excluded from the analysis, the estimated π_A

substantially decreased, although it was still higher than π_R. A possible explanation for this pattern is that the common ancestor of *A. halleri* and *A. lyrata* has not been under complete reproductive isolation since their initial divergence (Zhou *et al.*, 2007). In addition, nonzero gene flow was also detected between *A. l. lyrata* and *A. h. gemmifera* based on a population genetic analysis with MIGRATE. The gene trees in Fig. 3 suggest that there have been recent gene introgression events to *A. lyrata* and *A. halleri* although the amount was small. Investigating the pattern of speciation between sister taxa using genome-wide data shows that interspecific gene flow is much more prevalent than expected.

The occurrence of gene flow between species after speciation calls into question the identity of so-called 'good species'. Phylogenetic analyses that did not include the hybrid taxon *A. l. kamchatica* were still highly inconsistent across loci. Whereas at least 37.9% inconsistency was evident with the diploid hybrid included, 27.7% inconsistency occurred when it was excluded. In complexes of the type investigated here, the traditional view of species phylogeny may no longer be effective, as both branching and reticulate evolution work in concert in determining species diversification. The phylogeny of genomes is likely to be a mixture of the genealogical information of genes (or recombination units), and therefore cannot be determined by a simple, reductive view. Nevertheless, the fact that species tend to become diverged and differentiated indicates that natural selection counteracts the effect of gene flow.

Conclusions

In this study, we examined the effects of natural hybridization on genome evolution in *Arabidopsis* by looking at multilocus divergence among species. The sequence data from nuclear loci provided sufficient power to recover the genomic history among sister species. We found that what may be considered as a diploid form of *A. l. kamchatica* was composed of the two different genomes derived from *A. lyrata* and *A. halleri*. The evolutionary relationships of tetraploid and diploid forms of *A. lyrata kamchatica* remain to be explored via population approaches. Different lines of evidence supported the occurrence of gene flow between *A. lyrata* and *A. halleri*. Apparently, a period of divergence of 2–3 Myr may not be sufficient for attaining complete reproductive isolation in *Arabidopsis*. Rejecting the null model of no interspecific gene flow after speciation is a first step for investigating the roles of gene introgression in plant genome evolution.

Acknowledgements

We thank Thomas Mitchell-Olds for the DNA of *A. halleri* ssp. *halleri* and Hiroshi Akashi and Richard Abbott for editing the manuscript. We wish to acknowledge valuable comments from five anonymous reviewers. This work was supported by grants from the National Science Council, Taiwan.

References

Al-Shehbaz IA, O'Kane SL. 2002. Taxonomy and phylogeny of *Arabidopsis* (Brassicaceae). In: Somerville CR, Meyerowitz EM, eds. *The Arabidopsis book*. Rockville, MD: American Society of Plant Biologists, 1–22.

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

Baack EJ, Rieseberg LH. 2007. A genomic view of introgression and hybrid speciation. *Current Opinion in Genetics and Development* **17**: 513–518.

Bakker EG, Stahl EA, Toomajian C, Nordborg M, Kreitman M, Bergelson J. 2006. Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Molecular Ecology* **15**: 1405–1418.

Barow M, Meister A. 2003. Endopolyploidy in seed plants is differently correlated to systematics, organ, life strategy and genome size. *Plant, Cell & Environment* **26**: 571–584.

Barrier M, Bustamante CD, Yu J, Purugganan MD. 2003. Selection on rapidly evolving proteins in the *Arabidopsis* genome. *Genetics* **163**: 723–733.

Beck JB, Al-Shehbaz IA, O'Kane SL, Schaal BA. 2007. Further insights into the phylogeny of *Arabidopsis* (Brassicaceae) from nuclear Atmyb2 flanking sequence. *Molecular Phylogenetics and Evolution* **42**: 122–130.

Beerli P, Felsenstein J. 1999. Maximum likelihood estimation of migration rates and population numbers of two populations using a coalescent approach. *Genetics* **152**: 763–773.

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences, USA* **98**: 4563–4568.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **42**: 289–300.

Bomblies K, Weigel D. 2007. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Review Genetics* **8**: 382–393.

Chapman MA, Abbott RJ. 2010. Introgression of fitness genes across a ploidy barrier. *New Phytologist* **186**: 63–71.

Chiang YC, Hung KH, Moore SJ, Ge XJ, Huang S, Hsu TW, Schaal BA, Chiang TY. 2009. Paraphyly of organelle DNAs in *Cycas* Sect. Asiorientales due to ancient ancestral polymorphisms. *BMC Evolution Biology* **9**: 161.

