An Analysis of Microsatellite Loci in *Arabidopsis thaliana*: Mutational Dynamics and Application

V. Vaughan Symonds and Alan M. Lloyd[1]

Section of Molecular, Cell, and Developmental Biology and Institute for Cellular and Molecular Biology, University of Texas, Austin, Texas 78712

Manuscript received November 1, 2002
Accepted for publication June 23, 2003

ABSTRACT

Microsatellite loci are among the most commonly used molecular markers. These loci typically exhibit variation for allele frequency distribution within a species. However, the factors contributing to this variation are not well understood. To expand on the current knowledge of microsatellite evolution, 20 microsatellite loci were examined for 126 accessions of the flowering plant, *Arabidopsis thaliana*. Substantial variability in mutation pattern among loci was found, most of which cannot be explained by the assumptions of the traditional stepwise mutation model or infinite alleles model. Here it is shown that the degree of locus diversity is strongly correlated with the number of contiguous repeats, more so than with the total number of repeats. These findings support a strong role for repeat disruptions in stabilizing microsatellite loci by reducing the substrate for polymerase slippage and recombination. Results of cluster analyses are also presented, demonstrating the potential of microsatellite loci for resolving relationships among accessions of *A. thaliana*.

MICROSATELLITE loci are tandemly repeated DNA motifs of 1–6 bp in length; they are also referred to as simple sequence length polymorphisms (SSLPs), simple sequence repeats, simple tandem repeats, and variable number tandem repeats (VNTRs). These loci occur at high frequency in all eukaryotes examined (Katti *et al.* 2001) and at some lower frequency in prokaryotic genomes (Metzgar *et al.* 2001). The use of microsatellite loci as polymorphic DNA markers has expanded considerably over the past decade both in the number of studies (Estoup and Angers 1998) and in the number of organisms (Barker 2002), primarily due to their facility and power for population genetic analyses. Microsatellite loci are typically highly variable, even in organisms that otherwise display little genetic variation (Zwettler *et al.* 2002), are relatively straightforward to identify (Zane *et al.* 2002), and can be scored via many different methods. Although originally described from humans for use in genetic fingerprinting (Litt and Luty 1989), microsatellite locus use today includes genetic mapping (*e.g.*, McCouch *et al.* 1997; Sakamoto and Okamoto 2000), assessments of genetic diversity (Cruzan 1998; Driscoll *et al.* 2002), forensics (Gill *et al.* 1985; Kubo *et al.* 2002) and studies of human genetic disease proliferation (Cunniff 2001; Ranum and Day 2002).

Microsatellite loci increase and decrease in length due to polymerase slippage during DNA replication (Eckert *et al.* 2002) and recombination (Richard and Paques 2000), both of which are consequences of having a series of identical tandemly repeated units. With these phenomena in mind, discussions of microsatellite evolution primarily center around two models, the stepwise mutation model (SMM) and the infinite alleles model (IAM; Balloux and Lugon-Moulin 2002). In short, the SMM suggests that the mutation of microsatellite alleles occurs by the loss or gain of a single tandem repeat, and the IAM describes mutations involving the loss or gain of any number of repeats, but always generates new, previously unsampled alleles (see review by Estoup and Cornuet 1999). One commonality between these two models is that they consider only changes in tandem repeat number. More recently it has been suggested that microsatellite locus evolution is most strongly influenced by the balance between locus length and point mutation rate (Kruglyak *et al.* 1998). Specifically, longer microsatellite alleles are hypothesized to be more prone to generate new length variants than are shorter alleles (Wierdl *et al.* 1997). However, nonrepeat mutations (substitutions, insertions, and deletions) that interrupt perfect tandem repeats affect the function of length. The disruption of a set of tandem repeats by any process, including indels and point mutations, in effect lessens the number of perfectly repeated units and is expected to reduce the likelihood of locus evolution (Rolfsmeier and Lahue 2000). Despite advances in development of molecular evolution models and the widespread use of microsatellite markers, detailed analy-

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY295838–AY295871 and AY293992–AY294004.

[1]*Corresponding author:* Section of Molecular, Cell, and Developmental Biology, MBB 1.448b, 2500 Speedway, University of Texas, Austin, TX 78712. E-mail: lloyd@uts.cc.utexas.edu

sis of microsatellite evolution and the underlying forces remains limited to relatively few studies representing even fewer organisms (for examples, see Noor *et al.* 2001; Vigouroux *et al.* 2002).

Arabidopsis thaliana has long been a model genetic and molecular system for plant biology. Recently, natural variation within this species has come into focus (Alonso-Blanco and Koornneef 2000), expanding its utility toward addressing evolutionary and population biology questions. Unfortunately, the genetic infrastructure, including mapping data, in place for the few most commonly used accessions of *A. thaliana*, does not yet extend to the several hundred wild-collected accessions available. An examination of microsatellite variation within *A. thaliana*, therefore, serves at least two purposes: improving upon the genetic tools available for this model organism and expanding our knowledge of microsatellite evolution.

Previous studies on *A. thaliana* microsatellite loci have shown that they are abundant (Casacuberta *et al.* 2000; Katti *et al.* 2001) and highly variable (Innan *et al.* 1997; van Treuren *et al.* 1997; Clauss *et al.* 2002). However, these works are limited to <50 accessions and the studies minimally overlap in marker usage. To develop the utility of microsatellite loci among wild accessions and to investigate factors affecting mutation patterns at these loci, we have gathered size and sequence data for a diverse collection of *A. thaliana* accessions. We find substantial variability in mutation pattern among microsatellite loci and among accessions, most of which specifically conforms to neither the SMM nor the IAM. Contributing to this variation is sequence complexity and the presence of repeat disruptions within loci. Here we show that high-diversity loci tend to possess long stretches of contiguous repeats, while low-diversity loci either are uninterrupted with few total repeats or contain repeat interruptions that result in few contiguous repeats. Further, sequence data indicate that there is a wealth of intraspecific, potentially phylogenetically informative variation at these loci, an important point in a model system for which we possess little genealogical information.

MATERIALS AND METHODS

Plant materials: Genetic variation among 120 "wild" accessions and several commonly used reference accessions (including Col-0, L*er*, and WS) of *A. thaliana* was surveyed (Table 1). Line selection was based on global population coverage and, for a subset of lines, local proximity. That is, a few nested accessions including separate collections made from near the same location were selected. Although microsatellite size data exist for the three reference accessions, different size scoring methods tend to yield varying results (our personal observation). Therefore, the reference accessions were included in our analyses to derive data directly comparable with all other accessions included. Two stocks, Cal-0 and Tac-0, were generously provided by Johanna Schmitt and Lisa Dorn. Three of the reference accessions used were lab stocks. All remaining seed stocks were acquired from the *Arabidopsis* Biological Resource Center. Although all accessions of *A. thaliana* are reportedly nearly completely homozygous (Bergelson *et al.* 1998), all accessions included here underwent at least one round of additional selfing in our lab prior to genotyping. Seed for all lines were imbibed in water and vernalized at 4° for 3 days prior to germination at 22° under 24 hr light.

Microsatellite survey: Total DNAs were extracted from several rosette leaves of a single individual for each accession following a modified CTAB method (modified from Doyle and Doyle 1987). Approximately 50 ng of total DNA was used as template in individual microsatellite amplification reactions.

All lines were screened at 20 microsatellite marker loci. These loci were selected to provide approximately equal coverage across the genome at a density equivalent to that required for rough-scale mapping (approximately every 30 cM), taking into consideration both the distance between pairs of markers and the distance between centromere and chromosome end positions (Table 2). Primer sequences for all loci, which were originally described by Bell and Ecker (1994) and Lukowitz *et al.* (2000), were acquired through the *Arabidopsis* Information Resource (http://www.arabidopsis.org). Each locus was amplified by PCR and fluorescently labeled by one of two methods: either the forward primer in each reaction was labeled directly with one of the three dyes (D2, D3, and D4) used on Beckman-Coulter instruments or an M13 tailing scheme was followed as described by Boutin-Ganache *et al.* (2001), whereby the forward primer was 5′-tailed with the M13 forward sequence and used in conjunction with a 15-fold excess of a fluorescently labeled M13 forward primer. All primers used in amplification reactions were synthesized by ResGen (Invitrogen, San Diego). The switch was made to the M13 tailing scheme because it requires only three fluorescently labeled primers, rather than independently labeling all forward primers.

