RESEARCH ARTICLE

Unique nucleotide polymorphism of ankyrin gene cluster in *Arabidopsis*

JIANCHANG DU[1]*, XINGNA WANG[1,2], MINGSHENG ZHANG[1], DACHENG TIAN[1] and YONG-HUA YANG[1]*

[1] *State Key Laboratory of Pharmaceutical Biotechnology, Department of Biology, Nanjing University, Nanjing 210093, China*

[2] *Center for Drug Discovery and Design, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

Abstract

The ankyrin (*ANK*) gene cluster is a part of a multigene family encoding ANK transmembrane proteins in *Arabidopsis thaliana*, and plays an important role in protein–protein interactions and in signal pathways. In contrast to other regions of a genome, the *ANK* gene cluster exhibits an extremely high level of DNA polymorphism in an ~5-kb region, without apparent decay. Phylogenetic analysis detects two clear, deeply differentiated haplotypes (dimorphism). The divergence between haplotypes of accession Col-0 and Ler-0 (Hap-C and Hap-L) is estimated to be 10.7%, approximately equal to the 10.5% average divergence between *A. thaliana* and *A. lyrata*. Sequence comparisons for the *ANK* gene cluster homologues in Col-0 indicate that the members evolve independently, and that the similarity among paralogues is lower than between alleles. Very little intralocus recombination or gene conversion is detected in *ANK* regions. All these characteristics of the *ANK* gene cluster are consistent with a tandem gene duplication and birth-and-death process. The possible mechanisms for and implications of this elevated nucleotide variation are also discussed, including the suggestion of balancing selection.

[Du J., Wang X., Zhang M., Tian D. and Yang Y.-H. 2007 Unique nucleotide polymorphism of ankyrin gene cluster in *Arabidopsis*. *J. Genet.* **86**, 27–35]

Introduction

Ankyrin (ANK) repeats are commonly occurring protein motifs present in prokaryotes, eukaryotes and some viruses (Sedgwick and Smerdon 1999), and play a wide variety of roles in protein–protein interactions and in signal pathways (Li *et al.* 2005). Becerra *et al.* (2004) identified 105 ANK proteins classified in 16 groups of structurally similar proteins in *Arabidopsis*. Among them, ankyrin-transmembrane proteins are the most abundant group encoded by 40 genes, seven of which are linked and clustered tightly (Becerra *et al.* 2004). It is believed that this type of architecture of clustered gene order is nonrandomly distributed in eukaryote genomes (Lercher *et al.* 2002; Hurst *et al.* 2004).

Considerable data on the evolution of human major histocompatibility complex (MHC) and plant resistance (R) genes has been accumulated in previous studies. Resistance gene clusters appear to evolve more rapidly than other regions of the genome (Ronald 1998; Richter and Ronald 2000). Nevertheless, globin genes (Martin *et al.* 1996), homeobox genes (Krumlauf 1992), myosin heavy chain genes (Weiss *et al.* 1999) and T-cell receptor genes (Koop *et al.* 1994), which are arranged in tightly linked clusters, are highly conserved between related species. At least two models have been proposed to interpret the formation and maintenance of clustered gene families, of which one is concerted evolution (Richter *et al.* 1995; Parniske *et al.* 1997; Song *et al.* 1997; Wang *et al.* 1999). In this hypothesis, all member genes of a family are assumed to evolve as a unit, exchanging genetic information frequently, and are therefore homogenized in the absence of something like balancing selection (for example as a consequence of subfunctionalization of paralogues). An alternative hypothesis is the birth-and-death model. Birth in a birth-and-death process relates to the origin of paralogues, presumably by unequal crossing-over; death relates to the loss of paralogues, either by deletion or by degeneration.

*For correspondence. E-mail: changzi75@hotmail.com; yangyh@nju.edu.cn.

Keywords. ankyrin gene cluster; divergence; evolution; gene duplication; *Arabidopsis*.

This model suggests that members of a family evolve independently, and orthologues (or alleles) are predicted to be more similar than paralogues, especially when gene duplication occurred shortly after speciation (or population differentiation). Although the two models are conceptually different, they may not be distinguishable when sequence differences are small, or the rate of concerted evolution is very slow (Nei *et al.* 2000). However, recent investigations have shown that most non-rRNA genes, including highly conserved histone or ubiquitin genes, are subject to the birth-and-death process (Nei *et al.* 1997, 2000; Michelmore and Meyers 1998; Gu and Nei 1999; Piontkivska *et al.* 2002; Nei and Rooney 2005).

Compared to both the MHC and R genes, the genetic diversity and evolution of the *ANK* gene cluster has not been so extensively studied. Showing the highest degree of polymorphism among 876 randomly sequenced fragments, interspersed at an incidence, on average, of one every 50 kb in genomes of 96 *Arabidopsis* accessions, the estimate of average pairwise differences per site for *At4g03470* (*ANK4*, one member of this *ANK* gene cluster) is 0.0741 (Nordborg *et al.* 2005). The divergence, D_{xy}, between two haplotypes of Col-0 and Ler-0 accessions (Hap-C and Hap-L) per site is estimated to be 14.0%, even higher than the 10.5% average divergence between *A. thaliana* and *A. lyrata* (Stahl *et al.* 1999). Is this elevated nucleotide diversity specific to this gene member or found throughout the *ANK* gene cluster? What are the evolutionary forces shaping and maintaining this extremely high level of DNA diversity? How are the architectures of this type of a clustered gene family formed? To address these questions, we examined seven *ANK* clustered genes (*ANK1*, *ANK2*, *ANK3*, *ANK4*, *ANK5*, *ANK6*, *ANK7*), and sequenced corresponding regions of *ANK4* and *ANK5* in Ler-0 accession, based on the published Col-0 sequence. We also detected one 1.5-kb indel (insertion/deletion) polymorphic site from 14 *Arabidopsis* accessions, spanning both coding and noncoding regions of *At4g03480* (*ANK5*), and sequenced its flanking regions. Here we are especially interested in the factors (e.g. natural selection, genetic/genomic or demographic factors) structuring the pattern of polymorphism at this *ANK* gene cluster, as well as in whether the concerted evolution model or the birth-and-death process better explains evolution of the *ANK* gene cluster.

