J Mol Evol (2007) 64:237–247
DOI: 10.1007/s00239-006-0097-8

JOURNAL OF MOLECULAR EVOLUTION

Patterns of DNA Variation Among Three Centromere Satellite Families in *Arabidopsis halleri* and *A. lyrata*

Akira Kawabe, Deborah Charlesworth

Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories King's Buildings, West Mains Road, Edinburgh EH9 3JT, UK

Received: 19 April 2006 / Accepted: 3 October 2006 [*Reviewing Editor:* Dr. Brian Morton]

Abstract. We describe patterns of DNA variation among the three centromeric satellite families in *Arabidopsis halleri* and *lyrata*. The newly studied subspecies (*A. halleri* ssp. *halleri* and *A. lyrata* ssp. *lyrata* and *petraea*), like the previously studied *A. halleri* ssp. *gemmifera* and *A. lyrata* ssp. *kawasakiana*, have three different centromeric satellite families, the older pAa family (also present in *A. arenosa*) and two families, pAge1 and pAge2, that probably evolved more recently. Sequence variability is high in all three satellite families, and the pAa sequences do not cluster by their species of origin. Diversity in the pAge2 family is complex, and different from variation among copies of the other two families, showing clear evidence for exchange events among family members, especially in *A. halleri* ssp. *halleri*. In *A. lyrata ssp. lyrata* there is some evidence for recent rapid spread of pAge2 variants, suggesting selection favoring these sequences.

Key words: Centromere — Satellite — Repetitive sequence — *Arabidopsis halleri* — *Arabidopsis lyrata* — Gene conversion

Introduction

The satellite repetitive sequences are major components of noncoding sequence regions of higher eukaryote genomes (Charlesworth et al. 1994; Csink and Henikoff 1998). Satellite DNAs are repeats of 100– to 800–bp unit sequences, usually organized as tandem repeats. They can be highly variable within and between species, and repeats on the same chromosome vary. Satellite DNAs were once considered to be nonfunctional selfish elements that increase and decrease their frequency without any advantage or disadvantage for organisms. However, some satellite families have specific functions and are under natural selection. Centromeric satellite sequences are one such functionally important type of satellite in many eukaryote species (Csink and Henikoff 1998; A.E. Hall et al. 2004).

Centromere regions of various eukaryotes, both plants and animals, consist of highly repetitive satellite sequences (Choo 1997), although many exceptions have been reported (Lo et al. 2001; Saffery et al. 2003; Nagaki et al. 2004). Centromeric satellite sequences in many species have an important role for centromere functions, including chromosome segregation and chromosome positioning during interphase (Choo 1997). Misorientation or missegregation of chromosomes with nonfunctional centromeres causes aneuploidy and is thus severely deleterious for the offspring. Although it is not yet completely clear what determines that a region acts as a centromere, the DNA sequences and chromatin structure, including association with specific proteins, are all

Correspondence to: Akira Kawabe; *email:* Akira.Kawabe@ed.ac.uk

considered important. Species-specific centromeric satellite families appear to coevolve rapidly along with adaptive evolution of centromeric proteins (Csink and Henikoff 1998; Henikoff et al. 2001). Mutations in the relevant centromeric proteins could alter their binding to centromere satellite repeat variants, and copy numbers of favored satellite variants will increase over the generations. Variants in centromeric satellite sequences include not only nucleotide substitutions, but also changes caused by other mechanisms such as gene conversion, replication slippage, and unequal crossing-over (Charlesworth et al. 1994). The centromere regions of many organisms have reduced crossover frequencies (Choo 1998; Copenhaver et al. 1998), but exchange of DNA sequences may still occur. Gene conversion has been inferred in heterochromatic genome regions with low crossover frequencies (Jensen et al. 2002), and centromeric satellite repeats sometimes show clear evidence of unequal exchanges (Cabot et al. 1993).

In the present study, we analyzed nucleotide variation in three centromeric satellite families in *Arabidopsis halleri* and *lyrata* subspecies, which are the closest relatives of *A. thaliana* (Miyashita et al. 1998; Koch et al. 2000). *A. halleri* ssp. *gemmifera*, and *A. lyrata* ssp. *kawasakiana* have three different centromere satellite families (Kawabe and Nasuda 2005), with some chromosome specificity. pAa (similar to the centromere satellite found on all the *A. arenosa* chromosomes) is the only sequence on four of the eight *A. halleri* ssp. *gemmifera* chromosomes, as well as being found on one chromosome together with pAge1; pAge1 is also the sole satellite family on another chromosome, and the other two chromosomes carry exclusively pAge2 (Kawabe and Nasuda 2005). Of the three satellite families, pAa shows the greatest similarity to the *A. thaliana* pAL1 family, and we show evidence below (see Discussion) that pAa is older than pAge1 and pAge2.

Three related species studied previously, *A. thaliana* (Martinez-Zapater et al. 1986; Maluszynska and Heslop-Harrison 1991; Murata et al. 1994), *A. arenosa* (Kamm et al. 1995), and *A. griffithiana* (Heslop-Harrison et al. 2003), each have only one major satellite family on all the chromosomes. Recent sequencing of long stretches of satellite regions from BAC clones of *A. thaliana* relatives revealed that species-wide variability is higher than that within BAC clones in all species analyzed (S.E. Hall et al. 2005). *Sisymbrium irio* satellite sequences are unusually BAC clone-specific (species-wide sequence identity is 83%, versus a mean of about 90% within BACs [S.E. Hall et al. 2005]). FISH analyses using BAC clone probes showed chromosome-specific signal localizations similar to those of the *A. halleri* ssp. *gemmifera* satellite families. However, in other species studied, satellite identities are similar at both scales (about 90% either species-wide or within BAC clones [S. Hall et al. 2005]). This is similar to the findings within satellite families of *A. halleri* ssp. *gemmifera*, but the *A. halleri* ssp. *gemmifera* satellite families are much more highly diverged (divergence is about 30% [Kawabe and Nasuda 2005]). Thus the variety of satellite families in *A. halleri* and *A. lyrata* is unusual, perhaps indicating different selective histories, and/or different functionalities, of the three sequence families.