Amplification reactions were carried out in 10-μl volumes containing 1× PCR buffer (Invitrogen), 1.5 mm $MgCl_2$, 50 μm each dNTP, and either 600 nm each primer (for reactions with labeled forward primer) or 150 nm labeled M13 and reverse primers and 10 nm unlabeled forward primer (for M13 tailed primer scheme). Approximately 50 ng of each DNA extraction was used as template for individual locus amplification in a standard 96-well plate format. Standard amplification conditions consisted of 95° for 3 min, 30 cycles of denaturing at 94° for 1 min, annealing at 55° for 1 min, and polymerization at 72° for 1 min, followed by a final extension for 6 min at 72°. As the annealing temperature for the M13 primers is lower than that of the average SSLP primer, amplification conditions for the M13 scheme were modified by lowering the annealing temperature to 52°. Although two amplification schemes were used, the amplification conditions for each locus were consistent among all accession templates amplified and no significant difference in amplification rate was observed between the two protocols.

Microsatellite length polymorphisms were detected and scored by capillary electrophoresis on a Beckman-Coulter CEQ 2000XL DNA analyzer. Although all amplification reactions were carried out individually, the use of three different dyes allowed for the pool-plexing of samples during separation and allele sizing. Typically, the PCR products of three separate reactions for one individual, each labeled with a different dye (D2, D3, and D4) were pooled. The pooled products were then purified in vacuum filter plates (Millipore MANU030) at 20 in. Hg for 4 min (manufacturer's specifications) and subsequently eluted in 30 μl H_2O. A total of 1.25 μl of each cleaned, pooled sample was then added to 0.5 μl of 400-bp size standard (labeled with D1 dye) and 38 μl of sample loading

TABLE 1

List of the 126 accessions used in this study

Stock no.	Background	Stock no.	Background	Stock no.	Background
1394	No-0	6798	Mir-0	6843	Pt-0
1516	Sf-2	6799	Mt-0	6844	Ra-0
3081	No-0	6800	Mz-0	6845	Rd-0
6173	Est	6801	Na-1	6848	Rsch-0
6175	Condara	6803	Nd-0	6850	Rsch-4
6600	Aa-0	6804	Nie-0	6851	Ru-0
6601	Ag-0	6805	No-0	6852	Se-0
6607	Ba-1	6806	Np-0	6853	Sei-0
6608	Bay-0	6807	Nok-0	6854	Sap-0
6615	Bl-1	6808	Nok-1	6855	Sf-1
6616	Bla-1	6809	Nok-2	6856	Sav-0
6624	Bla-12	6810	Nok-3	6857	Sf-2
6625	Bla-14	6811	Nw-0	6859	Sg-2
6626	Br-0	6812	Nw-1	6860	Sh-0
6627	Bs-1	6813	Nw-2	6861	Si-0
6643	Bir-0	6814	Nw-3	6862	Sp-0
6659	Cal-0	6815	Nw-4	6863	St-0
6660	Can-0	6816	Ob-0	6864	Ste-0
6664	Chi-0	6818	Ob-2	6865	Stw-0
6665	Chi-1	6819	Ob-3	6867	Ta-0
6666	Chi-2	6820	Old-1	6868	Ts-1
6669	Co-1	6821	Old-2	6869	Ts-2
6672	Co-4	6822	Or-0	6870	Ts-3
6675	Cvi-0	6823	Ove-0	6871	Ts-5
6685	Dra-0	6824	Oy-0	6872	Ts-6
6686	Dra-1	6825	Pa-1	6873	Ts-7
6700	Est-0	6826	Pa-2	6874	Tsu-0
6701	Est-1	6827	Pa-3	6878	Ty-0
6702	Et-0	6828	Per-1	6879	Uk-1
6716	Gd-1	6829	Per-2	6880	Uk-3
6723	Gr-1	6830	Per-3	6884	Van-0
6734	Hau-0	6831	Pf-0	6885	Wa-1
6736	Hi-0	6832	Pi-0	6889	Wil-2
6745	Jl-3	6833	Pi-2	6891	Ws-0
6751	Kas-1	6834	Pla-0	6897	Wu-0
6754	Kil-0	6835	Pla-1	6898	X-0
6762	Kn-0	6836	Pla-3	6899	XX-0
6765	La-0	6838	Pn-0	Cal-0	Calvert
6769	Lc-0	6839	Po-0	Col	Col
6783	Ll-2	6840	Po-1	L*er*	L*er*
6784	Lm-2	6841	Pr-0	Tac-0	Tacoma
6789	Ma-0	6842	Pog-0	WS	WS

solution (Beckman-Coulter) in a well of a 96-well sample plate and overlaid with mineral oil. Each pool-plexed sample was separated on the CEQ using the standard Frag-1 method. This pool-plexing system resulted in the separation of products at three different loci simultaneously through a single capillary along with an internal size standard. Fragments were sized using the default fragment analysis protocol for the appropriate set of dyes used (AE2 or PA1 options).

Microsatellite data analyses: The CEQ raw data from each run were analyzed using the appropriate dye mobility calibration settings for each dye and the default fragment analysis settings for the 400-bp size standard. Alleles reported here reflect the amplification product size, as scored on a CEQ 2000XL DNA analyzer. Simply inferring the number of repeats from size data ignores potentially informative data from indels. Often alleles are sized on the basis of assumptions regarding the locus; for example, alleles at a dinucleotide repeat locus are often assumed to fall only into size classes 2 bp apart. However, our sequence data show real indels and real 1-bp differences among alleles at dinucleotide repeat loci in our data set. Therefore, we report all observed size classes, regardless of the repeat type.

Microsatellite cloning and sequencing: To investigate the nature of length variations within loci, several alleles were cloned and sequenced for six loci. Three loci were randomly selected from among the low-diversity loci (nga1107, nga1145, and nga129) and three from among the high-diversity loci (CIW7, nga172, and nga8). Individual alleles were amplified with unlabeled (no dye) forward and reverse primers as described in the preceding section from individual accessions. One microliter of PCR product was then added to a cloning reaction using the TOPO-TA cloning kit (Invitrogen). Colo-

nies with inserts were initially identified by blue/white screening, followed by PCR amplification from individual colonies and size confirmation on agarose gels. Multiple clones for each reaction were identified and plasmid DNA minipreparations were prepared from selective overnight liquid cultures. DNA minipreparations were carried out following a modified SDS protocol where DNA precipitation is preceded by separate phenol and chloroform extractions. Approximately 500 ng of vector with insert were used as template in sequencing reactions using either the T7 or the M13 reverse primer. Sequencing reactions were purified using Sephadex G-50 columns and the sequences were analyzed on an MJ Research (Watertown, MA) BaseStation DNA analyzer. Postrun data were processed using the Cartographer v. 1.2.4sg software (MJ Research). Sequence alignments for alleles of each locus were carried out using Megalign (DNASTAR, Madison, WI).

Associations between locus length and locus diversity: Associations between the genetic diversity of a locus and some measure of locus length, typically mean length, are commonly reported for microsatellite loci (Bachtrog *et al.* 2000; Moriguchi *et al.* 2003). To investigate this association for the loci examined here, the mean allele size was determined for each locus. That allele or the nearest in size was cloned and sequenced from multiple accessions for 10 loci, as described above. From these sequences and available Col sequence, the total number of repeats was counted or inferred for each locus by subtracting the shared, nonrepeat flanking sequence from the total locus length. Association strength between repeat number and locus diversity was assessed by calculating the correlation coefficients between the two; both Pearson product moment correlation and Spearman's coefficient of rank correlation were calculated. To examine the potential role of repeat interruptions on locus diversity, from those same sequences the largest number of contiguous repeats was counted. For example, in the following sequence, ACTGAGAGATTGAGAGAGACTT, the total number of repeats is seven, and the largest number of contiguous repeats is four. Again, association strength was determined by calculating correlation coefficients between the largest number of contiguous repeats and locus diversity. Because different repeat types often have different mutation rates (Bachtrog *et al.* 2000; Hile *et al.* 2000), for these analyses data were partitioned into two groups, according to repeat type: 15 GA repeat loci and 4 TA repeat loci; the one trinucleotide repeat locus in this study was omitted from these analyses. To further examine these relationships, data for the GA repeat loci were divided into groups of high and low locus diversity.

Genetic analyses: Gene diversity estimates for each locus were calculated by $n(1 - \Sigma p_i^2)/(n - 1)$, where n is the number of samples and p_i is the frequency of the ith allele, following the methods of Nei (1973) and Matsuoka *et al.* (2002). The value n is used here in place of $2n$ because all *A. thaliana* accessions are expected to be nearly completely homozygous due to inbreeding.

The fit of each locus' distribution to expected distributions under three different mutation models, the SMM, the IAM, and an intermediate two-phase model (TPM), was tested using the program BOTTLENECK (Cornuet and Luikart 1996). Because of sampling, data for all accessions were treated as a single population, which is not ideal, but is unavoidable. Observed allele frequencies and sample sizes were input parameters. These analyses provide a test statistic for the probability that an observed allele distribution with a given heterozygosity (gene diversity) was generated under each of the three mutation models.