Materials and methods

Identification of homologues at the ANK gene cluster

Information on seven paralogues at Col-0 *ANK* gene cluster was generated from annotation for the Columbia accession of *A. thaliana* (http://www.ncbi.nlm.nih.gov; released in August 2004). The genes we examined in this study, *ANK1*, *ANK2*, *ANK3*, *ANK4*, *ANK5*, *ANK6* and *ANK7*, occur in tandem on chromosome 4, and correspond to *At4g03440, At4g03450, At4g03460, At4g03470, At4g03480, At4g03490* and *At4g03500*, respectively. *At1g03670* and *ANK* gene cluster members (*ANK1–ANK7*) were duplicated genes (Becerra *et al.* 2004). We focussed here only on the tightly linked *ANK1–ANK7* locus.

To determine the counterparts in the Ler genome, full-length DNA of Col-0 *ANK1–ANK7* locus (including intergenic regions), as well as its 10-kb downstream flanking sequence, was initially extracted from GenBank (http://www.ncbi.nlm.nih.gov; released in August 2004), and then used as queries to search matched contigs in the Ler shotgun dataset (Jander *et al.* 2002; http://www.arabidopsis.org/Cereon). BLAST alignments with identity $<80\%$ were excluded from further investigation. To confirm our results, all sequences from candidate contigs were used in BLAST searches against the Col-0 full genome sequence. Any contig that produced effective hits not matching the right position was discarded. In the end, the target contigs were aligned with a reference sequence using Sequencher 4.0 (Gene Codes Co., Ann Arbor, USA). One or both sides of alignments were truncated properly, if necessary, especially where sequencing errors occurred more frequently.

Plant material preparation

About 15 accessions of *A. thaliana* from a worldwide sample were used (Col-0, USA; La-0, Poland; Fm-15, New York, USA; Kil-0, England; Hs-12, Massachusetts, USA; Yo-0, California, USA; Up-14, Michigan, USA; Tsu-0, Japan; Bg-4, Seattle, USA; Ler-0, Germany; Cs-25, Germany; Bur-0, England; Cvi-0, Cape Verde Islands; Pu2-7, Czech Republic; Sq-8, England). These accessions were obtained from the *Arabidopsis* Biological Resource Center (The Ohio State University, Columbus, USA), the Nottingham *Arabidopsis* Stock Centre (University of Nottingham, Loughborough, UK), and independent collectors. A minimum of 10 seeds from each *A. thaliana* accession were planted in small pots in soil consisting of 70% nutritive earth and 30% sand. Plants were cultivated in a 20°C growth chamber under a light–dark regime of 12:12 h, and transferred to a greenhouse after sprouting.

Genotyping

A 1514-bp length polymorphism was determined by three steps of Long PCR with LA Taq$^{\mathrm{TM}}$ DNA polymerase (TaKaRa), performed in 15 μl containing 15 ng of template DNA, 0.2 μM of each primer, 1.5 μl 10× PCR buffer, 1.5 μl of 25 mM $MgCl_2$, 2.4 μl of 2.5 mM of dNTP, and 0.75 U of Taq polymerase. A couple of primers, located at about 0.5 kb upstream or downstream from the target site, were designed according to the Col-0 sequence. Amplification was performed in a 96-well format by using an MJR PTC-200 thermocycler (MJ Research Inc., Waltham, USA). PCR reactions were initiated at 94°C for 2 min, followed by 30 cycles of denaturation at 94°C for 30 s, annealing at 50–56°C for 40 s, an extension at 68°C for 3 min, and a final extension step at 68°C for 10 min. Two different size bands would be expected, and were purified, and then sequenced to find the breakpoints.

DNA extraction, PCR and sequencing

DNA extraction was carried out with a modified CTAB procedure from *A. thaliana* mature leaves (Bergelson *et al.* 1998). Most part of *ANK4-ANK5* in Ler-0 was amplified by three PCR primer pairs designed from the Col-0 genome sequence (1542591U22, 5′ AAG AGC TGC CAC TAC TAG AAG A, and 1543947L22, 5′ GAA GCC ATG GAA CAC CAC TAA C; 1543884U22, 5′ TCT AGA TGA CTC CAT GTA GCA G, and 1545186L22, 5′ ATC GCC GAA CTA AAT CCC AAC C; 1545003U22, 5′ TCG TCT AGG TTT CCT CAC ATT T, and 1547271L22, 5′ ATG GAG ACA GCT GCC TGT CTA G). The last primer pair was also used to amplify another 13 worldwide accessions for *ANK5* homologous fragments. Linked segments overlapped tens of base pairs. PCR products were purified by 33% PEG 8000 and 5 M NaCl, and sequenced directly with both primers, using ABI/Perkin cycle sequencing Bigdye chemistry and an ABI 377 automated sequencer (Applied Biosystems, Foster City, USA). The new sequences reported in this manuscript have been deposited with the GenBank Data Library under accession numbers DQ842102–DQ842122.

Data analysis

Full-length gene sequences from seven *ANK* paralogues in Col-0 were aligned by the CLUSTALX package (Thompson *et al.* 1997), and alignments were optimized manually. Phylogenetic trees were constructed by the neighbour-joining (NJ) method (Saitou and Nei 1987), using PAUP* 4.0 (Swofford 2000). All newly generated sequences were aligned using Sequencher 4.0.5 (Gene Codes Co., Ann Arbor, USA), with slight manual correction for unambiguous polymorphism sites. We used the DnaSP 4.0 program (Rozas *et al.* 2003) to calculate population-genetic parameters such as nucleotide diversity (π, Nei 1987), the estimate of theta (θ, Watterson 1975), the average number of nucleotide differences per site between haplotypes (D_{xy}, Nei 1987), and the K_a/K_s ratio, to complete a sliding-window analysis, to determine silent and replacement sites. Tajima's (1989) test was used to examine compatibility with the neutral mutation theory. Without special explanation, indels were excluded in most calculations. Otherwise, when length polymorphism sites were considered in these algorithms, total number of polymorphic sites was increased by one site for each indel.