If these satellite sequences have centromere functions, and if natural selection operates to eliminate or increase variants of these repeat families, previously dominant families that are currently being eliminated from a genome should consist of old satellite sequences and might have high diversity and even non-species-specific variants. In contrast, favored families, which may have increased in abundance very recently, should have lower diversity and their variants should be species-specific and, perhaps, chromosome-specific. The main purpose of this study is to analyze differences in variation in the *A. halleri* versus *A. lyrata* satellite families and to investigate the evolutionary mechanisms involved.

Materials and Methods

Plant Materials

A. halleri ssp. *gemmifera* from Ashibi (Kyoto, Japan), *A. halleri* ssp. *halleri* from Pontresina (Switzerland), *A. lyrata* ssp. *lyrata* individual Ontario4 (Ontario, Canada), and *A. lyrata* ssp. *petraea* plant 99R11-2 (Esja mountain, Iceland) were used for analyses of all three satellite families. In addition to these plants, three plants from the Esja mountain population, two from the other three European populations, and one plant from the Ontario population were also used for pAge1 and pAge2 centromere satellite variation analyses. Total DNAs were isolated from dried leaves using a FastDNA kit (Q-BIOgene) according to the manufacturer's instruction.

Isolation of Centromeric Satellite Families

Three centromeric satellite families were PCR amplified as described by Kawabe and Nasuda (2005). We did not determine the chromosome locations or copy numbers. After agarose gel electrophoresis to identify dimers, trimers, and tetramers, bands were purified and cloned using the TOPO TA cloning kit (Invitrogen). Sequences of cloned PCR products were determined using the DYEnamic ET sequencing kit (Amersham). In the present study, sequences of at least 10 complete satellite units were newly obtained for each of the three satellite families in each species. DNA sequences of the newly determined centromere satellites were deposited in the GenBank databank with the accession numbers DQ872189–DQ872369.

Data Analyses

Sequences were aligned with the forward primer 5′ end as the first nucleotide. Unless otherwise specified, each unit from dimer, trimer, or tetramer sequences was treated as a separate sequence in

Table 1. Summary of nucleotide variation of the centromeric satellite sequences within species: divergence values between sequences were estimated with Jukes-Cantor correction

							Average difference/ bp within	
Family	Species	Population	No. of plants	No. of units[a]	Length (bp)[b]	S	Population	Plant
PAa	Gemmifera	Ashibi	1	6	178	48		0.119
	Halleri	Pontresina	1	15	174	63		0.098
	Petraea	Esja Mt.	1	10	176	44		0.068
	Lyrata	Ontario	1	15	174	54		0.097
	Arenosa		2	262	147	126	0.092	0.091
pAge1	Gemmifera	Ashibi	1	17	165	73		0.095
	Halleri	Pontresina	1	13	163	53		0.072
	Petraea	All	12	149	148	127	0.079	0.078
		Esja Mt.	4	50	159	94	0.078	0.070
		Wales	2	24	163	66	0.078	0.078
		Plech	2	26	163	83	0.100	0.102
		Stubbsand	2	24	155	57	0.085	0.080
		Karhumaki	2	25	161	71	0.067	0.069
	Lyrata	Ontario	2	22	161	68	0.090	0.088
pAge2 dimmers								
	Gemmifera	Ashibi	1	4	349	21		0.033
	Halleri	Pontresina	1	6	350	46		0.051
	Petraea	All	12	89	319	189	0.068	0.057
		Esja Mt.	4	30	346	102	0.057	0.054
		Wales	2	15	336	72	0.057	0.057
		Plech	2	15	338	118	0.078	0.068
		Stubbsand	2	15	346	69	0.059	0.060
		Karhumaki	2	14	346	70	0.047	0.046
	Lyrata	Ontario	2	13	342	86	0.077	0.068

[a] Excluding units with deletions longer than 9 bp.

[b] Length without alignment gaps.

the alignment and in our analyses (i.e., a dimer was treated as two monomer units; as described below, a few analyses of dimer and tetramer sequences of the pAge2 family treat dimers as the units, and this will be explicitly mentioned when it is relevant). Table 1 reports the lengths of sequence units excluding alignment gaps. The previously reported sequences, 5 pAa sequences from *A. arenosa* (Kamm et al. 1995) and 6 pAa, 19 pAge1, and 4 pAge2 sequences from *A. halleri* ssp. *gemmifera* (Kawabe and Nasuda 2005) were also included in the analyses, as well as 257 complete units of the pAa family of *A. arenosa* from S. Hall et al. (2005) whose sequences contained no ambiguous sites (N in the GenBank sequences).

For the complex pAge2 family described in detail below, trimer sequences were found in *A. halleri* ssp. *halleri* and *A. lyrata* ssp. *lyrata*, as well as dimers and tetramer sequences (consisting of two dimers), like those previously observed in *A. halleri* ssp. *gemmifera*. We first aligned monomer units of the dimer and tetramer sequences from *A. halleri* ssp. *gemmifera* and *A. lyrata* ssp. *petraea*. Characteristic differences between the sequences of the first and second monomer units of the dimer sequences (see Results) were then used to classify the pAge2 family sequences into "simple" and "complex." Simple pAge2 dimers are defined as those with distinguishable first and second units (see Fig. 1), first units evidently resembling first units of other dimers of other simple dimers and second units resembling second units of other dimers. Single nucleotide variants are also often present in these sequences. Complex dimer sequences resemble two units of the trimer sequences; in these sequences, at least one unit includes mosaic or chimeric structures that clearly mix parts of first and second units of simple dimers. One dimer sequence of *A. halleri* ssp. *halleri* plus one dimer and one tetramer sequence from *A. lyrata* ssp. *lyrata* had mosaic structures like those seen in trimers.

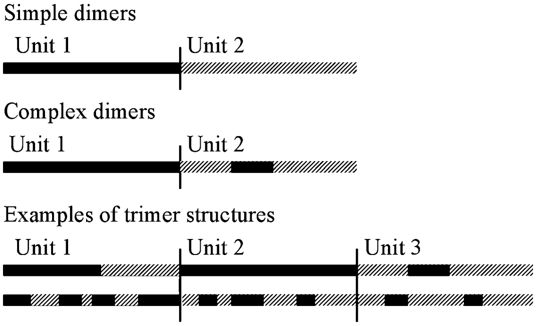

Fig. 1. Schematic diagram of pAge2 sequences, showing "simple" and "complex" dimers and two examples of trimer sequences. The two different units that make up simple dimers are represented by black and hatched lines. Trimer sequences are mosaics of first and second units of simple dimer sequences, and complex dimers are portions of trimer sequences; in the case illustrated, the dimer is composed of units 2 and 3 of the upper trimer sequence.