To describe the distribution of alleles for each locus, measures of skewness (g_1) and kurtosis (g_2) were calculated following Sokal and Rohlf (1995). Significant differences between low- and high-diversity loci were tested for by a simple *t*-test for each measure. Because of different mutation rates between dinucleotide and trinucleotide loci (Chakraborty *et al.* 1997; Sia *et al.* 1997), the GapAB locus was omitted from these analyses.

For similarity analyses, allele size class data were transformed into alphanumeric codes. From this transformed data set, pairwise distances were obtained on the basis of the proportion of shared alleles, as implemented in PAUP*4.0b10 (Swofford 2002). As the complete evolutionary history of *A. thaliana* accessions is partially reticulate and therefore cannot be accurately represented by a bifurcating tree, a majority-rule (70%) consensus tree of 1000 independent cluster analyses using unweighted pair group method using arithmetic averages (UPGMA) is presented to simply illustrate genetic similarity among accessions. In the course of building trees, cluster analyses have to randomly break ties between equivalent relationships. As a result, there is a stochastic component to resulting trees. One thousand independent UPGMA analyses were run on the complete data set and only relationships consistent with the 70% majority rule are presented to provide a more rigorous analysis and conservative tree. More detailed analyses aimed at reconstructing the intraspecific phylogeny of *A. thaliana* will be presented elsewhere.

Because low- and high-diversity loci may be influenced by differing mutation dynamics, we conducted a partition homogeneity test implemented in PAUP (Farris *et al.* 1994, 1995; Swofford 2002), which tests for the probability of significant conflict between data partitions with regard to phylogeny. The total data set was partitioned into two mutually exclusive groups, conservatively excluding the nga129 locus altogether because of its intermediate diversity measure. The low-diversity group included all loci with gene diversity measures <0.70 and the high-diversity partition included all loci with gene diversity measures >0.80.

RESULTS

Amplification fidelity: Amplification success varied both across the 20 loci and among the 126 accessions of *A. thaliana*. Amplification frequencies for each of the 20 loci investigated are listed in Table 2. Amplification success ranged from 77 to 98% across loci and from 70 to 100% among accessions (excluding four accessions; data not shown), with a total of 90% amplification success. No significant correlation was found between amplification success and any measure of locus diversity (analyses not shown). Of the 2526 marker-by-individual data points, only 4 (0.2%) were found to be heterozygous. This frequency is similar to that reported for 12 accessions of *A. thaliana* by Clauss *et al.* (2002). Amplification of two loci, GapAB and CIW10, consistently yielded two products for all accessions. In each case, the size of one of the products was constant among all accessions and the other varied. Considering the high level of gene duplication within *A. thaliana* (Vision *et al.* 2000), this observation may represent the simultaneous amplification of two distinct loci. In each case, only the variable allele was included in our analyses.

Allelic diversity within and among loci: There is a high degree of variation for allelic diversity among microsatellite loci (Figure 1). The most striking differences

TABLE 2

Microsatellite locus table

Marker	Chromosome	Position (cM)	Repeat type	% amplification success	Allele range	No. of alleles	Gene diversity	SMM	TPM	IAM
nga59	1	1.6	CT	92	111–192	31	0.94	$P < 0.05$	NS	NS
ZFPG	1	37.4	TC	98	127–236	26	0.87	$P < 0.01$	$P < 0.05$	NS
Centromere	1	62.0	—	—	—	—	—	—	—	—
nga128[a]	1	83.0	TC	90	177–227	13	0.90	NS	NS	NS
nga692	1	119.0	GA	83	106–152	25	0.90	$P < 0.01$	$P < 0.05$	NS
nga1145[a]	2	9.6	GA	94	208–239	11	0.45	$P < 0.001$	$P < 0.001$	$P < 0.005$
Centromere	2	19.0	—	—	—	—	—	—	—	—
nga1126	2	50.6	GA	93	182–221	17	0.87	$P < 0.05$	NS	NS
AthUbique[a]	2	82.0	CT	95	164–172	5	0.52	NS	NS	NS
nga172[a]	3	7.0	GA	92	152–244	31	0.95	NS	NS	NS
GaPAB	3	43.8	TTC	98	135–150	4	0.50	NS	NS	NS
Centromere	3	59.0	—	—	—	—	—			
nga707	3	78.0	TC	83	119–141	10	0.52	$P < 0.001$	$P < 0.01$	$P < 0.05$
CIW5[a]	4	5.3	TA	77	155–200	14	0.62	$P < 0.001$	$P < 0.001$	$P < 0.05$
Centromere	4	20.0	—	—	—	—	—	—	—	—
nga8	4	24.2	GA	91	122–222	38	0.96	NS	NS	NS
CIW7[a]	4	65.0	TA	87	126–180	22	0.92	$P < 0.05$	NS	NS
nga1139	4	83.4	TC	97	74–142	22	0.93	NS	NS	$P < 0.05$
nga1107	4	105.0	GA	90	134–155	9	0.41	$P < 0.001$	$P < 0.005$	$P < 0.05$
nga249	5	23.1	TC	98	115–139	11	0.49	$P < 0.001$	$P < 0.001$	$P < 0.05$
CDPK9	5	44.5	TC	87	86–179	22	0.86	$P < 0.005$	$P < 0.01$	NS
Centromere	5	70.0	—	—	—	—	—	—	—	—
CIW9[a]	5	88.0	TA	83	138–208	28	0.89	$P < 0.005$	$P < 0.005$	$P < 0.05$
nga129[a]	5	105.0	GA	83	179–205	14	0.71	$P < 0.001$	$P < 0.001$	NS
CIW10[a]	5	128.0	TA	83	136–187	20	0.93	NS	NS	$P < 0.05$

NS, not significant.

[a] Loci utilizing the M13 tailing scheme. True allele sizes for these loci are expected to be 19 bp smaller than those reported here because of the use of the M13 forward sequence.

are in the variation at a locus (number of alleles scored) and how that variation is distributed among alleles at a locus (gene diversity). These two measures are reported for all loci in Table 2. The average number of alleles detected per locus is 17.6 (range, 4–38). The average gene diversity estimate from our data is 0.76 (range, 0.41–0.96; Figure 2) and does not differ appreciably from that of 0.79, reported by Innan *et al.* (1997). Several different distribution patterns of allelic diversity are evident (Figure 1). For further analysis, loci were split into two very broad categories (Figure 2): high diversity (above the mean) and low diversity (below the mean).

High-diversity loci tend to be either somewhat normally distributed or strongly positively skewed ($\overline{X}$ skewness = 1.11). These loci also tend to have leptokurtotic distributions ($\overline{X}$ kurtosis = 1.71). Low-diversity loci show distribution patterns similar to those of the high-diversity loci, typically positively skewed ($\overline{X}$ skewness = 1.79) and leptokurtotic ($\overline{X}$ kurtosis = 3.98), but to a significantly greater degree ($P < 0.05$ for both tests). The tendency of microsatellite loci to mutate more frequently to larger allele sizes than to smaller sizes (becoming positively skewed) is well documented (Rubinsztein *et al.* 1999; Brohede *et al.* 2002).

Sequence results: As initially scored, PCR products for many loci displayed single-base-pair differences among alleles; however, our sequence data showed that ~95% of single-base-pair differences initially detected were artifactual. Reexamination of the original electropherograms determined that these discrepancies were attributable to the inconsistent nontemplate-dependent terminal transferase activity of *Taq* polymerase that adds a single deoxyadenosine (A) to the 3′ ends of PCR products. Although at a low frequency, instances of true single-base-pair differences were also revealed (*e.g.*, see alleles of locus nga129 in Figure 3). All sequenced size outliers proved to be the expected locus.