Results

Comparisons of homologue divergence

Given that the number of exons and the gene length varied a lot among *ANK* gene cluster paralogues in Col-0 (table 1), as well as the nucleotide divergence between these different paralogues, the sequence identity for any two ranged from 40.9% to 75.0% (table 2). We also calculated sequence similarity for *ANK1*, *ANK2*, *ANK3*, *ANK4* and *ANK5* alleles between Col-0 and Ler-0, and the values were 86.1% (*ANK4*), 90.9% (*ANK5*), and 99.4–100% (*ANK1*, *ANK2*, *ANK3*). In the region of *ANK6–ANK7*, the sequences were absent in the Ler shotgun dataset (about two-fold coverage of *Arabidopsis* genome). Because the dataset presumably covered at least some sequence from over 95% of all genes, *ANK6–ANK7* was probably deleted in the Ler-0 genome, and was not considered here (table 2). The average sequence identity between alleles at *ANK1–ANK5* was 95.2%, much higher than the 55.6% among paralogues, suggesting less frequent intralocus recombination or gene conversion, and independent evolution after tandem duplication at this cluster. Phylogenetic analysis (figure 1) further indicated that less DNA sequence homogeneity existed in paralogues, which was inconsistent with the rapid process of concerted evolution, but consistent with a birth-and-death process.

Table 1. Comparisons of paralogues from Col-0 *ANK* gene cluster.

Gene ID	Symbol	No. of exons	No. of introns	Gene length (bp)	Coding seq. length (bp)
At4g03440	*ANK1*	5	4	2935	2256
At4g03450	*ANK2*	3	2	2291	1926
At4g03460	*ANK3*	6	5	3708	2034
At4g03470	*ANK4*	3	2	2225	2052
At4g03480	*ANK5*	6	5	2848	1854
At4g03490	*ANK6*	6	5	3167	1764
At4g03500	*ANK7*	4	3	3480	1959

Related information was obtained from Col-0 annotation in GenBank, see Materials and methods.

Table 2. Pairwise sequence identity of *ANK* gene cluster coding sequences for Col-0 (C) and Ler-0 (L)

	C1	C2	C3	C4	C5	C6
L1	99.4					
L2		99.5				
L3			100.0			
L4				86.1		
L5					90.9	
C2	69.6					
C3	52.6	55.3				
C4	65.0	63.3	52.5			
C5	75.0	72.9	57.3	69.0		
C6	42.0	41.8	42.5	40.9	44.9	
C7	54.2	55.1	74.4	53.7	58.1	41.6

Data for L1, L2 and L3 were collected from http://www.arabidopsis.org/Cereon; data for L4 and L5 were obtained from the present study; data for L6 and L7 were unavailable, and were not considered here.

Pattern of nucleotide variation

Nordborg *et al.* (2005) sequenced part of *ANK4* started at base pair 1542316 (in chromosome 4 physical map) from 96 worldwide *A. thaliana* accessions, and observed an extremely high level of polymorphism, most of which was attributed to two clear highly differentiated haplotypes. The allele frequencies were estimated to be 0.62 and 0.38, respectively. To determine whether this enhanced variation was specific to this locus, or spread across the whole *ANK* gene cluster, we sequenced its counterpart in Ler-0 based on the Col-0 4632-bp DNA sequence, downstream from site 1542316. The sequenced regions spanned most part of *ANK4*, the intergenic region between *ANK4* and *ANK5*,

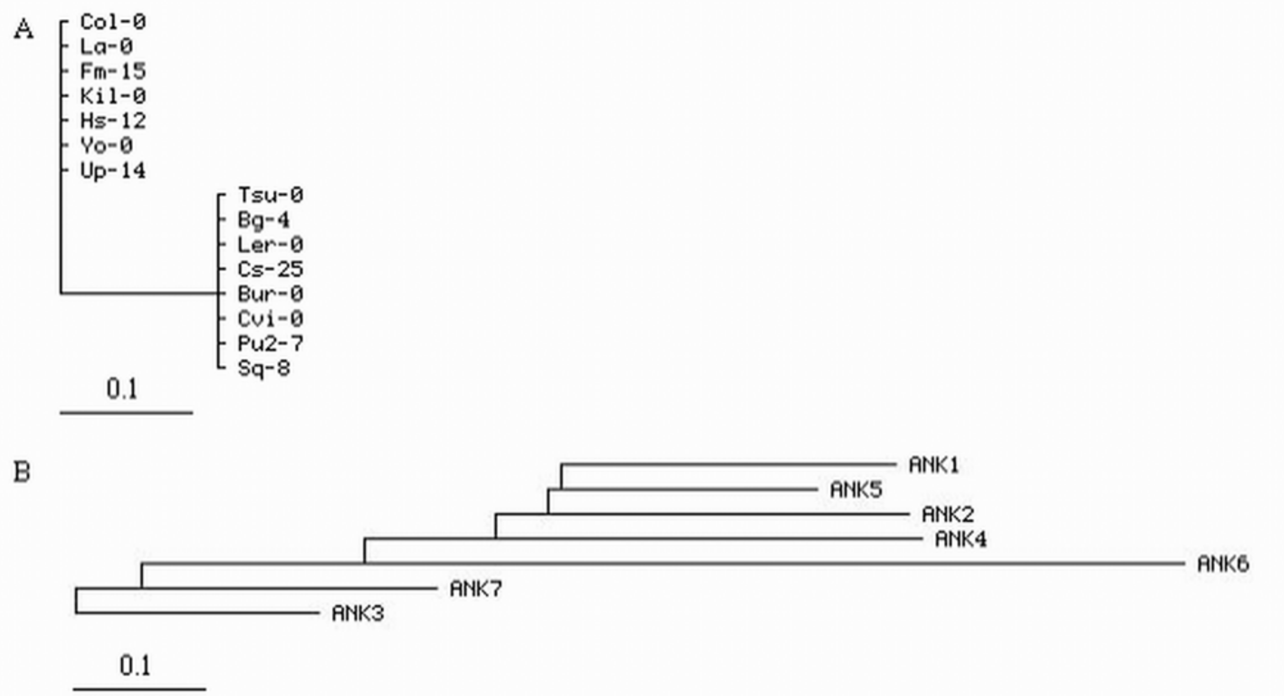

Figure 1. Relationships between allele classes and among paralogues at the *ANK* gene cluster in *A. thaliana*. (A) Neighbour-joining tree based on *ANK5* 2215 alignment base pairs from 15 accession sequences. (B) Neighbour-joining tree of seven paralogues from Col-0 accession.