Average numbers of nucleotide differences between sequences within individuals, divergence between species or subspecies, and minimum numbers of recombination events (Hudson and Kaplan 1985) were estimated with the DnaSP program (Rozas and Rozas 1999). Numbers of gene conversion events were estimated using GENECONV ver. 1.81 (Sawyer 1999). A few sequence units having deletions longer than 9 bp were excluded from these analyses. Phylogenetic trees were constructed by the neighbor-joining (NJ) method using JC-distances and pairwise deletion, using the MEGA2 program package (Kumar et al. 2002).

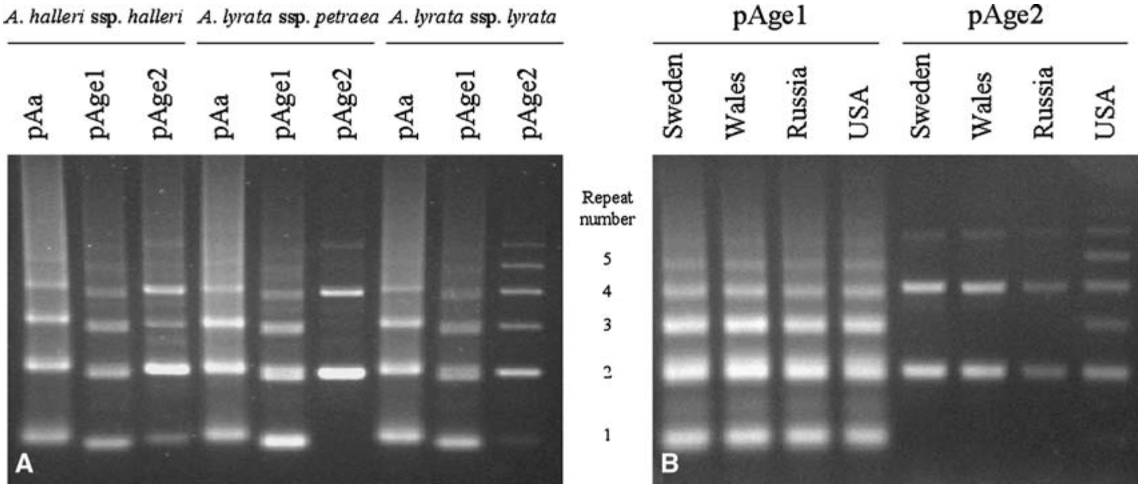

Fig. 2. PCR results for three centromeric satellite families for *A. halleri* ssp. *halleri*, *A. lyrata* ssp. *petraea*, and *A. lyrata* ssp. *lyrata*. **A** One plant each from *A. halleri* ssp. *halleri*, *A. lyrata* ssp. *petraea*, and *A. lyrata* ssp. *lyrata* for the pAa, pAge1, and pAge2 families. **B** pAge1 and pAge2 amplification of *A. lyrata* populations. The pAge2 family primers specifically amplify single units of the dimer sequence (Kawabe and Nasuda 2005).

Results

Existence of Three Centromeric Satellite Families in A. lyrata *and* A. halleri *Subspecies*

We first tested for the presence of each of the three centromeric satellite families in a single *A. lyrata* ssp. *petraea* plant from Iceland, an *A. lyrata* ssp. *lyrata* plant from Ontario, and an *A. halleri* ssp. *halleri* plant from Switzerland, using PCR with the primers described previously (Kawabe and Nasuda 2005). All three families (pAa, pAge1, and pAge2) previously found in *A. halleri* ssp. *gemmifera* and *A. lyrata* ssp. *kawasakiana* were obtained in ladder patterns typical for tandemly repeated sequence arrays (Fig. 2A). We sequenced multiple units from several plants from each of the taxa studied. The patterns from pAa and pAge1 amplifications are similar for all the taxa.

Complex Structure of pAge2 Family Sequences

For pAge2 the PCR results differ between the taxa studied. The previous study of *A. halleri* ssp. *gemmifera* and *A. lyrata* ssp. *kawasakiana* found only even numbers of pAge2 repeat units (dimers), and PCR using primers specific for single units confirmed this (Kawabe and Nasuda 2005). In *A. lyrata* ssp. *petraea*, PCR with the same primers showed the same strong amplification of even-numbered bands, indicating a predominance of ''simple'' dimers (see definition in the previous section and Figs. 1 and 2A). PCR amplifications of pAge2 family sequences in more plants from other European *A. lyrata* populations (at least two plants each from Iceland, Wales, Russia, Sweden, and Germany) showed the same pattern as the initial *A. lyrata* ssp. *petraea* plant studied (from Iceland), with only even-number repeat bands (Fig. 2B). In two of these taxa analyzed (*A. halleri* ssp. *gemmifera* and *A. lyrata* ssp. *petraea*), our new larger numbers of sequences reveal two clear peaks in the numbers of nucleotide differences between monomer sequences (Fig. 3); the peak with the larger number of differences corresponds to differences between the first and the second subunits of dimers, while variation within monomers of these units constitutes the peak with the smaller number of differences. The sequences of these complete dimers include 25 sites at which variants are specific to first or second units (Supplemental Fig. 1).

In contrast, PCR with further plants from two different North American populations (*A. lyrata* ssp. *lyrata*), the initially studied Ontario population and a population from Indiana, and with *A. halleri* ssp. *halleri*, yielded both even- and odd-number repeat bands, with almost the same strength of bands representing even and odd numbers of repeats. We determined sequences of 36 trimers (11 from *A. halleri* ssp. *halleri* and 25 from *A. lyrata* ssp. *lyrata*), together with 2 pentamer sequences, 7 dimers (5 from *A. halleri* ssp. *halleri* and 2 from *A. lyrata* ssp. *lyrata*), and 8 tetramers (1 from *A. halleri* ssp. *halleri* and 7 from *A. lyrata* ssp. *lyrata*) sequences. When multiple sequences were analyzed, plants of these taxa yielded clear peaks only when analysis was restricted to units from ''simple'' dimer sequences, indicating that these dimers are predominantly complete double units, with minor variants at individual nucleotide sites (Supplemental Fig. 1). Thus the regular simple dimeric form of sequence subunits observed in *A. halleri* ssp. *gemmifera* (Kawabe and Nasuda 2005) is still present in these taxa, as well as odd-numbered structures.

