Theor Appl Genet (2007) 114:223–235
DOI 10.1007/s00122-006-0425-9

ORIGINAL PAPER

Evolution of chloroplast mononucleotide microsatellites in *Arabidopsis thaliana*

Mattias Jakobsson · Torbjörn Säll · Christina Lind-Halldén · Christer Halldén

Received: 20 December 2005 / Accepted: 30 September 2006 / Published online: 23 November 2006

Abstract The level of variation and the mutation rate were investigated in an empirical study of 244 chloroplast microsatellites in 15 accessions of *Arabidopsis thaliana*. In contrast to SNP variation, microsatellite variation in the chloroplast was found to be common, although less common than microsatellite variation in the nucleus. No microsatellite variation was found in coding regions of the chloroplast. To evaluate different models of microsatellite evolution as possible explanations for the observed pattern of variation, the length distribution of microsatellites in the published DNA sequence of the *A. thaliana* chloroplast was subsequently used. By combining information from these two analyses we found that the mode of evolution of the chloroplast mononucleotide microsatellites was best described by a linear relation between repeat length and mutation rate, when the repeat lengths exceeded about 7 bp. This model can readily predict the variation observed in non-coding chloroplast DNA. It was found that the number of uninterrupted repeat units had a large impact on the level of chloroplast microsatellite variation. No other factors investigated—such as the position of a locus within the chromosome, or imperfect repeats—appeared to affect the variability of chloroplast microsatellites. By fitting the slippage models to the Genbank sequence of chromosome 1, we show that the difference between microsatellite variation in the nucleus and the chloroplast is largely due to differences in slippage rate.

Communicated by H. C. Becker.

Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00122-006-0425-9 and is accessible for authorized users.

M. Jakobsson · T. Säll
Department of Cell and Organism Biology, Genetics, Lund University, Lund, Sweden

C. Lind-Halldén
Department of Mathematics and Natural Sciences, Kristianstad University, Kristianstad, Sweden

C. Halldén
Department of Clinical Chemistry, Malmö University Hospital, Malmö, Sweden

Present Address:
M. Jakobsson (✉)
Bioinformatics Program, Department of Human Genetics, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Ave, Ann Arbor, MI 48109-2218, USA
e-mail: mjakob@umich.edu

Introduction

DNA is a focus of evolutionary studies because of its obvious relation to the phenotype of organisms via genes, but also in its own right because the genomes of species vary in size and constitution as a result of evolutionary processes. The fact that DNA occurs not only in the nucleus but also in the mitochondria and, in green plants, the chloroplast, makes it possible to compare DNA sequences that share the same species history, demography and outer environment but that differ in other aspects such as organization and cellular environment. In animals, for example, it has been found that the substitution rate of synonymous sites in the mitochondria is about 10 times higher than in nuclear DNA, whereas in plants the substitution rates for mitochondrial, chloroplast and nuclear DNA are found in the approximate proportions 1:3:12 (Li 1997).

Microsatellites, or simple sequence repeats (SSR), are abundant in all well-studied eukaryotes (Katti et al. 2001). SSRs consist of short sequence motifs that are tandemly repeated, and their use as genetic markers has expanded dramatically in the last decade (Ellegren 2004). Despite the popularity of SSRs, their mode of evolution is not fully understood. In the early 1990s, it was reported that changes in the number of repeats in SSRs are primarily due to slippage during replication, and that most changes occur one step at a time (Weber and Wong 1993). It then became natural to investigate the applicability of the stepwise mutation model (SMM; Ohta and Kimura 1973) to describe the mutational process of SSRs. However, the SMM is associated with certain inherent problems both in its original single-step version as well as in multi-step versions (e.g. the two phase model, Di Renzo et al. 1994). The SMM explains neither the often-observed correlation between SSR length and the degree of variation (Weber 1990); nor the apparent upward restriction on the length of SSRs.