Molecular variation at high-diversity loci: Individual alleles of three loci demonstrating high gene diversity were cloned and sequenced. An allelic alignment for a representative locus is shown in Figure 3 (Nga8). All 34 alleles sequenced from these loci were found to be either "perfect," that is, without interruptions of any kind within the repeated region (Nga172 and CIW7), or possessing nearly fixed interruptions in the extreme end of the repeat region (Nga8). With this one exception, the only source of size variation identified at these high-diversity loci was changes in repeat number. Although

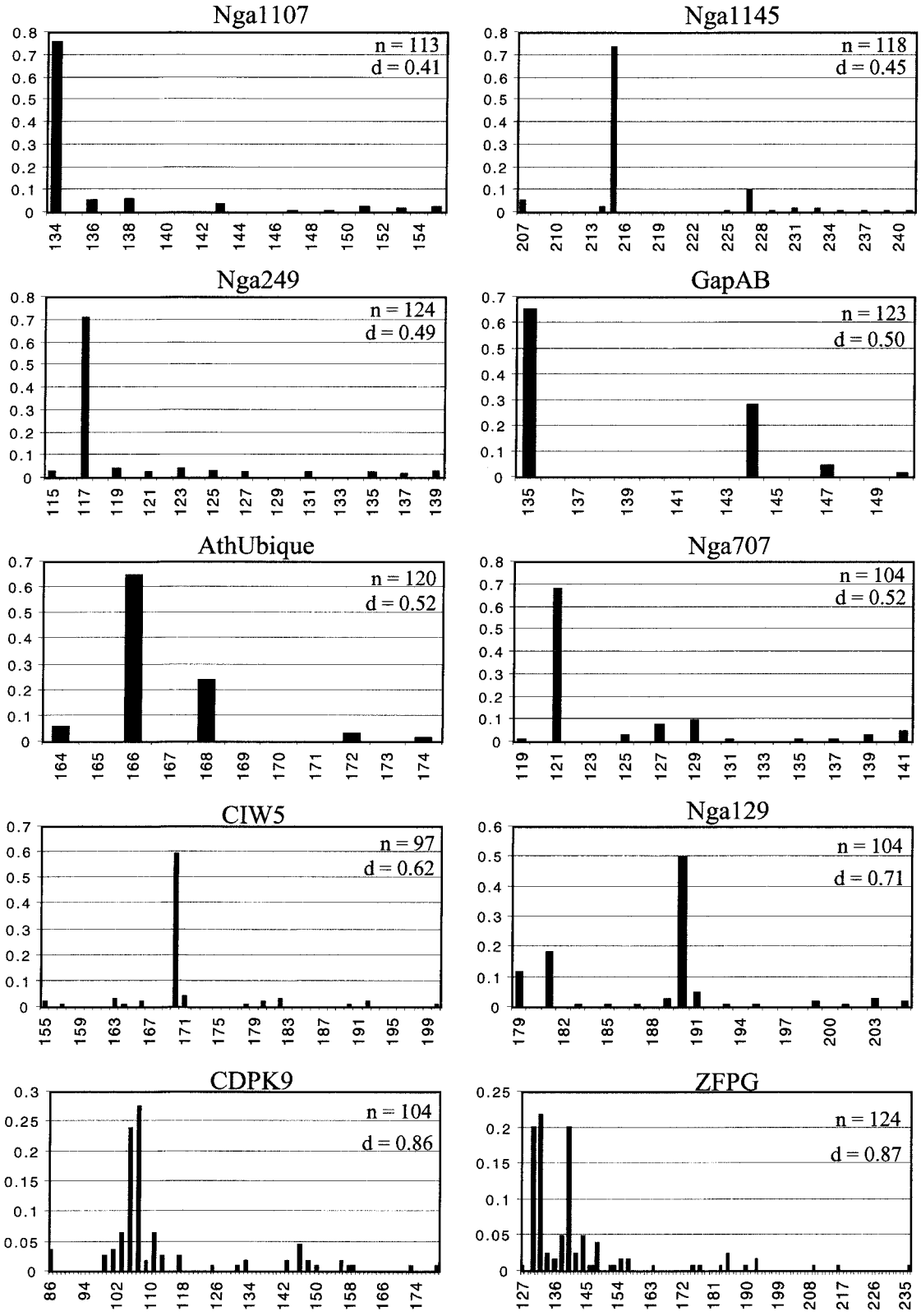

FIGURE 1.—Histograms showing allelic distributions for the 20 microsatellite loci examined. Allele size in base pairs is shown along the *x*-axis and the frequency of each allele class is displayed along the *y*-axis. Loci are arranged from lowest to highest gene diversity. Sample size (n) and gene diversity (d) are shown in the top right of each histogram.

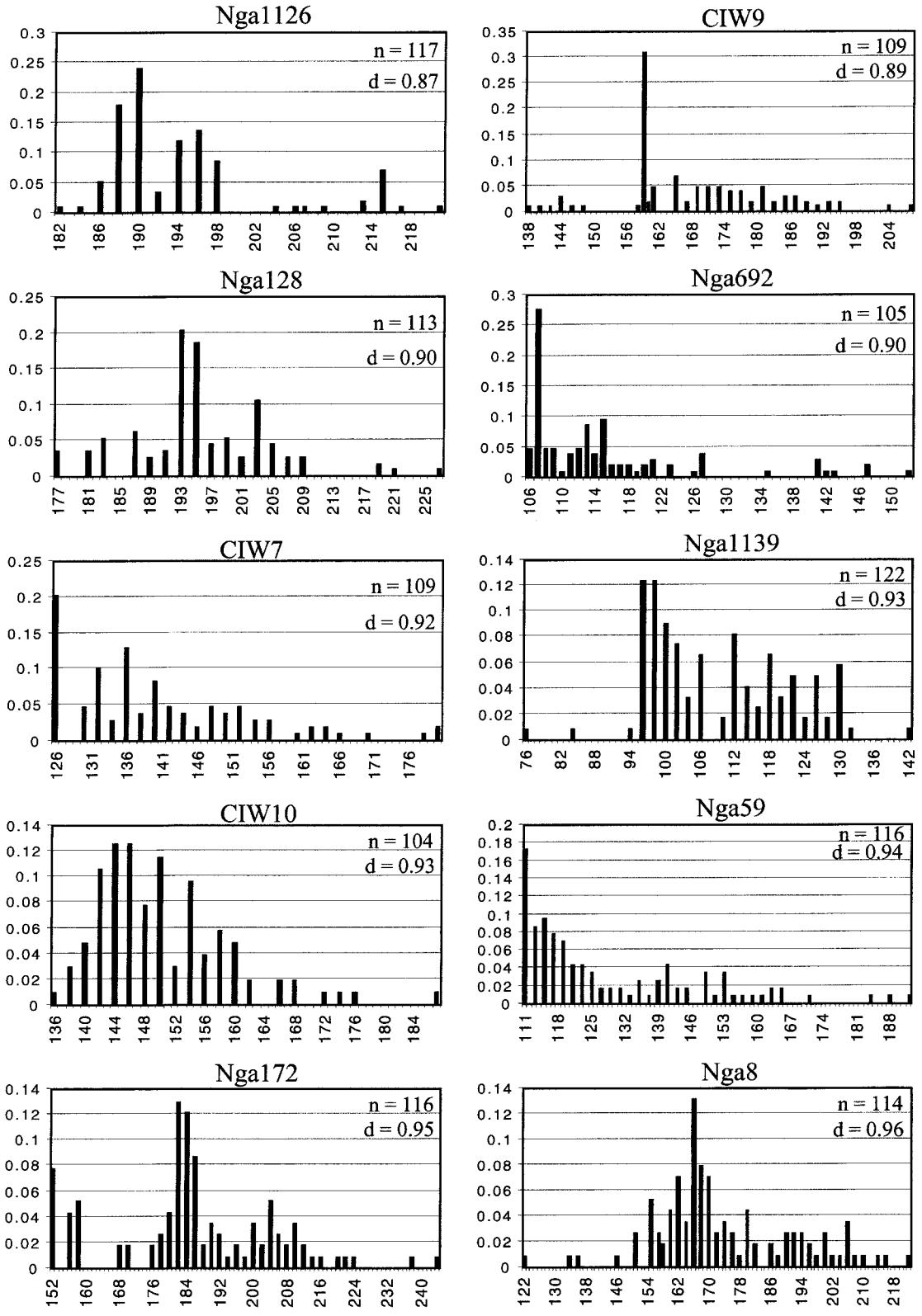

FIGURE 1.—*Continued.*

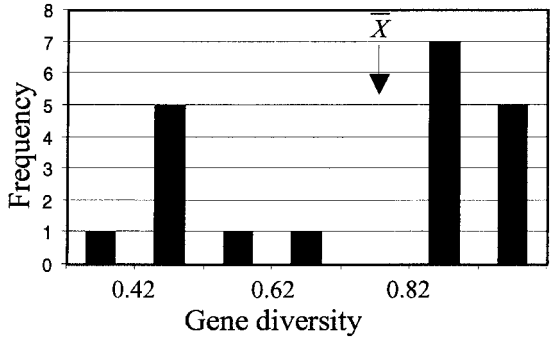

FIGURE 2.—Distribution of gene diversity measures among the 20 microsatellite loci. $\bar{X}$ indicates placement of the mean.

point mutations were identified in flanking regions, no insertions and deletions were revealed.

Molecular variation at low-diversity loci: The three low-diversity loci for which alleles were sequenced each revealed alleles with interruptions within the repeated region (Figure 3). In each case, the interruptions consisted of 2-bp insertions, back-to-back nucleotide substitutions, or some combination thereof; the origins of interruptions within tandemly repeated regions typically cannot be distinguished from among these possibilities.