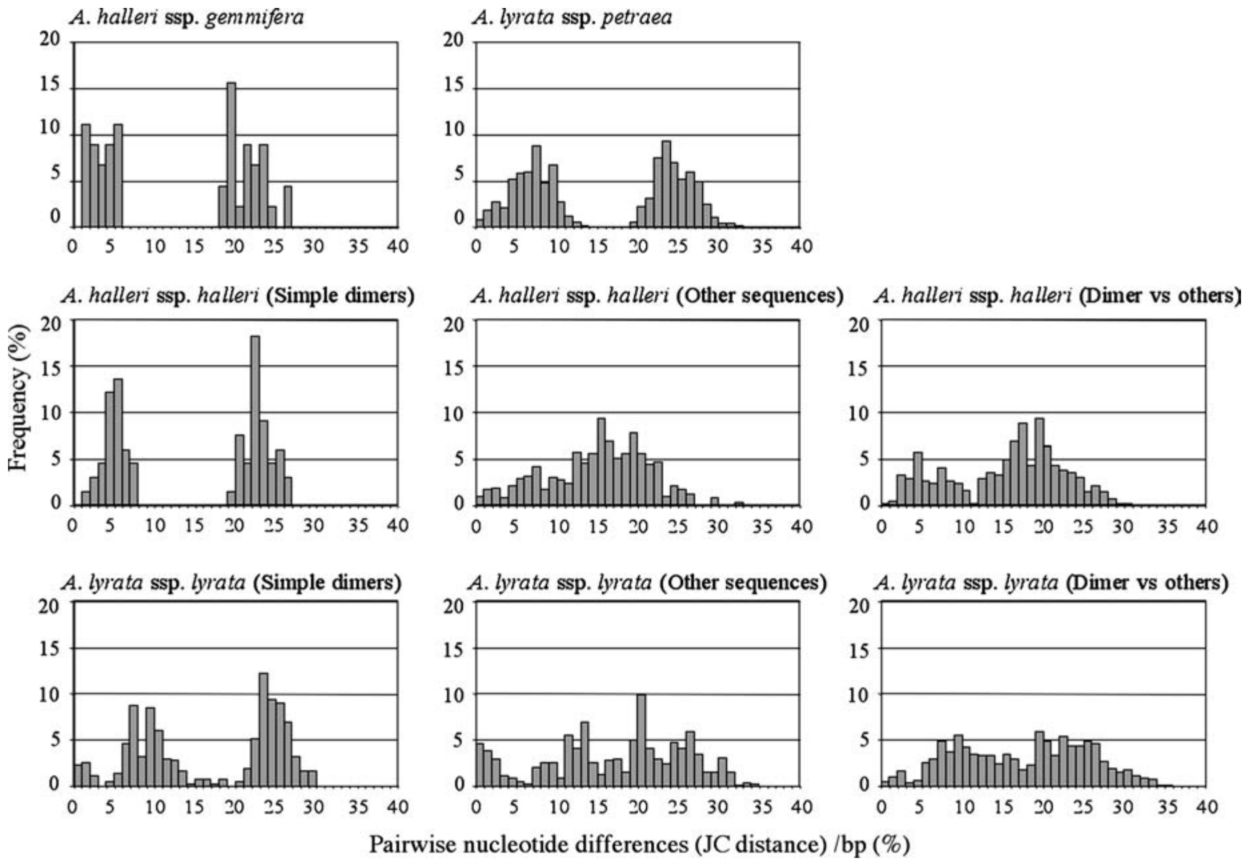

Fig. 3. Frequency distributions of pairwise nucleotide differences among pAge2 monomer units. Pairwise nucleotide differences per base pair with Jukes and Cantor correction were calculated in MEGA2 (Kumar et al. 2002) with the pairwise deletion option. Frequency distributions are shown with intervals of divergence of 1%.

The numbers of differences among units from odd-number repeat sequences, or between dimers and sequences with odd repeat numbers (Fig. 3), show no clear peaks and are generally smaller than between the first and the second units of simple dimer sequences, suggesting that these units differ less than do first and second subunits of dimers; these sequences thus have similarity to both first and second subunits of dimer sequences and can be classified as ‘‘complex,’’ following the ‘‘complex dimer’’ terminology defined in Fig. 1.

Among these sequences are some *A. lyrata* ssp. *lyrata* trimer sequences with chimeric structures that evidently derive from simple dimer units. Their sequences include long stretches similar to those of dimers from either *A. lyrata* ssp. *lyrata* or *petraea*. They can be classified into three groups, with different combinations of parts resembling portions of simple dimers and different junction regions (Supplemental Fig. 1). One type has high variation (7.60%) but this is based on only two sequenced clones. Within two trimer types, however, we find low variation (mean nucleotide diversity of 1.03%, based on 12 clones sequenced, and 2.31%, based on 11 clones). This is much less diversity than within simple pAge2 dimers (3%–7% depending on the species; see Table 1), and it suggests that these sequences rapidly and recently spread over the genome, particularly in *A. lyrata* ssp. *lyrata*.

Trimer sequences from *A. halleri* ssp. *halleri* are more complicated, although they cluster with other pAge2 sequences and differ only in the combinations of first and second units of simple dimers. The 11 trimer clones have some minor sequence differences at a few sites, and each clone has a different combination of variants characteristic of first and second units of simple dimer sequences.

Levels of Variation in Centromeric Satellite DNA Families

Including the sequences published previously, we analyzed a total of 308 pAa and 201 pAge1 monomer units (excluding 2 units with large deletions). For the pAa and pAge1 families, the average numbers of pairwise nucleotide differences per site are about 8%–10% for each family species-wide, within each population, or even within single plants (Table 1). For both these families, these values are similar to the mean raw nucleotide divergence values between sequences from different (sub)species (Dxy: Table 2).