Some attempts have been made to address the problem of an upper bound using versions of the SMM in which the possible lengths of the SSRs are constrained (Feldman et al. 1997; Garza et al. 1995). These models were thereby able to provide stationary distributions of the lengths of alleles. However, because it was not clear why an upper bound should exist, Kruglyak et al. (1998) and Durrett and Kruglyak (1999) developed a model that could explain the existence of an observed maximum SSR length. This model includes both a length-dependent mutation rate and point mutations that break up the SSR into two parts. As a consequence, the length distribution of the repeats can be balanced by slippage and by point mutations. Mutations due to slippage increase or decrease the sequence by one step at a time, the probability of each being equal. Kruglyak and co-workers were the first to present a model without an upper bound that nevertheless produced a stable distribution of allele lengths. Motivated by the observation that short SSRs appeared to be almost invariable, Calabrese et al. (2001) suggested an improvement of the 'Kruglyak model'. The model, henceforth called the 'Calabrese model', assumes that no slippage occurs in SSRs of length less than a fixed value. Additional models have been suggested by Sibly et al. (2001) and Calabrese and Durrett (2003).

The chloroplast in plants has attracted increasing attention (Palmer 1987). The low mutation rate has rendered chloroplast DNA particularly useful in interspecific studies, although in intraspecific studies this poses a problem that can partly be alleviated by the use of SSRs. The development and analysis of chloroplast SSRs have been pioneered in certain crops and forest trees (see e.g. Powell et al. 1995; Provan et al. 1996; Vendramin et al. 1999), and further investigations have been conducted by Provan (2000), Ishii et al. (2001) and Chen et al. (2002). A brief review is provided by Provan et al. (2001).

Arabidopsis thaliana is the most widely used plant model organism. The published *A. thaliana* chloroplast chromosome is 154,478 bp long and thus has a similar size to most other plant chloroplast genomes. It contains two single-copy sequences, a large single-copy (LSC) of 84,170 bp and a small single-copy (SSC) of 17,780 bp which are separated by two inverted regions, IRA and IRB, each of length 26,264 bp (Sato et al. 1999).

Because chloroplast genomes do not recombine, the chloroplast chromosome represents only a single repeat of the demographic history and its power as a genetic marker may be somewhat limited for certain investigations. Even so, chloroplast information is useful, especially when data from both the chloroplast and the nucleus can be compared.

We have specifically investigated two aspects of SSRs. First, we searched for factors intrinsic to the chloroplast that influence the level of variation in an empirical study. Secondly, we compared the distributions of all mononucleotide repeats in the chloroplast and in the chromosome 1 database sequence with the predictions made by models of SSR evolution described above. Finally, the levels of empirically estimated mutation rates were compared to the levels predicted by the Kruglyak and Calabrese models. Our results showed that the only significant indicator of the level of variation is the repeat length. We also found that models which included a length-dependent mutation rate provided a very good fit to the observed distribution of mononucleotide repeats.

Materials and methods

Plant material

Fifteen accessions of *A. thaliana* were analysed in the study (Table 1). DNA was extracted from fresh leaves of a single greenhouse-grown plant using the Plant DNeasy Kit from Qiagen.

Sequence analysis

A total of 60 fragments of the chloroplast genome (GenBank Accession number: NC_000932) containing

Table 1 Accessions for which 60 DNA fragments were sequenced

Name	Location
T:160	Västervik (Sweden)
T:81	Karhumäki (Russia)[a]
T:93	Tvärminne (Finland) [a]
T:104	Nurmes (Finland) [a]
Oy-0	Ostese (Norway)[b]
T:10	Lilla Edet (Sweden)
T:20	Tollarp (Sweden)
T:700	Anten (Sweden)
T:340	Höör (Sweden)
T:350	Klevshult (Sweden)
KAS-1	Kashmir (India)[b]
LIP-0	Lipowiec (Poland)[b]
Sv-0	Svebolle (Denmark)[b]
WIL-1	Wilma (Lithuania)[b]
BU-0	Burghaun (Germany)[b]

Except in the following cases, accessions have been collected by the authors: [a]Outi Savolainen, Oulu University; [b]Nobuharu Goto, The SENDAI *Arabidopsis* Seed Stock Centre

at least one SSR were chosen for sequencing. There are 234 mononucleotide repeats that have a length of 8 bp or more in the database-sequence of the *A. thaliana* genome (Table 2). The 60 fragments sequenced contained 81 of these nucleotide repeats (Table 2). These 60 fragments also contained 9 of all 16 di-, tri- and tetranucleotide repeats having more than five repeat units in the database-sequence of *A. thaliana*.