Table 3. Genetic variation at *ANK* gene cluster.

	Coding region				
	Synonymous	Nonsynonymous	Noncoding	Silent sites	Total sites
ANK4–ANK5 (n = 2)					
No. of sites	371.6	1191.4	1305	1651.6	2868
S	86	127	151	237	364
D_{xy}	0.2225	0.1094	0.1172	0.1409	0.1269
ANK1–ANK3 (n = 2)					
No. of sites	461.8	1464.2	6098	6545.8	8024
S	0	5	53	53	58
D_{xy}	0	0.0034	0.0087	0.0081	0.0072
ANK5 (n = 15)					
No. of sites	29.6	102.4	400	418.6	532
S	2	11	44	46	57
π	0.0361	0.0573	0.0603	0.0586	0.0571
θ	0.0208	0.0330	0.0348	0.0338	0.0330
D_{xy}	0.0676	0.1074	0.1131	0.1099	0.1071

S, π, θ, D_{xy} are number of segregating sites, average nucleotide pairwise difference, Watterson estimator from the number of segregating sites and the number of sequences, and nucleotide divergence, respectively. Col-0 and Ler-0 were selected to estimate haplotypic divergence in most cases.

Region	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	EX4	IN4	EX5	EX5	EX5	EX5	EX5	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT
Sites	0007	0010	0018	0029	0030	0035	0041	0052	0070	0090	0091	0137	0136	0164	0175	0180	0189	0228	0283	0295	0300	0309	0314	0831	0834	0839	0847	0850	0857	0861	0864	0870	0871	0873	0882	0889	0906	0909	0931	0932	0934	0940	0957
Col-0	A	T	A	A	C	-	C	C	i1	-	i2	G	i3	T	d4	C	C	i5	A	A	T	C	i6	G	GTA	i7	GT	i8	AT	G	C	T	G	-	T	i9	G	T	A	AT	A	G	A
La-0	.	.	.	.	.	-	.	.	i1	-	i2	.	i3	.	d4	.	.	i5	.	.	.	.	i6	.	.	i7	.	i8	.	.	.	.	.	-	.	i9	.	.	.	.	.	.	.
Fm-15	.	.	.	.	.	-	.	.	i1	-	i2	.	i3	.	d4	.	.	i5	.	.	.	.	i6	.	.	i7	.	i8	.	.	.	.	.	-	.	i9	.	.	.	.	.	.	.
Kil-0	.	.	.	.	.	-	.	.	i1	T	i2	A	i3	.	d4	.	.	i5	.	.	.	.	i6	.	.	i7	.	i8	.	.	.	.	.	-	.	i9	.	.	.	.	.	.	.
Hs-12	.	.	.	.	.	-	.	.	i1	-	i2	.	i3	.	d4	.	.	i5	.	.	.	.	i6	.	.	i7	.	i8	.	.	.	.	.	-	.	i9	.	.	.	.	.	.	.
Yo-0	.	.	.	.	.	-	.	.	i1	-	i2	.	i3	.	d4	.	.	i5	.	.	.	.	i6	.	.	i7	.	i8	.	.	.	.	.	-	.	i9	.	.	.	.	.	.	.
Up-14	.	.	.	.	.	-	.	.	i1	-	i2	.	i3	.	d4	.	.	i5	.	.	.	.	i6	.	.	i7	.	i8	.	.	.	.	.	-	.	i9	.	.	.	.	.	.	.
Tsu-0	G	G	G	T	T	ATCT	T	A	d1	-	d2	-	d3	C	i4	T	T	d5	G	-	A	A	d6	-	-	d7	-	d8	-	-	G	C	T	CT	-	d9	T	A	G	-	C	T	T
Bg-4	G	G	G	T	T	ATCT	T	A	d1	-	d2	-	d3	C	i4	T	T	d5	G	-	A	A	d6	-	-	d7	-	d8	-	-	G	C	T	CT	-	d9	T	A	G	-	C	T	T
Ler-0	G	G	G	T	T	ATCT	T	A	d1	-	d2	-	d3	C	i4	T	T	d5	G	-	A	A	d6	-	-	d7	-	d8	-	-	G	C	T	CT	-	d9	T	A	G	-	C	T	T
Cs-25	G	G	G	T	T	ATCT	T	A	d1	-	d2	-	d3	C	i4	T	T	d5	G	-	A	A	d6	-	-	d7	-	d8	-	-	G	C	T	CT	-	d9	T	A	G	-	C	T	T
Bur-0	G	G	G	T	T	ATCT	T	A	d1	-	d2	-	d3	C	i4	T	T	d5	G	-	A	A	d6	-	-	d7	-	d8	-	-	G	C	T	CT	-	d9	T	A	G	-	C	T	T
Cvi-0	G	G	G	T	T	ATCT	T	A	d1	-	d2	-	d3	C	i4	T	T	d5	G	-	A	A	d6	-	-	d7	-	d8	-	-	G	C	T	CT	-	d9	T	A	G	-	C	T	T
Pu2-7	G	G	G	T	T	ATCT	T	A	d1	-	d2	-	d3	C	i4	T	T	d5	G	-	A	A	d6	-	-	d7	-	d8	-	-	G	C	T	CT	-	d9	T	A	G	-	C	T	T
Sq-8	G	G	G	T	T	ATCT	T	A	d1	-	d2	-	d3	C	i4	T	T	d5	G	-	A	A	d6	-	-	d7	-	d8	-	-	G	C	T	CT	-	d9	T	A	G	-	C	T	T