Table 2. Levels of nucleotide differences in the centromeric satellite sequences between species: values are raw mean pairwise divergence (D_{xy})

Species compared		D_{xy}		
Species 1	Species 2	pAa	pAge1	pAge2 dimers
Arenosa vs other species				
Arenosa	Gemmifera	0.136	—	—
Arenosa	Halleri	0.120	—	—
Arenosa	Petraea	0.102	—	—
Arenosa	Lyrata	0.128	—	—
Average		0.122	—	—
Between *A. halleri* and *A. lyrata*				
Gemmifera	Petraea	0.093	0.129	0.066
Gemmifera	Lyrata	0.130	0.135	0.075
Halleri	Petraea	0.098	0.127	0.073
Halleri	Lyrata	0.112	0.134	0.081
Average		0.110	0.128	0.073
Between subspecies				
Gemmifera	Halleri	0.125	0.095	0.043
Petraea	Lyrata	0.101	0.087	0.083

Thus different units from the same species differ almost as much as units from different species. This indicates that the diversification of these sequences occurred before these species diverged from one another.

The pAge2 family has the highest level of variation of the three families when monomer units (357 units excluding 7 with large deletions or unsequenced regions) were used for the analyses (not shown), but this is due to differences between first and second monomer units of "simple" dimer sequences, not to differences between different sequences of such dimers. If we analyze simple pAge2 dimer sequences as dimers (112 sequences), the pAge2 family has an average of only 7.20% nucleotide differences, even including all species' sequences; within each species, these sequences have lower within-species diversity than either of the other two sequence families (Table 1). Divergence between species is also much less for pAge2 simple dimer sequences than for the pAa and pAge1 families (Table 2).

Given its longer divergence time from the other species, *A. arenosa* should have the most distinct sequences, while the two subspecies of either *A. halleri* or *A. lyrata* should differ the least. However, only the pAa family gives this result, having (slightly) greater divergence for *A. arenosa* than for the sequence comparisons involving the other species and subspecies, which all yield similar estimated values (Table 2). An absence of a relationship to the times when the taxa split suggest that this family is old, and has not recently increased in abundance in the *A. halleri* and *A. lyrata* genomes. The pAge1 family shows higher divergence between species than between subspecies (Table 2), but shows little differentiation between populations (Fig. 4), suggesting the absence of rapid homogenization between paralogous sequences within the species.

In the pAge2 family, however, sequences from some populations tend to cluster together, especially those from ssp. *petraea* (Fig. 4). The pAge2 sequences from the Plech and Karhumaki populations (both *A. lyrata* ssp. *petraea*) are similar, and both differ by less than 5% from the *A. halleri* sequences and, thus, cluster with them (Fig. 4). These sequences differ from those from the three other European populations (from Iceland, Sweden, and Wales), whose sequences are all similar and mostly differ by more than 5% from *A. halleri* sequences. The scattered phylogenetic positions of sequences of *A. lyrata* ssp. *lyrata* pAge2 from the Ontario population probably do not reflect this population's history but mainly reflect the existence of chimeric sequences ("complex" decayed dimers) and, perhaps, recombination between complex and simple sequences.

Sequence Exchanges in the Three Centromeric Satellite Families

The nucleotide variants within each satellite family suggest the action of crossing-over and/or gene conversion in several species, especially in the pAge2 family (Table 3). All datasets have at least one nucleotide pair for which all four combinations of variants were seen, indicating that some form of exchange has probably occurred, as already suggested above. The minimum estimated numbers of recombination events for the three families varied from 1 to 16 in the species studied. Although these estimates depend on the numbers of sequences in the dataset and numbers of parsimony informative sites, the different satellite families give similar results. It seems unlikely that these exchanges are due to reciprocal recombination, but gene conversion is possible.

For the pAge2 family, the presence of "decayed" trimer sequences suggests many sequence exchange events (most likely gene conversions), especially in the *A. halleri* ssp. *halleri* and *A. lyrata* ssp. *lyrata* dataset. The sequence combinations in *A. lyrata* ssp. *lyrata* and *A. halleri* ssp. *halleri* pAge2 trimer units suggest that exchange events initially occurred between simple dimer sequences (Supplemental Fig. 1). However, fixed variants are found between the dimer and the trimer sequences, even within the same subspecies, suggesting that sequence exchange between the different categories of sequences may be limited (although the small number of sequences analyzed overestimates the number of fixed differences). The trimer-specific variants are often shared between different subunits (i.e., not confined to the first or any other specific subunit of trimers); this suggests se-

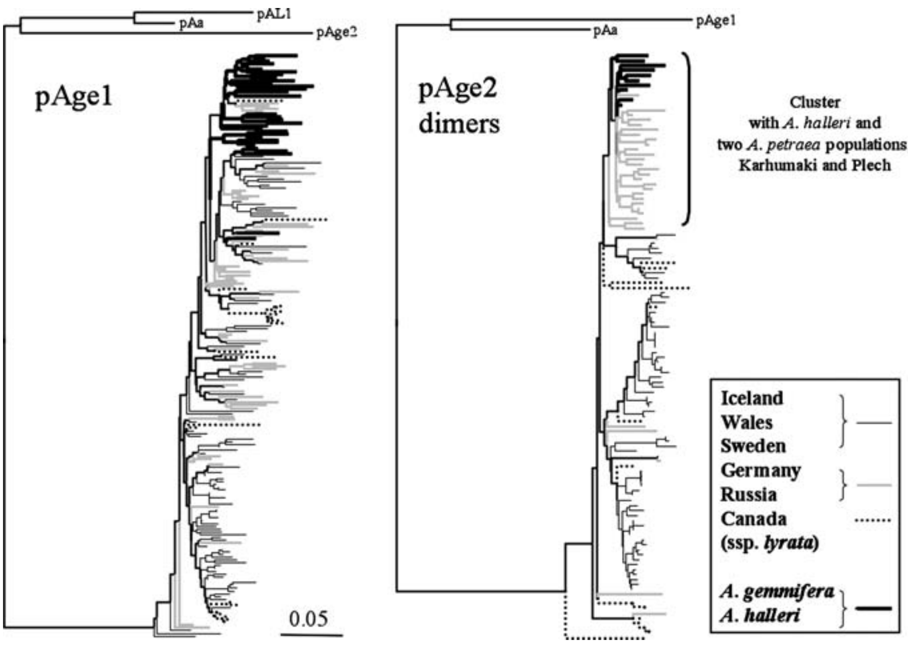

Fig. 4. Phylogenetic relationships of the pAge1 and pAge2 families. pAge1 monomer (left) and pAge2 dimer (right) sequences were used for tree constructions. The figure shows NJ trees using Jukes and Cantor distances. Branches of units from different species or populations are distinguished by thick black lines (*A. halleri* subspecies), gray shaded lines (Plech from Germany and Karhumaki from Russia), thin black lines (Esja Mt. from Iceland, Stubbsand from Sweden, and Clogwyn D'ur Arddu, Wales), and dotted lines (Ontario, Canada). Consensus sequences from other families are also included as outgroups. A distance bar is shown at the bottom. A version with each population in a different color is supplied as Supplementary Fig. 2.

quence exchanges also between different trimer sequence types, and even between subunits of a single trimer sequence. Although we have not analyzed long stretches of pAge2 units and do not know how the dimeric and other types of units are arranged in the genome, these results suggest that the different types of pAge2 units may represent an early stage in the differentiation of a new satellite sequence family.