Of the 17 alleles of the nga129 locus that were sequenced, only one size class (the most common) revealed an interruption. The 190-bp allele possesses a CT doublet within the repeat region, along with a 2-bp mutation, that immediately flanks the 3′ end of the microsatellite locus. These two mutations were always found to be linked. That is, no alleles were sequenced that possess one mutation and not the other. This pair of mutations was found only in the 190-bp allele, and all 8 alleles of this size that were sequenced are identical. The remaining variation detected among alleles at this locus appears to be the result of varying repeat number only.

Upon sequencing 25 alleles from the nga1145 locus, an AA interruption three repeat units from the 3′ end of the locus was discovered. Unlike the nga129 locus, this interruption is evident in many alleles (size classes), rather than in only the most common allele. Other sources of variation at this locus include a unique GG mutation, immediately flanking the AA interruption, and a single finding of an apparent duplication event composed of the entire microsatellite locus (accession no. 6672). For this locus also, all remaining allelic diversity appears to be due to repeat-number variation.

As with most loci, the primary source of size variation is change in repeat number for the nga1107 locus also; however, this locus is the most complex with regard to interruptions. It consists of four GA repeat regions, separated by 11-, 14-, and 2-bp interruptions, from the 5′ to 3′ ends, respectively. The second of these three interruptions appears to be a complex of successive VNTR loci (GCGC/TT/AAA/CCC/TA). Excepting its absence from one line, however, no sequence variation was uncovered within this complex among the seven accessions sequenced. Interestingly, the accession missing this insertion is the common reference strain, Col-0. Col-0 also lacks the second insertion and, despite these deletions (or lack of insertions), possesses the longest allele sampled at this locus due to many more repeats.

Relationship between contiguous repeat length and gene diversity for all loci: Correlation analyses show a general positive relationship between number of repeats possessed by the mean allele of a locus and locus diversity; however, this relationship varies depending on how the data are partitioned (Table 3). For the comparisons made, Pearson's and Spearman's correlation coefficients are in general agreement; therefore, unless stated otherwise, discussion applies to results of both tests. For the 15 loci with GA repeats, the total number of repeats possessed by the mean allele does positively correlate with locus diversity. However, the number of uninterrupted repeats demonstrates a stronger and (for Pearson's) more significant correlation with locus diversity. Analyses including only high- or low-diversity loci show the same trend, with one clear exception; for low-diversity loci, the total number of repeats in the mean allele shows *no* significant correlation with locus diversity. The four TA repeat loci show the same general positive correlation between locus diversity and repeat number, again, with the number of contiguous repeats being more tightly correlated with diversity than the total number of repeats. The smaller sample size for TA repeat loci precluded more detailed analyses.

Testing the SMM, TPM, and IAM: Results of mutation model tests are shown in Table 2. Of the 20 loci examined, 5 potentially fit all three models of evolution tested and 6 display distributions that do not differ significantly from the expected distribution under any of the three models tested (SMM, TPM, and IAM). Only 3 loci rejected two of the three models, suggesting the third as a reasonable fit. On average, low-diversity loci show a much higher model rejection rate than do high-diversity loci (Table 4), although the relative rejection rate among tests is consistent between the two sets of loci. Consistent with other reports (see review by ELLEGREN 2000), the SMM was the most frequently rejected model (13/20 loci), although most loci show hallmarks of SMM-like evolution (Figure 1). Under the SMM and TPM, there is a general heterozygosity deficiency (19/20 and 17/20 loci, respectively), and under the IAM, there are an equal number of loci with heterozygosity excess and heterozygosity deficit.

Performance of microsatellite data in cluster analyses: To evaluate the performance of *A. thaliana* microsatellite loci for estimating intraspecific relationships, a majority-rule consensus tree based on 1000 UPGMA cluster analyses was generated (Figure 4). Because of the inclusion of particular pairs and sets of accessions, many relationships could be predicted. For example,

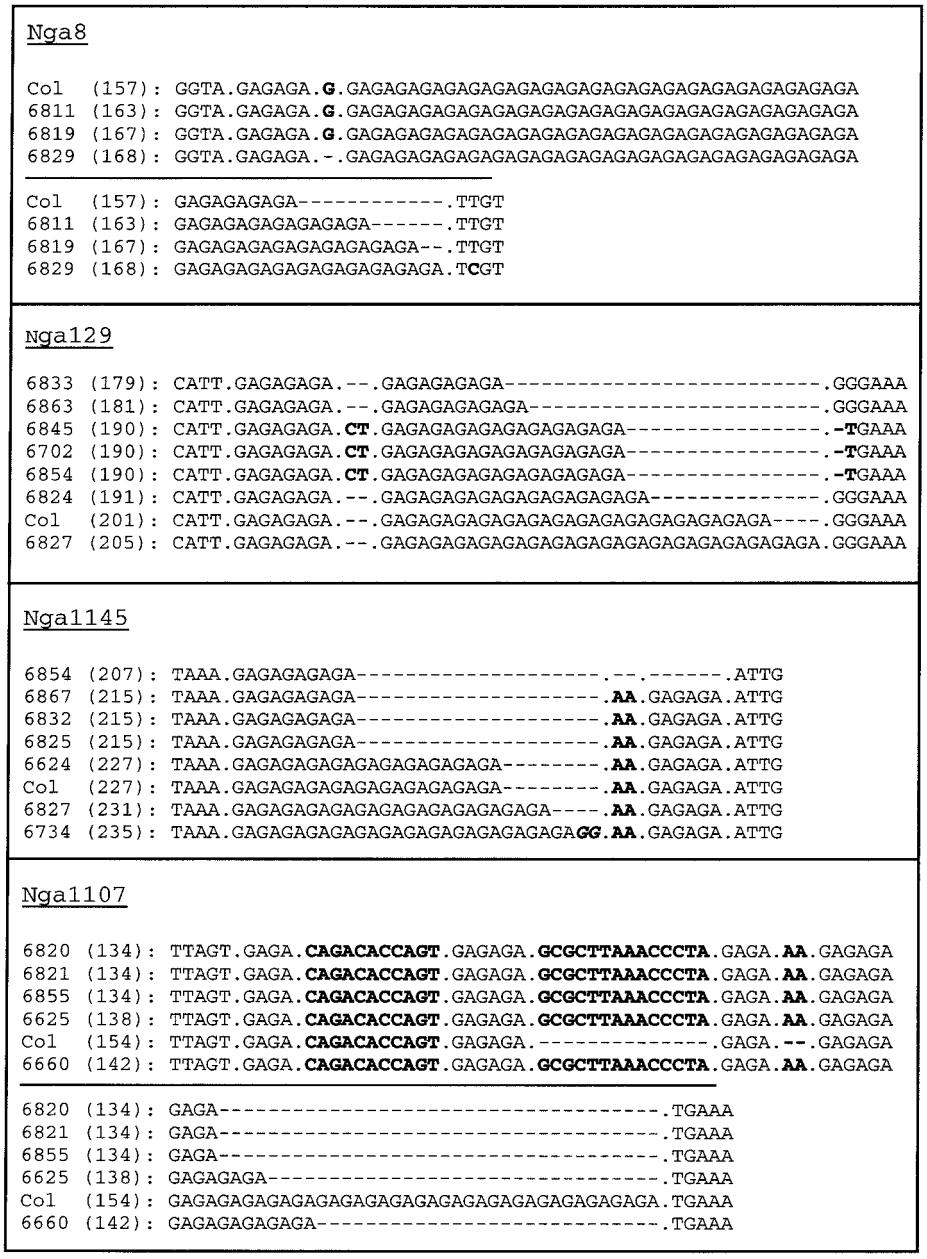

FIGURE 3.—Sequence alignments for representative high (nga8) and low (nga129, nga1145, and nga1107) gene diversity loci. Interruptions within repeat regions are highlighted in boldface type. As repeat-number variation was the only source of variation revealed at two of the high-diversity loci sequenced, CIW7 and nga172, only sequences from a representative locus, nga8, are shown. Dashes indicate gaps and dots serve to break up the sequence to aid viewing.

groups of accessions collected from the same locale were included (*e.g.*, Nok-0. Nok-1, Nok-2, etc.) and were expected to cluster together. The cluster analysis presented here reveals many groupings that are consistent with predicted associations. A selection of expected clusters are highlighted in Figure 4 and are discussed below. Interestingly, partition homogeneity tests revealed no significant incongruence between low- and high-diversity loci ($P = 0.90$).

DISCUSSION

Over the past decade, the frequency of microsatellite locus use has increased considerably (see reviews by ESTOUP and ANGERS 1998; ELLEGREN 2000). Despite this newfound popularity, detailed examinations of microsatellite mutation patterns and the forces that generate and maintain microsatellite diversity remain restricted to relatively few organisms. Here we present analyses of allelic variation at 20 loci for 126 accessions of *A. thaliana*. We provide sequence data supporting a strong role for repeat interruptions that result in relatively short repeat segments in stabilizing microsatellite loci and present a cluster analysis of all accessions.