The PCR/sequencing primers were designed using the OLIGO software (Appendix A). PCR-reactions were performed in a 25 µl mixture containing 0.5 ng of template DNA, a 1× PCR reaction buffer (Applied Biosystems), 2.5 mM $MgCl_2$, 0.4 µM of each primer (DNA technology A/S), 200 mM of each dNTP (Amersham Pharmacia Biotech) and 0.75 U of AmpliTaq Gold (Applied Biosystems). The PCR products were purified using the QIAquick 96 PCR Purification Kit from Qiagen. Sequencing of both strands was performed using labelled dye-terminators from Beckman (CEQ Dye Terminator Cycle Sequencing Quick Start Kit). Unincorporated dye-terminators were removed from the sequencing reactions by means of EtOH purification, in accordance with the supplier's recommendations. The sequencing was carried out on a Beckman CEQ 2000 sequencer, using short capillary arrays (CEQ Separation Capillary Array, 33–75B).

Table 2 The number and fraction of all nucleotide repeats in the *A. thaliana* chloroplast genome that were sequenced

Mononucleotide repeat length	GenBank	Sequenced	Fraction sequenced (%)
8	107	19	18
9	58	15	26
10	29	14	48
11 or more	40	33	83
Di-, tri-, tetra-nucleotide repeats	16	9	56

The total number of nucleotide repeats of a particular number of repeat units found in the GenBank sequence (NC_000932) indicated under 'GenBank'

The sequences obtained from the 15 accessions investigated were aligned for each locus, with the lengths of the repeats (number of repeat units) scored using Phred quality values (Phred-phrap package from CodonCode, Dedham, MA, USA) and the Sequencher software from GeneCode (Ann Arbor, MI, USA). Every repeat consisting of five or more repeat units was scored as a repeat. All polymorphic SSRs were submitted to The Arabidopsis Information Resource (TAIR: http://www.arabidopsis.org) as polymorphisms under the following TAIR accession numbers: 1005468356–1005468496.

The SSR distributions

The *A. thaliana* chloroplast genome and chromosome 1 was searched for mono- and dinucleotide repeats using the EMBOSS package (Rice et al. 2000) to obtain the complete distributions of these SSRs. The expected distribution of mononucleotide repeats for a particular genome or sequence was computed using the following expression: $X(n) = p(\mathrm{Y})^n[1\text{-}p(\mathrm{Y})]^2 \times m$, where $p(\mathrm{Y})$ is the fraction of the nucleotide Y in the genome of consideration, n is the repeat length in units and m is the total number of bases in the genome. The expected distribution of dinucleotides was computed using a similar expression, $X(n) = [p(\mathrm{Y})p(\mathrm{Z})]^n[1\text{-}p(\mathrm{Y})][1\text{-}p(\mathrm{Z})] \times m$, where $p(\mathrm{Y})$ and $p(\mathrm{Z})$ are the fraction of the nucleotide Y and nucleotide Z ($\mathrm{Z} \neq \mathrm{Y}$) in the genome of consideration, n is the repeat length in units and m is the total number of bases in the genome.

Quantifying variation

The level of variation for different types of microsatellites was quantified by computing the number of alleles and the gene diversity. The gene diversity, H, was calculated as $H = n\left(1 - \sum_i p_i^2\right)/(n-1)$, where p_i is the relative frequency of allele i, and n is the total number of samples (Nei 1987). In order to test for deviations from neutrality, a test of gene diversity excess described by Cornuet and Luikart (1996) was employed. The test assumes balance between mutations and genetic drift and that microsatellite mutations occur

according to the SMM. The number of loci having excess gene diversity is known under the null hypothesis of neutrality.

Testing factors that may affect levels of variation

First, the relation between microsatellite variation and the length of the alleles was analysed by calculating the proportion of loci that were variable, given a sample of a single allele of specified length. For example, among the 60 sequenced fragments, there were a total of 25 repeat loci in which at least one of the 15 accessions had an allele of length exactly 8 bp, and 9 of these repeat loci were variable, whereas the remaining 16 repeat loci were invariable.