Discussion

Multifamily Structure of Centromere Satellite Families in A. halleri *and* lyrata *Subspecies*

All *A. halleri* and *lyrata* subspecies have three centromeric satellite families, indicating that the time required for fixation of a major centromere satellite family is not shorter than the history of *A. halleri* and *lyrata*, which diverged from one another recently (Ramos-Onsins et al. 2004), certainly much more recently than the divergence of their common ancestor from *A. thaliana* (estimated, based on a molecular clock, to be ~5 MYA [Koch et al. 2000]).

Fixation of a new satellite sequence in a species requires that a variant must arise in a single unit and then spread throughout its local array on its chromosome of origin and then to all arrays across all chromosomes in the species. Fixation times will therefore be very long unless selection favors variants (Ohta and Dover 1984; Charlesworth et al. 1994). Theoretical studies (reviewed by Charlesworth et al. 1994) show that fixation times of new sequences in multicopy sequence arrays depend on unequal crossover frequencies, gene conversion frequencies, copy numbers, and bias in the gene conversion process (Ohta and Dover 1984; Stephan 1989). Due to the huge numbers of satellite repeats in the *A. thaliana* centromere, even the first stage of spread within a chromosome will require a long time in the absence of selection. In *A. thaliana*, the centromere regions include about 2 Mbp consisting of repeats of the 180-bp pAL1 satellite family (Haupt et al. 2001; Hosouchi et al. 2002). There are thus about 10,000 copies in each chromosome, and this is likely to apply also to *A. halleri* and *A. lyrata*, whose genomes are larger than that of *A. thaliana* (Johnston et al. 2005), although recent comparative sequence analyses of pericentromeric region showed a much larger pericentromeric intergenic region in *A. thaliana* than related species (A.E. Hall et al. 2006). The species studied here were not studied by A.E. Hall et al. (2006), but it is likely that similar results would be found, since it appears clear that the exceptional species is *A. thaliana*.

The second stage, spread to centromeres of other chromosomes, requires interchromosomal exchanges between units, which are probably rare events (Morgante et al. 1997; Heslop-Harrison et al. 1999; Schindelhauer and Schwarz 2002), although the presence of the same satellite sequence families in the centromeres of different chromosomes indicates some interchromosomal movement for the satellite families studied here (Kawabe and Nasuda 2006). Recombination only slightly increases fixation times, predominantly affecting the variability among members of the multigene family. Since centromere regions have low rates of reciprocal recombination, the largest influences on fixation times of the variants studied here should come from the gene conversion rate and amount of bias in the conversion process,

Table 3. Estimated numbers of recombination and gene conversion events

Satellite sequence family	Species or subspecies	No. of units	No. of parsimony informative sites	Minimum no. of recombination events	No. of gene conversion events
pAa	Gemmifera	6	8	3	0
	Halleri	15	31	6	2
	Petraea	10	11	4	0
	Lyrata	15	31	3	0
	Arenosa	262	108	9	0
pAge1	Gemmifera	17	24	5	2
	Halleri	13	14	6	0
	Petraea	149	85	11	3
	Lyrata	22	25	5	1
pAge2	Gemmifera	10	33	1	0
	Halleri	44	69	9	4
	Petraea	182	100	16	3
	Lyrata	121	93	6	264
pAge2 dimers	Gemmifera	4	3	1	0
	Halleri	6	5	1	0
	Petraea	89	117	16	7
	Lyrata	13	45	5	0

which has effects similar to natural selection (Ohta and Dover 1984).

Assuming that crossing-over does not occur in the centromere region, the expected time to fixation under neutrality can be roughly estimated simply from the effective population size, the repeat number, and the gene conversion rate per repeat unit (Nagylaki and Petes 1982; Ohta 1983). This is of interest, because it can allow one to infer whether selection has been driving replacements of satellite sequences by variant sequences. Using equation (8) of Ohta (1983), the predicted fixation time within one chromosome, without selection or biased gene conversion, is roughly given by the copy number divided by the gene conversion rate (if this value is lower than N_e generations) or by $4N_e$ (if the copy number/gene conversion rate ratio exceeds N_e generations). N_e can be estimated from silent site diversity, which is between 1% and 2% for this species (Wright et al. 2003; Ramos-Onsins et al. 2004); with plausible mutation rates per base pair of 1.5×10^{-8} and 6.5×10^{-9} (Wright et al. 2003), we obtain N_e values up to 10^6. The divergence time between *A. thaliana* and *A. halleri* and *A. lyrata*, whose satellite families are different, is estimated to be about 5 MYA (Koch et al. 2000), which is about 10^6 generations, assuming one generation per year. Assuming about 10^4 copies of the satellite sequences in each chromosome, the gene conversion rate must therefore exceed 10^{-2} for fixation of different satellite families in these species under genetic drift alone. Gene conversion rates have not been estimated for centromere regions in these species but, for noncentromeric regions, are probably as low as the mutation rate, 10^{-8} to 10^{-9} per site per generation (Innan 2004). Using this value (i.e., assuming that the rate is not much higher in centromere regions), the fixation of a new satellite sequence within one chromosome is estimated to require about 10^{10} generations, much larger than the species' divergence time. Thus, either selection or biased gene conversion is necessary to explain complete substitution of satellite sequences in the available time.