Amplification fidelity: Each primer pair successfully primed amplification in an average of 90% of all accessions examined. This amplification rate is similar to that reported in other studies of *A. thaliana* microsatellites (INNAN *et al.* 1997; CLAUSS *et al.* 2002). The consistency of amplification rate among reports and repeated at-

TABLE 3

Correlation coefficients (*r*)

		All[a]		Low *d*[b]		High *d*[c]	
Repeat	Test	U[d]	T[e]	U	T	U	T
GA/TC	Pearson's	0.78***	0.70**	0.86*	0.29	0.84**	0.83**
GA/TC	Spearman's	0.93***	0.86***	0.81	−0.06	0.76*	0.76*
TA/AT	Pearson's	0.89	−0.63	—	—	—	—
TA/AT	Spearman's	0.95	−0.20	—	—	—	—

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.
[a] All loci of a particular repeat type.
[b] Low-diversity loci only ($d < 0.75$).
[c] High-diversity loci only ($d > 0.75$).
[d] Largest no. of uninterrupted repeats at a locus.
[e] Total no. of repeats at a locus.

tempts at amplifying individual null alleles (no amplification product) would seem to argue that the remaining null alleles are mainly due to sequence divergence within priming sites or deleted loci, rather than to spurious amplification failure. Furthermore, accessions suspected to be closely related show similar null allele patterns (data not shown).

Forces affecting mutation patterns: As has been reported in other systems (Bachtrog *et al.* 2000; Brohede *et al.* 2002), several different patterns of allelic distribution were revealed among the microsatellite loci of *A. thaliana*. The mutation models typically invoked to explain microsatellite distribution patterns are the SMM, the IAM, or some combination thereof (*e.g.*, the TPM). However, observed distribution patterns rarely fit the stringent SMM (Shriver *et al.* 1993; Ellegren 2000) and empirical evidence documenting independent identical mutations argue against the IAM (Brohede *et al.* 2002; Thuillet *et al.* 2002). Indeed, more than one-half of the loci examined here have distributions that either differ significantly from and thus reject all models or fit all models equally well (Table 2), effectively supporting none. An alternative to simple mutation dynamics in explaining the observed model-fit results is that some aspect of population demography has resulted in the observed allele distributions. Two of the mutation models support this alternative, showing strong trends toward heterozygosity deficit (17/20 loci for the TPM and 19/20 for the SMM), a finding that is consistent with hypotheses regarding the relatively recent and rapid expansion of *A. thaliana* global populations (Sharbel *et al.* 2000), while only under the IAM is the assumption of equilibrium met. Unfortunately, both mutation dynamic and demographic interpretations are compromised by violations of certain test assumptions, specifically that the sample represents a single contiguous population at mutation-drift equilibrium.

Beyond these models, it has been suggested that microsatellite locus equilibrium is a balance between polymerase slippage rate and mutation rate (Kruglyak *et al.* 1998; Schug *et al.* 1998). In short, the longer the string of uninterrupted repeats (*e.g.*, AGAGAGAG), the more likely is the generation of new alleles via slippage and recombination. Any mutation within the repeated region that causes an interruption (*e.g.*, AGAG***TTT***A GAG) will effectively split the original repeat region into two shorter segments. This is expected to increase locus stability (*i.e.*, reduce the generation of new alleles), simply by reducing the substrate for polymerase slippage and recombination. This model would seem to fit our typical finding of repeat interruptions in low-diversity loci and longer stretches of uninterrupted repeats within high-diversity loci.

To investigate this further, we examined the relationship between mean allele length and locus diversity for different data partitions (Table 3). If repeat disruptions stabilize loci simply by breaking them into smaller segments, then the degree of stability conferred should be dependent upon the lengths of the resulting repeat segments. This was tested by comparing the strengths of association between locus diversity and (1) the total number of repeats possessed by the mean allele at a locus and (2) the largest number of contiguous repeats possessed by the mean allele. The results show that gene diversity is more strongly correlated with the number of contiguous repeats than with the total number of repeats (Table 3); the number of contiguous repeats accounts for 12% (all GA repeats), 66% (low-diversity GA repeats), and 40% (TA repeats) more of the observed variation in genetic diversity, as determined by comparing coefficients of determination (r^2). The nature of this difference becomes evident when high- and low-diversity loci are examined separately; this was possible only for the GA repeat loci, where sample size was sufficient. The correlation with diversity turns out to be identical for total repeat number and contiguous repeat number for high-diversity loci. This is a result of high-diversity loci tending not to be interrupted, which

TABLE 4

Frequency of mutation model test rejection

	Test			
Locus type	SMM	TPM	IAM	Average
Low diversity[a]	0.86	0.86	0.71	0.81
High diversity	0.58	0.33	0.25	0.37

[a] Excluding the trinucleotide locus, GapAB.

means that the total number of repeats is equal to the contiguous number of repeats. Conversely, low-diversity loci demonstrate *no* (Spearman's) and very weak (Pearson's) relationships between total number of repeats and diversity (Table 3), whereas including only the number of contiguous repeats yielded some of the strongest associations with diversity observed among all complete and partitioned data sets. This provides strong evidence supporting a role for repeat disruptions in locus stability, one that is highly dependent upon placement of the interruption and the lengths of the remaining contiguous repeats. Because several of the low-diversity loci are without interruptions, this tight relationship also indicates that interrupted loci with few contiguous repeats behave in a manner similar to that of uninterrupted loci with few total repeats. As marker selection is often governed by criteria such as gene diversity, contiguous repeat number for mean allele size may provide a valuable predictor of marker utility. How broadly this relationship holds will require similar analyses in other organisms.

Size homoplasy: At any taxonomic level, the issue of size homoplasy in microsatellite data sets is an important and complicated one (Estoup *et al.* 2002). Size homoplasy can arise in a number of ways. Given the high mutation rate estimates for microsatellite loci (Hancock 1999), convergence on repeat number via slippage is likely the most common type and, unfortunately, impossible to detect *a posteriori*. As such, its frequency in *A. thaliana* cannot be addressed in our analysis. Another type of homoplasy involves mutations within the microsatellite locus other than changes in repeat number that result in size convergence. Through sequencing we have detected nonslippage mutations (*e.g.*, point mutations and insertions) within repeat regions that have led to size homoplasy. Each of the low-diversity loci possessed 2-bp repeat interruptions that could easily be misinterpreted as repeat-number variation from size data alone (see Figure 3). These findings suggest a strong potential for this type of size homoplasy. Fortunately, these cases are easily detected via sequencing. A third homoplasy type for microsatellite loci involves DNA insertions and deletions flanking the repeat region (Grimaldi and Crouau-Roy 1997). Our sequence data revealed predominantly point mutations in the immediate flanking regions; no insertions or deletions were discovered in 84 sequenced alleles (data not shown). Interestingly, a large-scale analysis of microsatellite marker loci in maize (Matsuoka *et al.* 2002) has shown that the most common source of variation is indels flanking the repeat locus. This sharp contrast in intraspecific sources of size variation underscores the need for more detailed microsatellite studies.

Cluster analyses: Previous efforts toward genealogy reconstruction within *A. thaliana* have resulted in somewhat well-resolved phylogenies including few accessions (Innan *et al.* 1997; van Treuren *et al.* 1997; Bergelson *et al.* 1998) or trees including many accessions with minimal resolution (Sharbel *et al.* 2000). Resolution in the former is likely due to the type of analysis presented; neighbor-joining approaches to tree building yield fully resolved trees, regardless of the level of support. Reports of low-resolution trees likely result from more rigorous analyses, but use markers with low mutation rates relative to the time scale involved.

A. thaliana accessions are derived from natural populations that likely have histories involving interpopulation gene flow and recombination. Because of this, their reticulate evolutionary history cannot be fully represented by analyses that yield bifurcating trees. However, to provide some reference of similarity among many *A. thaliana* accessions, a cluster analysis is presented here and selections of the results are discussed below. The tree presented (Figure 4) is not proposed as a phylogeny, but instead as a tentative framework and test of genealogical signal. This tree is a majority rule consensus of 1000 independent UPGMA runs and shows only relationships with strong support (*i.e.*, only relationships that occur in 70% or more of all independent runs are represented), while relationships with weak support are collapsed back to a central node. The finding of strongly supported clusters and unresolved relationships between clusters likely reflects the presumed reticulate history of populations within the species and recent independent evolution of separate lineages. Below we briefly discuss a few of the more interesting results.