To determine if the physical location (in the non-coding part of the cp genome) of a microsatellite affected its level of variability, we tested whether the order of the loci along the cp chromosome influenced the level of variation. This was done by means of an ordinary runs-test (Sokal and Rohlf 1995) in which for each locus it was noted whether the level of variation was above or below the average. The distribution of 'runs', i.e. groups of consecutive loci on the same side of the average was then compared with the expected distribution given complete independence among loci.

To determine whether interruptions of mononucleotide repeats reduce the level of variation in the non-coding DNA, we searched for imperfect mononucleotide repeats that could be compared to perfect mononucleotide repeats. We identified all cases of imperfect repeats in which 8, 9, 10 and 11 bp were interrupted by a single nucleotide, and the longer of the two adjacent sequences was 6 bp at most (for example TTTTCTTTTT is an imperfect repeat of length 10 bp). There were 24 such imperfect repeats of length 8 bp, 19 of length 9 bp, 3 of length 10 bp and 4 of length 11 bp. The level of variability for each length category of imperfect repeat was then compared to the corresponding (one bp shorter) length category of perfect repeats. For example, the level of variation of imperfect mononucleotides of length 8 bp was compared to the level of variation of perfect repeats of length 7 bp.

In a second approach to test if interruptions affect the level of variation, the number of bases S identical to a given repeat, beyond the first base on either side of the repeat, was counted. For example, if on both sides a stretch of eight A's was limited by two non-A bases (e.g. CTAAAAAAGG), then the count S was 0, whereas if on both of the two sides there was one non-A, followed by an A and then a non-A (e.g. ATAAAAAAAAGA), S would be 2. The correlation of H and S was computed.

Using a tree-based model to estimate the slippage rate from polymorphism data

The slippage rate per year (=generation) of cpDNA mononucleotide repeats was estimated from the variation detected among the 15 accessions of *A. thaliana*. It was estimated in the following way for each repeat locus that was sequenced. We first obtained an unrooted neighbour-joining tree on the basis of a distance matrix, using all variable mononucleotide loci available (result not shown). This was done in order to estimate the genealogy of the 15 accessions. The distance matrix was calculated according to a strict one-step SMM so that the distance between x_1 repeat units and x_2 repeat units equaled $|x_1 - x_2|$. The total branch length of the genealogy was found to be 2.6 times the largest pairwise distance within the genealogy. Using an estimate of the time to the most recent common ancestor (MRCA) from a set of *A. thaliana* accessions that overlapped with the present 15 accessions (Säll et al. 2003), we computed an estimate of the total branch length of the genealogy. Using an upper estimate of the time to the MRCA of 400,000 years (Säll et al. 2003) yields a total branch length of 2.08 million years. We then counted the minimum number of mutational steps, assuming single-unit steps, required to explain the observed variation among the 15 accessions, given the estimated genealogy. This was done for each locus separately. The mutation rate for each repeat locus per generation was then calculated by dividing the number of mutational steps by the total branch length of the genealogy. Note that since the estimate of the total length of the genealogy is based on external information, the estimates of the mutation rate are independent of the effective population size. This method of estimating the slippage rate is tree-model-based and uses polymorphism data. However, to contrast this estimate to estimates of the slippage rate from slippage models, this tree-model-based estimate will henceforth be denoted as the 'empirical estimate' of the slippage rate.

Slippage models

Two different slippage models were tested. The first was the 'Kruglyak model' (Kruglyak et al. 1998) which involves a continuous Markov chain in which the states of the chain are positive integers that correspond to the number of repeat units in a microsatellite allele. In this model at most one transition of three types can occur in each generation. First, single-step changes can occur, with changes of +1 unit and –1 unit each happening at the rate $(r–1) \times b$, where r is the number of repeat units