However, pairing of nonhomologous regions or the action of mobile elements could increase sequence exchange rates in centromere regions, making fixation possible in a shorter time than estimated above. Also, in centromere regions, other processes, including sister chromatid exchanges and nonhomologous end joining between physically distant arrays, or even between nonhomologous chromosomes, could cause exchanges similar to those due to gene conversion events. Physical clustering of centromeres in nuclei or ectopic exchange between tranposable elements could potentially cause sequence exchanges between chromosomes. The effects of these processes are not treated separately in the models on which our calculation is based, but they are equivalent to an increased gene conversion rate and will, thus, reduce the fixation time estimated above (Nagylaki and Petes 1982; Ohta 1983). However, this possibility seems implausible, because the gene conversion rate in the centromere regions required for replacement of the satellite sequences of an *Arabidopsis* centromere under neutrality in the available time would have to be extremely high (10^{-2} per repeat). Moreover, as just explained, the fixation time of repetitive sequences dispersed among two or more chromosomes will be even longer than that in a single chromosome (Ohta and Dover 1983). All three satellite families are the sole sequences in at least one chromosome of the species studied here (Kawabe and Nasuda 2005). Thus centromere satellite sequences may have been replaced too rapidly to have occurred by genetic drift alone. A biased increase in certain satellite variants is thus necessary to explain the homogenization of sequences throughout the whole *A. thaliana* genome.

This analysis also suggests that the shared satellite families between *A. halleri* and *A. lyrata* probably represent a transient stage. Several other species, including some plants (Harrison and Helop-Harrison 1995; Ananiev et al. 1998; Gindullis et al. 2001), have multifamily centromere satellite sequences, often with some sequences specific to certain chromosomes, as observed in *A. halleri* and *A. lyrata*. As in the genus *Arabidopsis*, relatives of such species usually have just a single, completely different, satellite sequence in all their chromosomes. Coexistence of two or more centromere satellite families within species thus tends to last for evolutionary times shorter than the divergence times of the species studied, consistent with our finding that shared variants are found only in closely related *Arabidopsis* species. Most cases of satellite families shared between species are in recently diverged animal or plant species (e.g., Mestrovic et al. 1998; Nijman and Lenstra 2001). Cases involving species with estimated divergence times of more than 10 MYA have been reported, but centromere functions of the sequences have not been verified (Vershinin et al. 1996; Robles et al. 2004).

Species and Population Specificity of Different Satellite Sequences

The different levels of similarity of satellite sequences within and between species suggest occasional species-specific amplification of variants. Among the three satellite families of *A. halleri* and *A. lyrata*, pAge2 sequences have the greatest species and population specificity, while pAa shows almost no species-specific sequence variants. These differences are partly related to the history of the different satellite families. *A. arenosa*, a close relative of *A. halleri* and *A. lyrata*, has only the pAa family, suggesting that this sequence type evolved in a common ancestor, long enough ago to spread throughout the chromosomes of this species, and is thus probably the oldest of the three families present in *A. halleri* and *A. lyrata*. This would imply that pAge1 evolved from a centromere region initially dominated by pAa, in an ancestor of *A. halleri* and *A. lyrata*, and that pAge2 emerged and increased subsequently. A recent origin of the pAge2 family is supported by the fact that it shows the lowest levels of PCR amplification (suggesting a low copy number) and the clearest band patterns (Fig. 2). The clear pAge2 bands indicate that length variation is low among sequences in this family, consistent with our sequence analyses and suggesting a recent common ancestor; in contrast, smeared patterns for pAa and pAge1 suggest repeat units of highly variable lengths, i.e., greater age.

The low diversity among *A. lyrata* ssp. *lyrata* pAge2 trimers (with two groups of sequences that are both almost-monomorphic combinations of simple pAge2 dimer units) suggests selectively favored variants that are increasing in frequency too rapidly for exchange events to generate variation among sequence copies.

Restriction of Sequence Exchanges to Within Families

If the abundance of one complex dimer sequence variant increases at a given centromere, for any reason, the region involved (presumably initially just one centromere) will initially contain mixtures of sequences with different structures.

Despite the evidence of recombination (probably due to gene conversion), especially for the pAge2 family, we found no clones, even those containing trimer or tetramer sequences, that include partial or complete unit sequences of the other satellite families. If exchanges occur between different chromosomes or distantly localized satellite units, chimeric sequences of two or more families might be observed if larger numbers of sequences were analyzed; our current dataset only includes a small fraction of all satellites. The failure to detect any such chimeric structures suggests that the mechanisms causing exchanges of centromeric satellite sequences are restricted to physically close locations. Only one chromosome of *A. halleri* ssp. *gemmifera* has two different satellite families in a single centromere, and so physical separation may explain restriction of exchanges to similar sequence units, consistent with the results suggesting that exchange events between distantly located satellite units are generally infrequent in *Arabidopsis* species (S.E. Hall et al. 2005).

Exchanges may also be limited by sequence divergence. In many organisms, including plant species, recombination occurs only between similar sequences (Dooner and Martinez-Ferez 1997; Dooner 2004). The *A. halleri* and *lyrata* centromeric satellite families' sequences may mostly be different enough (about 30% nucleotide divergence) to restrict between-family exchange events. However, some exchanges are detected between different pAge2 dimer units, with divergence of about 20%.

Among the *A. halleri* ssp. *halleri* pAge2 sequences, many types of trimers were found, and almost every clone is unique (contrasting with the pAge2 results from *A. lyrata* ssp. *lyrata*). Double or multiple events have occurred in short stretches of sequences in the centromere regions containing pAge2, which cannot be explained simply by recombination between units.

We detected only a few gene conversion events other than those just described in *A. halleri* ssp. *halleri* and *A. lyrata* ssp. *lyrata* pAge2. The high variability and short length of unit sequences of pAa and pAge1 make it difficult to detect gene conversion within these families, but the generally conserved dimer structure of pAge2 sequences also suggests limitation of sequence exchanges, even in this satellite family.

Observations of extant sequence variation, or even direct experiments monitoring mutation processes, cannot allow us to fully determine the evolutionary processes that destroy strict dimer units and generate other more complex structures. Nevertheless, the presence of dimer sequences must lead to generation of new satellite variants, allowing for coevolution with centromere proteins. The system in these *Arabidopsis* species may therefore provide opportunities in the future for testing whether selective events have indeed driven the evolutionary changes observed in their centromere satellite sequences.

Acknowledgments. We thank S. Preuss for technical assistance and the Institute of Evolutionary Biology, Edinburgh University, sequencing service for sequencing.