The relationship between the two most-utilized reference strains, Col-0 and L*er*, remains unresolved. These two accessions are purportedly derived from the same seed stock, although details of that original stock remain elusive (Nottingham *Arabidopsis* Stock Center; http://nasc.nott.ac.uk). The accumulation of mutations due to either irradiation (in the L*er* line) or generations in cultivation likely cannot explain this finding as L*er* does show strong similarity to La-0 (6765), which was also derived from the above-mentioned stock. The Col-0 genotype does not match identically with any other accessions examined in our lab. In addition, our Col-0 DNA sequences match those in the database so that seed or DNA contamination in our lab also would not appear to explain this finding. Given the low levels of both phenotypic (personal observation) and genetic similarity between Col-0 and L*er*, it would appear that

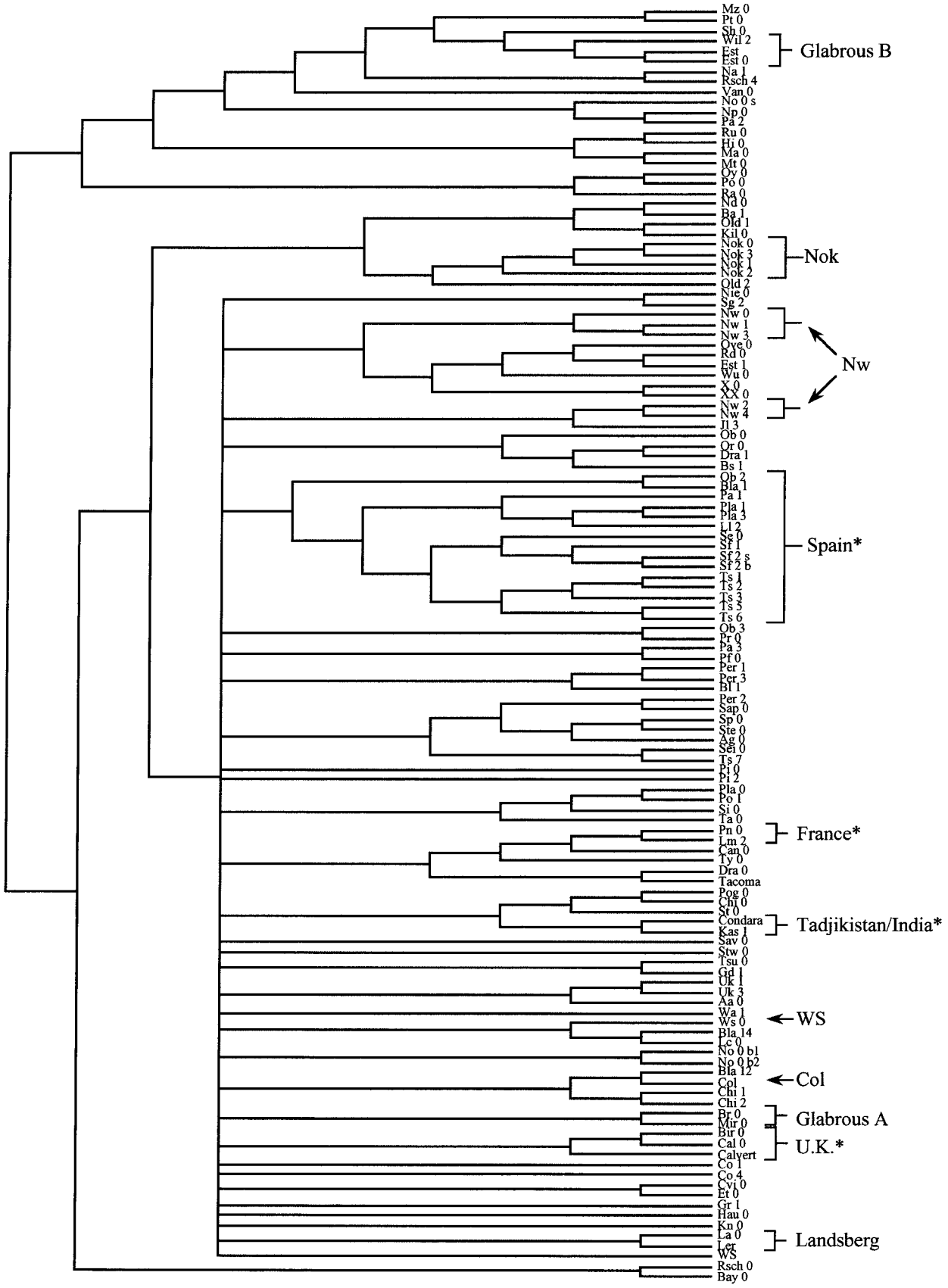

FIGURE 4.—Majority-rule (70%) consensus tree derived from 1000 independent UPGMA runs. Clusters and accessions of particular interest that are discussed in the text are denoted by brackets. Examples of accessions that cluster together according to geographic origin are marked with an asterisk. For details on UPGMA analyses, see MATERIALS AND METHODS.

the original stock was more heterogeneous than originally suspected; this is also in accord with reports of strong sequence divergence between the two accessions (*e.g.*, Noel *et al.* 1999; Borevitz *et al.* 2003).

Notwithstanding the exception just discussed, accessions originating from a common seed stock or collection site typically cluster together (*e.g.*, the Nok cluster). However, instances where all accessions from a locality do not cluster together (*e.g.*, the two NW clusters) are also evident. In all, 70% of expected groupings were resolved. Again, these findings may be due to local populations that are quite heterogeneous.

Past reports on *A. thaliana* genealogies have shown little to no correspondence between geographic origin and relatedness. This has been suggested to be the result of recolonization of central and northern Europe from glacial refugia (Sharbel *et al.* 2000). Although much of the consensus tree presented here shows similar incongruence between geography and genetic similarity, this finding is not ubiquitous; particular clusters show consistent biogeographic trends. For example, the "Spain" cluster illustrates close associations among many independent accessions collected from throughout Spain. Likewise, accessions from India and Tadjikistan cluster together, as do collections from several proximal geographic regions (examples are denoted with an asterisk in Figure 4).

For cases in which similarity is incongruent with geography, there are two likely explanations: (1) the resolved relationship is correct and explanations for the pattern observed must be sought or (2) the genealogy is incorrect and a more appropriate marker is required. Differentiating between these two is, of course, not always simple. One approach to this problem is to seek corroborating or refuting evidence for specific genealogical hypotheses. For example, our results show strong similarity between the Br-0 (6626) accession from Czechoslovakia and Mir-0 (6798) from Italy (Glabrous A cluster in Figure 4). It happens that both of these lines are glabrous (lacking hairs), a relatively uncommon phenotype among wild-derived accessions. Others have reported sequence data showing that these two accessions share the same allele at the GL1 locus (Hauser *et al.* 2001), which (when knocked out) is responsible for the glabrous phenotype. In addition, there is a second microsatellite-based cluster that contains glabrous accessions (Glabrous B cluster), which according to Hauser *et al.* (2001) share a defective GL1 locus. Taken together, strong support exists for these particular relationships. Although not well resolved across the entire tree, it is clear that microsatellite data possess signal useful for reconstructing the evolutionary history of this group and warrant further investigation.

Conclusions: This analysis reveals several important aspects of microsatellite evolution and application in *A. thaliana*. Most loci examined support no individual mutation model. Instead, it appears that sequence interruptions within the repeat region of microsatellite loci have a strong influence on the potential diversification of loci and should be taken into consideration in the construction of new microsatellite mutation models. Specifically, the magnitude of the effect of repeat interruptions is proportional to the lengths of the remaining intact repeat regions. Additionally, microsatellite loci of *A. thaliana* possess a high level of intraspecific phylogenetic signal. As these marker data are potentially of broad use, they can be accessed at http://www.esb.utexas.edu/arabidopsis2010/.

We are grateful to U. Mueller, D. Levin, R. Jansen, D. Hillis, J. Tate, V. Godoy, and two anonymous reviewers for helpful discussions and comments during the preparation of this manuscript. Additionally, we thank G. Stein and A. Ellington for technical assistance and facility use, respectively. This material is based on work supported by the National Science Foundation under grant no. MCB-0114976.

LITERATURE CITED

Alonso-Blanco, C., and M. Koornneef, 2000 Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. Trends Plant Sci. **5:** 22–29.

Bachtrog, D., M. Agis, M. Imhof and C. Schlötterer, 2000 Microsatellite variability differs between dinucleotide repeat motifs—evidence from Drosophila melanogaster. Mol. Biol. Evol. **17:** 1277–1285.

Balloux, F., and N. Lugon-Moulin, 2002 The estimation of population differentiation with microsatellite markers. Mol. Ecol. **11:** 155–165.

Barker, G. C., 2002 Microsatellite DNA: a tool for population genetic analysis. Trans. R. Soc. Trop. Med. Hyg. **96:** S21–S24.

Bell, C. J., and J. R. Ecker, 1994 Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics **19:** 137–144.

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

Borevitz, J. O., D. Liang, D. Plouffe, H. S. Chang, T. Zhu *et al.*, 2003 Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res. **13:** 513–523.