Region	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT
Sites	1960	1961	1976	1981	1991	1992	1993	2000	2030	2039	2041	2055	2058	2067	2083	2097	2099	2111	2120	2121	2122	2128	2129	2134	2138	2143	2146	2166	2167	2173	2198	2215	2217	2219	2220	2221	2222	2225	2239	2242	2245	2247	2254	2258
Col-0	C	T	G	C	G	C	T	i10	C	T	A	T	-	G	d11	T	A	A	T	T	T	A	C	T	T	T	TT	G	T	d12	-	T	C	T	T	T	C	A	C	C	C	C	GCT	T
La-0	.	.	.	.	.	.	.	i10	.	.	.	.	-	.	d11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	d12	-	.	.	.	.	.	.	.	.	.	.	.	.	.
Fm-15	.	.	.	.	.	.	.	i10	.	.	.	.	-	.	d11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	d12	-	.	.	.	.	.	.	.	.	.	.	.	.	.
Kil-0	.	.	.	.	.	.	.	i10	.	.	.	.	-	.	d11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	d12	-	.	.	.	.	.	.	.	.	.	.	.	.	.
Hs-12	.	.	.	.	.	.	.	i10	.	.	.	.	-	.	d11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	d12	-	.	.	.	.	.	.	.	.	.	.	.	.	.
Yo-0	.	.	.	.	.	.	.	i10	.	.	.	.	-	.	d11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	d12	-	.	.	.	.	.	.	.	.	.	.	.	.	.
Up-14	.	.	.	.	.	.	.	i10	.	.	.	.	-	.	d11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	d12	-	.	.	.	.	.	.	.	.	.	.	.	.	.
Tsu-0	A	A	T	T	C	T	G	d10	A	-	G	C	GA	A	i11	G	T	T	A	A	C	G	T	A	C	C	-	A	C	i12	TA	C	T	C	-	A	G	G	T	T	T	G	.	C
Bg-4	A	A	T	T	C	T	G	d10	A	-	G	C	GA	A	i11	G	T	T	A	A	C	G	T	A	C	C	-	A	C	i12	TA	C	T	C	-	A	G	G	T	T	T	G	.	C
Ler-0	A	A	T	T	C	T	G	d10	A	-	G	C	GA	A	i11	G	T	T	A	A	C	G	T	A	C	C	-	A	C	i12	TA	C	T	C	-	A	G	G	T	T	T	G	.	C
Cs-25	A	A	T	T	C	T	G	d10	A	-	G	C	GA	A	i11	G	T	T	A	A	C	G	T	A	C	C	-	A	C	i12	TA	C	T	C	-	A	G	G	T	T	T	G	.	C
Bur-0	A	A	T	T	C	T	G	d10	A	-	G	C	GA	A	i11	G	T	T	A	A	C	G	T	A	C	C	-	A	C	i12	TA	C	T	C	-	A	G	G	T	T	T	G	.	C
Cvi-0	A	A	T	T	C	T	G	d10	A	-	G	C	GA	A	i11	G	T	T	A	A	C	G	T	A	C	C	-	A	C	i12	TA	C	T	C	-	A	G	G	T	T	T	G	-	C
Pu2-7	A	A	T	T	C	T	G	d10	A	-	G	C	GA	A	i11	G	T	T	A	A	C	G	T	A	C	C	-	A	C	i12	TA	C	T	C	-	A	G	G	T	T	T	G	.	C
Sq-8	A	A	T	T	C	T	G	d10	A	-	G	C	GA	A	i11	G	T	T	A	A	C	G	T	A	C	C	-	A	C	i12	TA	C	T	C	-	A	G	G	T	T	T	G	.	C

Figure 2. DNA polymorphisms detected at *ANK5* in 15 worldwide accessions of *A. thaliana*. Different regions are indicated as EX (exon), IN (intron), and INT (intergenic). The numbers above reference sequence Col-0 denote polymorphic site positions in alignment. A dot represents identical nucleotide to Col-0; i/d means insertion or deletion of some DNA sequences. Indels with length less than 5 bp are displayed directly, others are given here: i1/d1, TAT CAC TTG TGA GAG CGA TG; i2/d2, TTG AAC CGT CCT GGT AAT AAA ATC CAA GGA AAA ACC TCT ACC TTA; i3/d3, CTT CAC AAT TGG AAG GGA GAA AAT C; i4/d4, CGC A; i5/d5, AGT TCT TTG CTA GAA; i6/d6, 1514 bp; i7/d7, CGT AGT A; i8/d8, TAG AGG; i9/d9, AAA AAT ACT GAC AG; i10/d10, TGC CTT GCA GGG CAA TAA AAC CTC; i11/d11, TGC CAA A; i12/d12, TAT CTT TGT TGC TGG AGG CA.

and a part of *ANK5*. We therefore aligned the two sets of sequences, and subsequently calculated different genetic parameters (table 3).

A total of 364 nucleotide polymorphisms were detected in 4793 aligned base pairs (excluding 1925 alignment gaps). Nucleotide divergence over all sites was 0.1269, much higher than the typically observed <0.01 in the whole genome of *Arabidopsis* (Bergelson *et al.* 1998). The ratio of K_a to K_s was 0.98 for *ANK5* and 0.43 for *ANK4* (table 4), indicating different evolutionary patterns in different gene regions. The largest indel was a 1514-bp deletion observed in Ler-0, relative to Col-0. To determine whether this length polymorphism was fixed between Hap-C and Hap-L, or was unique in the Ler-0 genome, we sequenced the counterparts in another 13 worldwide accessions, and aligned them against the Col-0 2215-bp sequence across this length polymorphic site. A total of 57 substitution sites and 30 length polymorphic sites were observed in 2259 aligned base pairs (figure 2).

Surprisingly, nucleotide substitution sites were all fixed substitutions, without any rare alleles or shared mutations at this locus. The estimates of nucleotide diversity (π) and Watterson theta (θ) over all sites were 0.0571 and 0.0330, respectively (table 3), significantly incompatible with the neutral hypothesis (Tajima's $D = 3.175, P < 0.001$). The estimate of D_{xy} was 0.1071 for all sites, and 0.1099 for silent sites (table 3).

In contrast to *ANK4–ANK5*, sequence divergence between Col-0 and Ler-0 in the *ANK1–ANK3* region was estimated to be 0.0072 for all sites and 0.0081 for silent sites (table 3), which is in the range of *Arabidopsis* genome-level data. The lack of elevated nucleotide diversity extended to a distance of at least 8 kb, except for a short divergence peak (spanning ~0.3 kb intergenic region between *ANK1* and *ANK2*; data not shown). Such different polymorphic patterns observed in members of a gene cluster have rarely been reported and will be discussed in detail in the next section.

Table 4. Estimates of K_a/K_s ratio at *ANK5* and *ANK4* loci.