Chiang TY, Schaal BA. 2006. Phylogeography of plants in Taiwan and the Ryukyu Archipelago. *Taxon* **55**: 31–41.

Clauss MJ, Koch MA. 2006. Poorly known relatives of *Arabidopsis thaliana*. *Trends in Plant Science* **11**: 449–459.

Dart S, Kron P, Mable BK. 2004. Characterizing polyploidy in *Arabidopsis lyrata* using chromosome counts and flow cytometry. *Canadian journal of Botany* **82**: 185–197.

Dodd RS, Afzal-Rafii Z, Kashani N, Budrick J. 2002. Land barriers and open oceans: effects on gene diversity and population structure in *Avicennia germinans* L. (Avicenniaceae). *Molecular Ecology* **11**: 1327–1338.

Elven R, Murray DF, Razzhivin VY, Yurtsev BA. 2007. *Panarcticflora checklist*. Oslo: National Centre for Biosystematics, University of Oslo.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Galbraith DW, Dolezel J, Lambert G, Macas J. 1998. DNA and ploidy analyses in higher plants. In: Robinson JP ed. *Current protocols in cytometry*. New York, NY, USA: Wiley-Blackwell, 7.6.1–7.622.

Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E. 1983. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* **220**: 1049–1051.

Hall AE, Fiebig A, Preuss D. 2002. Beyond the *Arabidopsis* genome: opportunities for comparative genomics. *Plant Physiology* **129**: 1439–1447.

Hayata B. 1911. Materials for a flora of Formosa. *Journal of the College of Science Imperial University, Tokyo* **30**: 1–471.

Hegi G. 1986. *Illustrierte Flora von Mittel-Europa. Pteridophyta und Spermatophyta. Band 4 Angiosperma Dicotyledones 2, Teil 1*. Munich, Germany: Carl Hanser, 230–301.

Henry IM, Dilkes BP, Comai L. 2007. Genetic basis for dosage sensitivity in *Arabidopsis thaliana*. *PLoS Genetics* **3**: e70.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.

Inoue K. 1971. *Arabis lyrata* ssp. *kamchatica* (Fisch. *ex* DC.) Hulten in Formosana. *Journal of Japanese Botany* **46**: 32.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**: 111–120.

King MG, Roalson EH. 2009. Discordance between phylogenetics and coalescent-based divergence modelling: exploring phylogeographic patterns of speciation in the *Carex macrocephala* species complex. *Molecular Ecology* **18**: 468–482.

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J, Hey J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**: 1913–1931.

Koch MA, Matschinger M. 2007. Evolution and genetic differentiation among relatives of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **104**: 6272–6277.

Lawton-Rauh A, Robichaux RH, Purugganan MD. 2007. Diversity and divergence patterns in regulatory genes suggest differential gene flow in recently derived species of the Hawaiian silversword alliance adaptive radiation (Asteraceae). *Molecular Ecology* **16**: 3995–4013.

Lexer C, Kremer A, Petit RJ. 2006. Shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. *Molecular Ecology* **15**: 2007–2012.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**: 96–99.

Mahelka V, Fehrer J, Krahulec F, Jarolímová V. 2007. Recent natural hybridization between two allopolyploid wheatgrasses (*Elytrigia*, Poaceae): ecological and evolutionary implications. *Annals of Botany* **105**: 249–260.

Mallet J. 2007. Hybrid speciation. *Nature* **446**: 279–283.

Mayr E. 1954. *Animal species and evolution*. Cambridge, MA, USA: Belknap Press.

Mitchell-Olds T. 2001. *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends in Ecology and Evolution* **16**: 693–700.

Mitchell-Olds T, Clauss MJ. 2002. Plant evolutionary genomics. *Current Opinion in Plant Biology* **5**: 74–79.

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

Muir G, Schlötterer C. 2005. Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.). *Molecular Ecology* **14**: 549–561.

Murray MG, Thompson WF. 1980. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research* **8**: 4321–4325.

Nei M. 1987. *Molecular evolutionary genetics*. New York, NY, USA: Columbia University Press.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences, USA* **76**: 5269–5273.

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

O'Kane SL, Al-Shehbaz IA. 2003. Phylogenetic position and generic limits of *Arabidopsis* (Brassicaceae) based on sequences of nuclear ribosomal DNA. *Annals of the Missouri Botanical Garden* **90**: 603–612.