Boutin-Ganache, I., M. Raposo, M. Raymond and C. F. S Deschepper, 2001 M13-tailed primers improve the readability and usability of microsatellite analyses performed with two different allele-sizing methods. Biotechniques **31:** 24–28.

Brohede, J., C. R. Primmer, A. Moller and H. Ellegren, 2002 Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. Nucleic Acids Res. **30:** 1997–2003.

Casacuberta, E., P. Puigdomenech and A. Monfort, 2000 Distribution of microsatellites in relation to coding sequences within the Arabidopsis thaliana genome. Plant Sci. **157:** 97–104.

Chakraborty, R., M. Kimmel, D. N. Stivers, L. J. Davison and R. Deka, 1997 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94:** 1041–1046.

Clauss, M. J., H. Cobban and T. Mitchell-Olds, 2002 Cross-species microsatellite markers for elucidating population genetic structure in Arabidopsis and Arabis (Brassicaceae). Mol. Ecol. **11:** 591–601.

Cornuet, J. M., and G. Luikart, 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics **144:** 2001–2014.

Cruzan, M., 1998 Genetic markers in plant evolutionary ecology. Ecology **79:** 400–412.

Cunniff, C., 2001 Molecular mechanisms in neurologic disorders. Semin. Pediatr. Neurol. **8:** 128–134.

Doyle, J. F., and J. L. Doyle, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf material. Phytochem. Bull. **19:** 11–15.

Driscoll, C. A., M. Menotti-Raymond, G. Nelson, D. Goldstein and S. J. O'Brien, 2002 Genomic microsatellites as evolutionary chronometers: a test in wild cats. Genome Res. **12:** 414–423.

Eckert, K. A., A. Mowery and S. E. Hile, 2002 Misalignment-mediated DNA polymerase beta mutations: comparison of microsatellite and frame-shift error rates using a forward mutation assay. Biochemistry **41:** 10490–10498.

Ellegren, H., 2000 Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. **16:** 551–558.

Estoup, A., and B. Angers, 1998 Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations, pp. 55–86 in *Advances in Molecular Ecology* (Nato Sciences Series), edited by G. R. Carvalho. IOS Press, Amsterdam/Washington, DC.

Estoup, A., and J. Cornuet, 1999 Microsatellite evolution: inferences from population data, pp. 49–65 in *Microsatellites: Evolution and Applications*, edited by D. B. Goldstein and C. Schlötterer. Oxford University Press, New York.

Estoup, A., P. Jarne and J. M. Cornuet, 2002 Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol. Ecol. **11:** 1591–1604.

Farris, J. S., M. Kellersjo, A. G. Kluge and C. Bult, 1994 Testing significance of congruence. Cladistics **10:** 315–320.

Farris, J. S., M. Kellersjo, A. G. Kluge and C. Bult, 1995 Constructing a significance test for incongruence. Syst. Bot. **44:** 570–572.

Gill, P., A. J. Jeffreys and D. J. Werrett, 1985 Forensic application of DNA "fingerprints." Nature **318:** 577–579.

Grimaldi, M. C., and B. Crouau-Roy, 1997 Microsatellite allelic homoplasy due to variable flanking sequences. J. Mol. Evol. **44:** 336–340.

Hancock, J. M., 1999 Microsatellites and other simple sequences: genomic context and mutational mechanisms, pp. 1–9 in *Microsatellites: Evolution and Applications*, edited by D. Goldstein and C. Schlötterer. Oxford University Press, New York.

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

Hile, S. E., G. Yan and K. A. Eckert, 2000 Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. Cancer Res. **60:** 1698–1703.

Innan, H., R. Terauchi and N. T. Miyashita, 1997 Microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana*. Genetics **146:** 1441–1452.

Katti, M. V., P. K. Ranjekar and V. S. Gupta, 2001 Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. **18:** 1161–1167.

Kruglyak, S., R. T. Durrett, M. D. Schug and C. F. Aquadro, 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95:** 10774–10778.

Kubo, S., Y. Fujita, Y. Yoshida, K. Kangawa, I. Tokunaga *et al.*, 2002 Personal identification from skeletal remain by D1S80, HLA DQA1, TH01 and polymarker analysis. J. Med. Invest. **49:** 83–86.

Litt, M., and J. A. Luty, 1989 A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. **44:** 397–401.

Lukowitz, W., C. S. Gillmore and W. Scheible, 2000 Positional cloning in *Arabidopsis*. Why it feels good to have a genome initiative working for you. Plant Physiol. **123:** 795–805.

Matsuoka, Y., S. E. Mitchell, S. Dresovich, M. Goodman and J. Doebley, 2002 Microsatellites in *Zea*—variability, patterns of mutations, and use for evolutionary studies. Theor. Appl. Genet. **104:** 436–450.

McCouch, S. R., X. Chen, O. Panaud, S. Temnykh, Y. Xu *et al.*, 1997 Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant. Mol. Biol. **35:** 89–99.

Metzgar, D., E. Thomas, C. Davis, D. Field and C. Wills, 2001 The microsatellites of Escherichia coli: rapidly evolving repetitive DNAs in a non-pathogenic prokaryote. Mol. Microbiol. **39:** 183–190.

Moriguchi, Y., H. Iwata, T. Ujino-Ihara, K. Yoshimura, H. Taira *et al.*, 2003 Development and characterization of microsatellite markers for Cryptomeria japonica D. Don. Theor. Appl. Genet. **106:** 751–758.

Nei, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70:** 3321–3323.

Noel, L., T. L. Moores, E. A. van Der Biezen, M. Parniske, M. J. Daniels *et al.*, 1999 Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. Plant Cell **11:** 2099–2112.

Noor, M. A., R. M. Kliman and C. A. Machado, 2001 Evolutionary history of microsatellites in the obscura group of Drosophila. Mol. Biol. Evol. **18:** 551–556.

Ranum, L. P., and J. W. Day, 2002 Dominantly inherited, non-coding microsatellite expansion disorders. Curr. Opin. Genet. Dev. **12:** 266–271.

Richard, G. F., and F. Paques, 2000 Mini- and microsatellite expansions: the recombination connection. EMBO Rep. **1:** 122–126.

Rolfsmeier, M. L., and R. S. Lahue, 2000 Stabilizing effects of interruptions on trinucleotide repeat expansions in Saccharomyces cerevisiae. Mol. Cell. Biol. **20:** 173–180.

Rubinsztein, D. C., B. Amos and G. Cooper, 1999 Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Philos. Trans. R. Soc. Lond. B Biol. Sci. **354:** 1095–1099.

Sakamoto, T., and N. Okamoto, 2000 Microsatellite linkage map of rainbow trout and its application for QTL analysis. Tanpakushitsu Kakusan Koso **45:** 2872–2879 (in Japanese).

Schug, M. D., K. A. Wetterstrand, M. S. Gaudette, R. H. Lim, C. M. Hutter *et al.*, 1998 The distribution and frequency of microsatellite loci in Drosophila melanogaster. Mol. Ecol. **7:** 57–70.

Sharbel, T. F., B. Haubold and T. Mitchell-Olds, 2000 Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol. Ecol. **9:** 2109–2118.

Shriver, M. D., L. Jin, R. Chakraborty and E. Boerwinkle, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

Sia, E. A., R. J. Kokoska, M. Dominska, P. Greenwell and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol. Cell. Biol. **17:** 2851–2858.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*. W. H. Freeman, New York.

Swofford, D. L., 2002 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, version 4. Sinauer Associates, Sunderland, MA.

Thuillet, A. C., D. Bru, J. David, P. Roumet, S. Santoni *et al.*, 2002 Direct estimation of mutation rate for 10 microsatellite loci in durum wheat, Triticum turgidum (L.) Thell. ssp durum desf. Mol. Biol. Evol. **19:** 122–125.

van Treuren, R., H. Kuittinen, K. Karkkainen, E. Baena-Gonzalez and O. Savolainen, 1997 Evolution of microsatellites in Arabis petraea and Arabis lyrata, outcrossing relatives of Arabidopsis thaliana. Mol. Biol. Evol. **14:** 220–229.

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

Vision, T. J., D. G. Brown and S. D. Tanksley, 2000 The origins of genomic duplications in Arabidopsis. Science **290:** 2114–2117.

Wierdl, M., M. Dominska and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146:** 769–779.

Zane, L., L. Bargelloni and T. Patarnello, 2002 Strategies for microsatellite isolation: a review. Mol. Ecol. **11:** 1–16.

Zwettler, D., C. P. Vieira and C. Schlötterer, 2002 Polymorphic microsatellites in Antirrhinum (Scrophulariaceae), a genus with low levels of nuclear sequence variability. J. Hered. **93:** 217–221.

Communicating editor: V. Sundaresan