	K_s	K_a	K_a/K_s
ANK5	0.1103	0.1076	0.98
ANK4	0.2786	0.1192	0.43

K_s, Number of synonymous substitutions per synonymous site; K_a, number of nonsynonymous substitutions per nonsynonymous site.

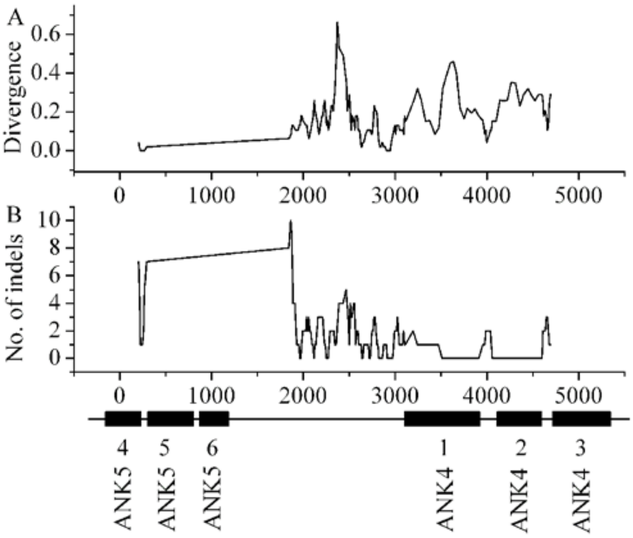

Figure 3. Sliding-window plots along *ANK* gene cluster regions. (A) Divergence between Col-0 and Ler-0 accessions (representing Hap-C and Hap-L) at silent sites. Windows include 100 silent sites, with successive displacements of 25 sites. The black boxes represent *ANK* coding regions, marked with corresponding positions. (B) Indel number distribution in different regions. The nearly horizontal line from site 313 to site 1826 indicates a large indel polymorphism detected in this study.

To determine the distribution of nucleotide substitutions, silent divergence between Col-0 and Ler-0 was estimated by sliding a window of 100 silent sites across each of the studied regions (figure 3,A). The peak at site 2464 in the intergenic region was due to 24 unmatched base pairs. Similar results were observed in the coding region of *ANK4*. Three short divergence peaks were positioned at sites 3242, 3598 and 4262, where the divergence estimates were 0.32, 0.46 and 0.35, respectively (data from sliding-window analysis).

Variation in insertions and deletions

The pattern of indels contains valuable population-genetic information, and plays an important role in genome evolution. Recent investigations clarified that the majority of divergence between closely related DNA samples was due to indels (Britten *et al.* 2003). Indels were also detected throughout the tandemly duplicate *Arabidopsis ATTI* genes (Clauss and Mitchell-Olds 2004), with an indel peak roughly coinciding with the peak of nucleotide polymorphism. To address the distribution of indel numbers at the *ANK* loci, a sliding-window analysis with 100 silent sites across each of the studied regions was done (figure 3,B). As expected, the number of indels in a given window was positively associated with the number of nucleotide substitutions in the intergenic region ($r = 0.45$, 100 point sets; $P << 0.001$; figure 3B), which is consistent with data for the *ATTI* loci. Surprisingly, a negative relationship between indels and nucleotide polymorphisms was observed in the *ANK4* gene region ($r = -0.58$, 52 point sets; $P < 0.001$; figure 3,B).

Characteristics of indels in coding regions of *ANK5* and *ANK4* were also examined. Except for the 1514-bp indel length polymorphism at *ANK5*, the other four indel sizes (1, 4, 86 and 4 bp) were not found to be multiples of three, suggesting that *ANK5* had no function at least within one allele class. In *ANK4*, three (3, 6 and 3 bp) out of four indel lengths were frame-conservative, and the only exception (38-bp indel) was potentially a frame-shift mutation. The features of indels at *ANK5* and *ANK4*, coupled with an analysis of K_a/K_s ratios (0.98 and 0.43, see table 4) indicated that both had evolved into pseudogenes in one or both allele classes after gene duplications. No length polymorphism was detected in the coding regions of *ANK1*, *ANK2* and *ANK3*.

Discussion

Data analysis from Nordborg *et al.* (2005) showed that 76 out of 92 polymorphic sites in 566 aligned base pairs (excluding indels) were fixed between two major haplotypes with intermediate frequencies, suggesting that this extremely high level of diversity at the *ANK* gene cluster might arise from maintenance of DNA dimorphism. Complete linkage disequilibrium (LD) and the total number of polymorphic sites fixed at *ANK5* further enhanced this assumption (figure 2). Successive sliding-window analysis for *ANK4–ANK5* showed no apparent decayed nucleotide polymorphism (figure 3,A). However, this high level of polymorphism quickly disappeared in regions of *ANK1-ANK3*, in spite of the <4-kb distance of the most proximal site from the elevated polymorphic region. We could therefore divide the entire *ANK* gene cluster into three discrete domains: (i) the 5-kb large dimorphic domain containing *ANK4–ANK5*, (ii) a common polymorphic domain spanning *ANK1–ANK3*, and (iii) a domain covering *ANK6–ANK7*, which presumably involved a long deletion event.

Unfortunately, there has been no evidence for mechanisms for maintenance of DNA dimorphism in about half of the previously published dimorphic loci in *A. thaliana* (Kawabe *et al.* 1997; Kawabe and Miyashita 1999; Purugganan and Suddith 1999; Kuittinen and Aguadé 2000; Hauser *et al.* 2001). A higher mutation rate (Yoshida *et al.* 2003), an introgression from another species, or two isolated populations fused before expansion (Teeter *et al.* 2000), as well as recombination and gene conversion (Innan *et al.* 1996; Haubold *et al.* 2002), have been invoked to explain dimorphism. A neutral process with no recombination was also shown to explain dimorphic haplotypes (Aguadé 2001). However, the predominant hypothesis invoked balancing selection (Hanfstingl *et al.* 1994; Tian *et al.* 2002) or frequency-dependent selection (Stahl *et al.* 1999). We noticed that the pronounced intraspecific haplotype divergence at the *Arabidopsis RPP5* gene cluster was maintained by frequency-dependent selection (Noel *et al.* 1999). In contrast, data for the *Arabidopsis ATTI* gene cluster were not supportive of an extant balanced polymorphism at divergence peak regions for nonfunctional *ATTI5*$^+$, less likely a selected target (Clauss and Mitchell-Olds 2004).