of a particular microsatellite allele, and b is the slippage rate per repeat unit. Second, point mutations can cut an uninterrupted microsatellite into two parts, which happen at a rate of a. The model then only keeps track of the part to the ''left'' of the point mutation so that a microsatellite of length r will after such a point mutation assume one of the following states, 1, 2, ..., r–1. Third, a transition from state 1 of a microsatellite to state 2 can occur at some low rate c. This third type of transition is necessary in order to prevent state 1 from being an absorbing state. Durrett and Kruglyak (1999) showed that this Markov chain model leads to a stationary distribution of microsatellite lengths. The stationary distribution of the microsatellite lengths, π_i, $i > 0$, where i is the number of repeat units, can be computed by solving the following equations (Eqs. 1 and 2 in Kruglyak et al. 1998):

$$c\pi_1 = b\pi_2 + a\sum_{j=2}^{\infty}\pi_j,$$

$$b(i-1)\pi_i = bi\pi_{i+1} + ia\sum_{j=i+1}^{\infty}\pi_j,\ i>1$$

This model leads to a distribution that is a function of the parameters a, b and c. Since microsatellites are usually defined as having more than some given number of repeat units, often 4, the theoretical and the empirical distribution can be conditioned on, for example, >4 repeat units. Thus, the parameter c is not relevant when the microsatellites are assumed to have at least d repeat units, with $d \geq 2$. It then follows from Eq. 2 in Kruglyak et al. 1998, (dividing it by a) that the only parameter needed to compute the theoretical stationary distribution is the ratio $k = b/a$, or the ratio of the slippage rate per repeat unit to the point mutation rate. The empirical distribution can then be fitted to the theoretical distribution and if an estimate of the point mutation rate a is assumed, an estimate can be obtained for b. Using previously published estimates of the neutral substitution rate for the chloroplast of 2.9×10^{-9} (Säll et al. 2003) as an approximation of the mutation rate, we were therefore able to obtain estimates of b. Because of the heterogeneous GC content of the different regions in the chloroplast and because different GC contents lead to quite different expected distributions of nucleotide repeats, we limited the analysis to the non-coding single-copy DNA.

The second model to be tested was the 'PS/PM model' developed by Calabrese et al. (2001). It was also used to fit the nucleotide distributions and to estimate the slippage rate. This model, which is an extension of the 'Kruglyak model', involves the assumption that slippage does not occur in sequences with fewer than or equal to κ repeat units for some constant κ. For sequences of length $l > \kappa$ the mutation rate due to slippage is $2b(l - \kappa)$.

Given the quantity of data available for length distributions of mono-nucleotide repeats, even models that provide a good fit will be rejected if tested using, for example, an ordinary χ^2 goodness-of-fit test. Our approach was instead to calculate the sum of squared differences between the observed values and the values predicted by the model as a measure of how well the model predicts the data. This value can then be used to compare the fit of the model in relation to other models.

Results

A total of 244 cpDNA repeat loci with a length of five repeat units or more were studied in 15 *A. thaliana* accessions. Of these loci, 235 were mononucleotide repeats, and the 9 others were dinucleotide repeats or repeats with longer repeat units. Of the 235 mononucleotide repeat loci, 164 were found in the non-coding and the other 71 were found in the coding DNA (Table 3). The repeats found in coding DNA varied between 5 and 13 bp in length, with an average length of 7.0 bp. None of these showed any variation across the 15 accessions. Thus, only the non-coding DNA repeats were considered in the analyses that follow.

Factors influencing mononucleotide SSR variability

Effect of allele length

In analysing the 164 mononucleotide loci found in the non-coding DNA, our aim was to determine what

Table 3 Breakdown of the 235 chloroplast mononucleotide repeats that were sequenced in 15 *A. thaliana* accessions

Number of loci		
Category	Mean allele length ≥5	Mean allele length ≥7
All loci	235	112
In coding sequence	71	27
Variable	0	0
Invariable	71	27
In non-coding sequence	164	85
Variable	42	41
Invariable	122	44