Osada N, Wu CI. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics* **169**: 259–264.

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aguade M. 2004. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. *Genetics* **166**: 373–388.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**: 1645–1656.

Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* **301**: 1211–1216.

Rosenberg NA, Tao R. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. *Systematic Biology* **57**: 131–140.

Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D, Gaut BS. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE* **3**: e2411.

Ruggiero MV, Jacquemin B, Castric V, Vekemans X. 2008. Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in *Arabidopsis halleri*. *Genetic Research* **90**: 37–46.

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

Savolainen O, Langley CH, Lazzaro BP, Fréville H. 2000. Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Molecular Biology and Evolution* **17**: 645–655.

Schmickl R, Jorgensen MH, Brysting AK, Koch MA. 2010. The evolutionary history of the *Arabidopsis lyrata* complex: a hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evolution Biology* **10**: 98.

Shimizu KK, Fujii S, Marhold K, Watanabe K, Kudoh H. 2005. *Arabidopsis kamchatica* (Fisch. ex DC.) K. Shimizu & Kudoh and *A. kamchatica* subsp. *kawasakiana* (Makino) K. Shimizu & Kudoh, new combinations. *Acta Phytotax Geobot* **56**: 163–172.

Shimizu-Inatsugi R, Lihová J, Iwanaga H, Kudoh H, Marhold K, Savolainen O, Watanabe K, Yakubov VV, Shimizu KK. 2009. The allopolyploid *Arabidopsis kamchatica* originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. *Molecular Ecology* **18**: 4024–4048.

Sokal RR, Rohlf FJ. 1994. *Biometry: the principles and practice of statistics in biological research*, 3rd edn. New York, NY, USA: Freeman.

Städler T, Roselius K, Stephan W. 2005. Genealogical footprints of speciation processes in wild tomatoes: demography and evidence for historical gene flow. *Evolution* **59**: 1268–1279.

Strasburg JL, Rieseberg LH. 2008. Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris* – large effective population sizes and rates of long-term gene flow. *Evolution* **62**: 1936–1950.

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L *et al.* 2008. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research* **36**: D1009–D1014.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**: 599–607.

Takahata N, Nei M. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**: 325–344.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**: 1596–1599.

Tateishi N, Ozaki Y, Okubo H. 2007. Occurrence of ploidy variation in *Camellia×vernalis*. *Journal of the Faculty of Agriculture, Kyushu University* **52**: 11–15.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673–4680.

Tiffin P, Gaut BS. 2001. Sequence diversity in the tetraploid *Zea perennis* and the closely related diploid *Z. diploperennis*: insights from four nuclear loci. *Genetics* **158**: 401–412.

Trenel P, Hansen MM, Normand S, Borchsenius F. 2008. Landscape genetics, historical isolation and cross-Andean gene flow in the wax palm, *Ceroxylon echinulatum* (Arecaceae). *Molecular Ecology* **17**: 3528–3540.

Wang Z, Zhang J. 2009. Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genetics* **5**: e1000329.

Wendel JF. 2000. Genome evolution in polyploids. *Plant Molecular Biology* **42**: 225–249.

Woerner AE, Cox MP, Hammer MF. 2007. Recombination-filtered genomic datasets by information maximization. *Bioinformatics* **23**: 1851–1853.

Wright SI, Lauga B, Charlesworth D. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Molecular Biology and Evolution* **19**: 1407–1420.

Yatabe Y, Kane NC, Scotti-Saintagne C, Rieseberg LH. 2007. Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics* **175**: 1883–1893.

Zhou R, Zeng K, Wu W, Chen X, Yang Z, Shi S, Wu CI. 2007. Population genetics of speciation in nonmodel organisms: I. Ancestral polymorphism in mangroves. *Molecular Biology and Evolution* **24**: 2746–2754.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Different flow cytometry spectra patterns of *Arabidopsis lyata* ssp. *kamchatica.*

Fig. S2 Likelihood surface of the congruence test.

Table S1 The genealogical and genome information of 98 genes

Table S2 Classification of congruent and introgressed genes based on the ontology category

Table S3 Observed numbers of genealogies with *Arabidopsis lyrata* ssp. *kamchatica* excluded from phylogenetic analyses

Table S4 Interspecies gene flow estimated as $2N_em$ between *Arabidopsis lyrata* ssp. *lyrata* and *Arabidopsis halleri* ssp. *gemmifera*

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.