In the case of the *ANK* gene cluster examined here, the data show some characteristic signals of balanced polymorphism, such as a high level of nucleotide diversity, maintenance of intermediate-frequency alleles, a reduction in the number of haplotypes, and a high level of LD (Cork and Purugganan 2005). Moreover, full-length cDNA of *ANK1*, *ANK2* and *ANK7* were present in Col-0 (http://www.arabidopsis.org). In spite of the potential deletion of *ANK6–ANK7* in Ler-0, the functional *ANK7* in Col-0 could also qualify the indel polymorphism locus as the potential target of balancing selection. In addition, we do find the matched regions at ~1-kb distance downstream from Col-0 *ANK7* in the Ler shotgun dataset, and observed 87% sequence identity in a 235-bp alignment. The estimate was increased continuously until 98% with the distance extended in the next 5-kb range (BLAST results, data not shown). These additional observations were also concordant with the balancing-selection model, although our data were insufficient to unequivocally confirm this hypothesis.

A second possible explanation for the observed deviation from a neutral equilibrium process are alternative genetic/genomic or demographic factors. Indeed, large-scale genome sequencing of 96 individuals of *A. thaliana* revealed a clear global population structure, as well as a pattern of isolation by distance (Nordborg *et al.* 2005). In contrast to selective effects on some specific loci, however, demographic perturbation is expected to affect all loci. Thus, the two contrasting polymorphism patterns at *ANK1–ANK3* and *ANK4–ANK5*, as well as the unique nucleotide polymorphism downstream of *ANK7*, would tend to exclude this hypothesis.

Recombination and gene conversion may also generate nucleotide diversity in dimorphic regions. Recombination events were indeed detected at many surveyed *Arabidopsis* loci (Hanfstingl *et al.* 1994; Innan *et al.* 1996; Stahl *et al.* 1999; Tian *et al.* 2002). However, such recombinational shuffling cannot create nucleotide substitutions, and, therefore, could not have been responsible for the observed distinct dimorphic haplotypes. It raised the possibility that Ler-0 had undergone a recombination event in the common polymorphic domain that altered the original haplotypic structure, but we considered it unlikely, as the chance of outcrossing between different *Arabidopsis* accessions in a 99% inbreeding species is remote. The average effective population recombination rate in *Arabidopsis* was estimated to be 6×10^{-4} per meiosis between adjacent bases (Tian *et al.* 2002). Indeed, additional sequencing of six other *Arabidopsis* accessions evenly sampled from Hap-C and Hap-L detected no polymorphic substitutions in a 514-bp alignment at *ANK3* (La-0, Up-14, Yo-0, Cs-25, Bur-0, Bg-4; data not shown). Altogether, it appears that recombination and gene conversion are unlikely explanations for the observed dimorphism.

It was important to differentiate the birth-and-death process from that of concerted evolution. Concerted evolution would tend to homogenize genes within a haplotype (or species) through frequent interlocus recombination and gene conversion. There would be no obvious allelic relationship between genes in different haplotypes. Paralogues would be more similar than orthologues (or alleles) (Michelmore and Meyers 1998). Birth-and-death evolution, however, assumes that new genes are created by repeated gene duplication and that some of the duplicate genes are maintained in the genome for a long time, whereas others are deleted or become nonfunctional through accumulation of deleterious mutations (Nei and Hughes 1992; Ota and Nei 1994; Nei *et al.* 1997; Nei and Rooney 2005). Substantial evidence supports the birth-and-death hypothesis at the *ANK* gene cluster, rather than concerted evolution. First, paralogues were less similar than dimorphic alleles, and seemed to have evolved independently. Second, no intralocus or gene conversion was detected at *ANK5*, which was grouped into two clear haplotypes. Third, pseudogenes were indeed detected at the *ANK* loci.

Until recently, how nonrandom gene clusters were formed and maintained in a genome was not well understood (Hurst *et al.* 2004). Our systematic investigation on the *ANK* gene cluster has provided some valuable clues to this issue. The data indicate that the members of the *ANK* gene cluster most probably evolved by the gene duplication and birth-and-death process. Paralogues diverged remarkably after duplication events, and some of them became pseudogenes after long-term evolution. Our results have not only confirmed the data from Nordborg *et al.* (2005) but also detected three different polymorphic domains. The second domain harbours a high degree of dimorphic haplotypes. Also, the characteristics of DNA polymorphism left some footprints of balancing selection. Nevertheless, functional genetic analysis of the *ANK* gene cluster and studies on other clusters in *A. thaliana* are required to further determine the main evolutionary forces acting on this locus.

Acknowledgements

We are grateful to anonymous reviewers for valuable advice and helpful comments on earlier versions of the manuscript. This work was supported by National Science Fund of China (30470925 and 40371071) and '863' (2004AA218023) to Y.-H. Y.

References

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

Becerra C., Jahrmann T., Puigdomènech P. and Vicient C. M. 2004 Ankyrin repeat-containing proteins in *Arabidopsis*: characterization of a novel and abundant group of genes coding ankyrin-transmembrane proteins. *Gene* **340**, 111–121.

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

Britten R. J., Rowen L., Williams J. and Cameron R. A. 2003 Majority of divergence between closely related DNA samples is due to indels. *Proc. Natl. Acad. Sci. USA* **100**, 4661–4665.

Clauss M. J. and Mitchell-Olds T. 2004 Functional divergence in tandemly duplicated *Arabidopsis thaliana* trypsin inhibitor genes. *Genetics* **166**, 1419–1436.

Cork J. M. and Purugganan M. D. 2005 High-diversity genes in the *Arabidopsis* genome. *Genetics* **170**, 1897–1911.

Gu X. and Nei M. 1999 Locus specificity of polymorphic alleles and evolution by a birth-and-death process in mammalian MHC genes. *Mol. Biol. Evol.* **16**, 147–156.

Hanfstingl U., Berry A., Kellogg E. A., Costa J. T. III, Rudiger W. and Ausubel F. M. 1994 Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection. *Genetics* **138**, 811–828.

Haubold B., Kroymann J., Ratzka A., Mitchell-Olds T. and Wiehe T. 2002 Recombination and gene conversion in a 170-kb genomic region of *Arabidopsis thaliana*. *Genetics* **161**, 1269–1278.