References

Ananiev EV, Phillips RL, Rines HW (1998) Chromosome-specific molecular organization of maize (Zea Mays L.) centromeric regions. Proc Natl Acad Sci USA 95:13073–13078

Cabot EL, Doshi P, Wu ML, Wu CI (1993) Population genetics of tandem repeats in centromeric heterochromatin: unequal crossing over and chromosomal divergence at the *Responder* locus of *Drosophila melanogaster*. Genetics 135:477–487

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215–220

Choo KHA (1997) The centromere. Oxford University Press, New York

Choo KHA (1998) Why is the centromere so cold? Genome Res 8:81–82

Copenhaver GP, Browne WE, Preuss D (1998) Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads. Proc Natl Acad Sci USA 95:247–252

Csink AK, Henikoff S (1998) Something from nothing: the evolution and utility of satellite repeat. Trends Genet 14:200–204

Dooner HK (2004) Extensive interallelic polymorphisms drive meiotic recombination into a crossover pathway. Plant Cell 14:1173–1183

Dooner HK, Martinez-Férez IM (1997) Recombination occurs uniformly in the *bronze* gene, a recombination hotspot in the maize genome. Plant Cell 9:1633–1646

Gindullis F, Desel C, Galasso I, Schmidt T (2001) The large-pscale organization of the centromeric region in Beta species. Genome Res 11:253–265

Hall AE, Keith KC, Hall SE, Copenhaver GP, Preuss D (2004) The rapidly evolving field of plant centromeres. Curr Opin Plant Biol 7:108–114

Hall AE, Kettler GC, Preuss D (2006) Dynamic evolution at pericentromeres. Genome Res 16:355–364

Hall SE, Luo S, Hall AE, Preuss D (2005) Differential rates of local and global homogenization in centromere satellites from *Arabidopsis* relatives. Genetics 170:1913–1927

Harrison GE, Heslop-Harrison JS (1995) Centromeric repetitive DNA in the genus *Brassica*. Theor Appl Genet 90:157–165

Haupt W, Fischer TC, Winderl S, Fransx P, Torres-Ruiz A (2001) The CENTROMERE1 (CEN1) region of Arabidopsis thaliana: architecture and functional impact of chromatin. Plant J 27:285–296

Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293:1098–1102

Heslop-Harrison JS, Murata M, Ogura Y, Schwarzacher T, Motoyoshi F (1999) Polymorphism and genomic organization of repetitive DNA from centromeric regions of *Arabidopsis thaliana*. Plant Cell 11:31–42

Heslop-Harrison JS, Brandes A, Schwarzacher T (2003) Tandemly repeated DNA sequences and centromeric chromosomal regions of *Arabidopsis* species. Chromosome Res 11:241–253

Hosouchi T, Kumekawa N, Tsuruoka H, Kotani H (2002) Physical map-based sized of the centromeric regions of Arabidopsis thaliana chromosomes 1, 2, and 3. DNA Res 9:117–121

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

Innan H (2004) Theories for analyzing polymorphism data in duplicated genes. Gen Genet Sys 79:65–75

Jensen MA, Charlesworth B, Kreitman M (2002) Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. Genetics 160:493–507

Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, Lopez R, Price HJ (2005) Evolution of genome size in Brassicaceae. Ann Bot 95:229–235

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

Kawabe A, Nasuda S (2005) Structure and genomic organization of centromeric repeat in Arabidopsis species. Mol Genet Genomics 272:593–602

Kawabe A, Nasuda S (2006) Polymorphic chromosomal specificity of centromere satellite families in *Arabidopsis halleri* ssp. *gemmifera*. Genetica 126:335–342

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

Kumar S, Tamura K, Jacobsen I, Nei M (2000) Molecular Evolutionary Genetics Analysis, version 2.0. Pennsylvania and Arizona State universities, University Park and Tempe

Lo AW, Magliano DJ, Sibson P, Kalitsis P, Craig JM, Choo KHA (2001) A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA. Genome Res 11:448–457

Maluszynska J, Heslop-Harrison JS (1991) Localization of tandemly repeated DNA sequences in *Arabidopsis thaliana*. Plant J 1:159–166

Martinez-Zapater J, Estelle MA, Somerville CR (1986) A highly repeated DNA sequence in *Arabidopsis thaliana*. Mol Gen Genet 204:417–423

Mestrovic N, Plohl M, Mravinac B, Ugarkovic D (1998) Evolution of satellite DNAs from the genus Palorus—experimental evidence for the "library" hypothesis. Mol Biol Evol 15:1062–1068

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

Morgante M, Jurman I, Shi L, Zhu T, Keim P, Rafalski JA (1997) The STR120 satellite DNA of soybean: organization, evolution and chromosomal specificity. Chromosome Res 5:363–373

Murata M, Ogura Y, Motoyoshi F (1994) Centromeric repetitive sequence in *Arabidopsis thaliana*. Jpn J Genet 69:361–370

Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang J (2004) Sequencing of a rice centromere uncovers active genes. Nat Genet 36:138–145

Nagylaki T, Petes TD (1982) Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics 100:315–337

Nijman JJ, Lenstra JA (2001) Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeat. J Mol Evol 52:361–371

Ohta T (1983) Time until fixation of a mutant belonging to a multigene family. Genet Res 41:47–55

Ohta T, Dover GA (1983) Population genetics of multigene families that are dispersed into two or more chromosomes. Proc Natl Acad Sci USA 80:4079–4083

Ohta T, Dover GA (1984) The cohesive population genetics of molecular drive. Genetics 108:501–521

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

Robles F, De La Herran R, Ludwig A, Rejon CR, Rejon MR, Garrido-Ramos MA (2004) Evolution of ancient satellite DNA in sturgeon genomes. Gene 338:133–142

Rozas J, Rozas R (1999) *DnaSP* version 3: a integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Saffery R, Sumer H, Hassan S, Wong LH, Craig JM, Todokoro K, Anderson M, Srafford A, Choo KHA (2003) Transcription within a functional human centromere. Mol Cell 12:509–516

Sawyer SA (1999) GENECONV: a computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University, St. Louis; available at: http://www.math.wustl.edu/~sawyer

Schindelhauer D, Schwarz T (2002) Evidence for a fast intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous α-satellite DNA array. Genome Res 12:1815–1826

Stephan W (1989) Tandem-repetitive noncoding DNA: forms and forces. Mol Biol Evol 6:198–212

Vershinin AV, Alkhimova EG, Heslop-Harrison JS (1996) Molecular diversification of tandemly organized DNA sequences and heterochromatic chromosome regions in some Triticeae species. Chromosome Res 4:517–525

Wright SI, Lauga B, Charlesworth D (2003) Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol Ecol 12:1247–1263