Hauser M. T., Harr B. and Schlotterer C. 2001 Trichome distribution in *Abidopsis thaliana* and its close relative *A. lyrata*: molecular analysis of the candidate gene *GLABROUS1*. *Mol. Biol. Evol.* **18**, 1754–1763.

Hurst L. D., Pál C. and Lercher M. J. 2004 The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310.

Innan H., Tajima F., Terauchi R. and Miyashita N. T. 1996 Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. *Genetics* **143**, 1761–1770.

Jander G., Norris S. R., Rounsley S. D., Bush D. F., Levin I. M. and Last R. L. 2002 *Arabidopsis* map-based cloning in the post-genome era. *Plant Physiol.* **129**, 440–450.

Kawabe A. and Miyashita N. T. 1999 DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana*. *Genetics* **153**, 1445–1453.

Kawabe A., Innan H., Terauchi R. and Miyashita N. T. 1997 Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana*. *Mol. Biol. Evol.* **14**, 1303–1315.

Koop B. F., Rowen L., Wang K., Kuo C. L., Seto D., Lenstra J. A. *et al.* 1994 The human T-cell receptor TCRAC/TCRDC (C alpha/C delta) region: organization, sequence, and evolution of 97.6 kb of DNA. *Genomics* **19**, 478–493.

Krumlauf R. 1992 Evolution of the vertebrate Hox homeobox genes. *BioEssays* **14**, 245–252.

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

Lercher M. J., Urrutia A. O. and Hurst L. D. 2002 Clustering of housekeeping genes provides a unified model of gene order in the human genome *Nat. Genet.* **31**, 180–183.

Li J., Ji C., Zheng H., Fei X., Zheng M., Dai J. *et al.* 2005 Molecular cloning and characterization of a novel human gene containing four ankyrin repeat domains. *Cell Mol. Biol. Lett.* **10**, 185–193.

Martin D. I., Fiering S. and Groudine M. 1996 Regulation of beta-globin gene expression: straightening out the locus. *Curr. Opin. Genet. Dev.* **6**, 488–495.

Michelmore R. W. and Meyers B. C. 1998 Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**, 1113–1130.

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

Nei M. and Hughes A. 1992 Balanced polymorphism and evolution by the birth and death process in the MHC loci. In *Proceedings of the Eleventh Histocompatibility Workshop and Conference* (ed. K. Tsuji, M. Aizawa and T. Suzuki), pp. 27–38. Oxford University Press, Oxford.

Nei M. and Rooney A. P. 2005 Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152.

Nei M., Gu X. and Sitnikova T. 1997 Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci. USA* **94**, 7799–7806.

Nei M., Rogozin I. B. and Piontkivska H. 2000 Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc. Natl. Acad. Sci. USA* **97**, 10866–10871.

Noel L., Moores T. L., van der Biezen E. A., Parniske M., Daniels M. J., Parker J. E. *et al.* 1999 Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. *Plant Cell* **11**, 2099–2111.

Nordborg M., Hu T. T., Ishino Y., Jhaveri J., Toomajian C., Zheng H. *et al.* 2005 The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, 1–11.

Ota T. and Nei M. 1994 Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Mol. Biol. Evol.* **11**, 469–482.

Parniske M., Hammond-Kosack K. E., Golstein C., Thomas C. M., Jones D. A. *et al.* 1997 Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. *Cell* **91**, 821–832.

Piontkivska H., Rooney A. P. and Nei M. 2002 Purifying selection and birth-and-death evolution in the histone H4 gene family. *Mol. Biol. Evol.* **19**, 689–697.

Purugganan M. D. and Suddith J. I. 1999 Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. *Genetics* **151**, 839–848.

Richter T. E. and Ronald P. C. 2000 The evolution of disease resistance genes. *Plant Mol. Biol.* **42**, 195–204.

Richter T. E., Prior T. J., Bennetzen J. L. and Hulbert S. H. 1995 New rust resistance specificities associated with recombination in the *Rp1* complex in maize. *Genetics* **141**, 373–381.

Ronald P. C. 1998 Resistance gene evolution. *Curr. Opin. Plant Biol.* **1**, 294–298.

Rozas J., Sanchez-DelBarrio J. C., Messeguer X. and Rozas R. 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497.

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

Sedgwick S. G. and Smerdon S. J. 1999 The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem. Sci.* **24**, 311–316.

Song W. Y., Pi L. Y., Wang G. L., Gardner J., Holsten T. and Ronald P. C. 1997 Evolution of the rice *Xa21* disease resistance gene family. *Plant Cell* **9**, 1279–1287.

Stahl E. A., Dwyer G., Mauricio R., Kreitman M. and Bergelson J. 1999 Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis. Nature* **400**, 667–671.

Swofford D. L. 2000 *PAUP*: phylogenetic analysis using parsimony*. Sinauer, Sunderland.

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

Teeter K., Naeemuddin M., Gasperini R., Zimmerman E., White K. P., Hoskins R. *et al.* 2000 Haplotype dimorphism in a SNP collection from *Drosophila melanogaster*. *J. Exp. Zool.* **288**, 63–75.

Thompson J. D., Gibson T. J., Plewniak F., Jeanmougin F. and Higgens D. G. 1997 The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.

Tian D., Araki H., Stahl E., Bergelson J. and Kreitman M. 2002 Signature of balancing selection in *Arabidopsis. Proc. Natl. Acad. Sci. USA* **99**, 11525–11530.

Wang S., Magoulas C. and Hickey D. 1999 Concerted evolution within a trypsin gene cluster in *Drosophila. Mol. Biol. Evol.* **16**, 1117–1124.

Watterson G. A. 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276.

Weiss A., Mcdonough D., Wertman B., Acakpo-satchivi L., Montgomery K., Kucherlapati R. *et al.* 1999 Organization of human and mouse skeletal myosin heavy chain gene clusters is highly conserved. *Proc. Natl. Acad. Sci. USA* **96**, 2958–2963.

Yoshida K., Kamiya T., Kawabe A. and Miyashita N. T. 2003 DNA polymorphism at the *ACAULIS5* locus of the wild plant *Arabidopsis thaliana. Genes Genet. Syst.* **78**, 11–21.

Received 2 April 2006