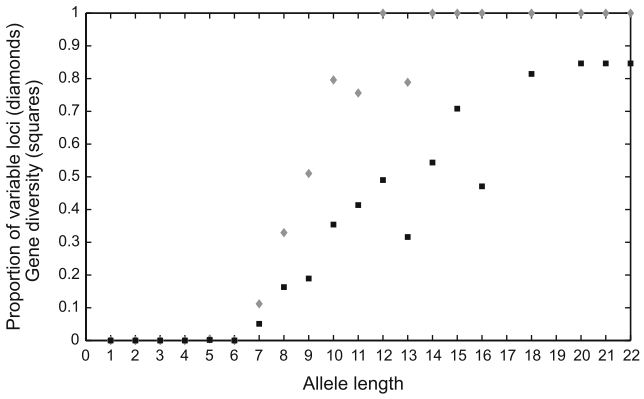

Fig. 1 The proportion of non-coding mononucleotide repeat loci that were variable in 15 *A. thaliana* accessions (*diamonds*) and gene diversity (*squares*), given the presence of at least one allele of specified length

factors intrinsic to the chloroplast chromosome influence the level of microsatellite variation in the loci. First the effect of allele length was examined. Figure 1 shows the proportion of loci that were variable, given the presence of at least one allele of specified length. A clear pattern is evident. There is a threshold at an allele length of 7 bp below which there are almost no variable loci, and a transition interval extending from 7 to 12 bp, where the proportion of variable loci increases. Above 12 bp, almost all loci are variable. A similar analysis of the relation between allele length and average gene diversity was also carried out for the same loci that were investigated above. As can be seen in Fig. 1 the level of gene diversity increases continuously with allele length without reaching a level of 1 (the upper limit).

Effect of physical location in the cp genome

Another aspect of sequence variation investigated here was the effect of spatial organisation. As mentioned earlier, the chloroplast genome is divided into four macrostructural regions: LSC, SSC, IRA and IRB. There are 85 repeat loci in the non-coding DNA if we only consider repeat loci with a mean allele length of 7 or more. The numbers of sequenced loci and variable loci were 70 and 34, respectively, in the LSC, 12 and 5 in the SSC, and 3 and 2 each in the IRA and the IRB. No significant difference between LSC and SSC was obtained, the average H being 0.24 for LSC and 0.18 for SSC ($P = 0.57$, Wilcoxon two-sample test). Thus, the macro-region location of a microsatellite in the chloroplast does not appear to affect variability.

To determine if the physical location of a microsatellite affects the level of variability, we tested whether the order of the loci along the cp chromosome influences the level of variation in any way. This was done by means of an ordinary runs-test (Sokal and Rohlf 1995) in which for each locus it was noted whether the level of variation was above or below the average. No relationship between variation and ordering of the loci was found ($t_s = 0.1433$, $P = 0.89$). Thus, the level of variation in a specific locus appears to be independent of variability at surrounding loci.

Effect of imperfect repeat patterns

We investigated whether imperfect repeats, i.e. tandem repeats interrupted by a single base, showed a different pattern of variation than perfect—or uninterrupted—repeats. Among the 60 fragments that we sequenced, there were a total of 50 repeat loci which, in the database sequence, were interrupted by a single nucleotide where the longer of the two adjacent sequences was at most 6 bp (24 repeat loci of length 8 bp, 19 of length 9 bp, 3 of length 10 bp and 4 of length 11 bp). None of these interrupted sequences were found to represent variable loci among the 15 accessions we sequenced. In comparison, the number of uninterrupted repeat loci that we found to be variable among the 15 accessions was 20 of 80, including 3 of 32 repeat loci of length 7 bp, 5 of 19 repeat loci of length 8 bp, 5 of 15 repeat loci of length 9 bp, and 7 of 14 repeat loci of length 10 bp.

In a second approach, we analysed the flanking sequences of perfect repeat loci. For a given repeat, the number of identical bases that extend beyond a single base pair interruption was counted. The correlation between the count S and gene diversity was $r_S = 0.05$ ($P = 0.54$). Thus, no indication of any relationship between the length of the adjacent number of nucleotides of the same type and the average gene diversity was found.

Levels of variation and deviation from neutrality

There were 85 non-coding mononucleotide repeat loci in our dataset that had a mean allele length of 7 bp or more. Of these loci, 41 were variable (Table 3). Among 79 non-coding mononucleotides with mean allele length between 5 and 7 only one variable mononucleotide was found (Table 3). The nucleotide distribution of the 85 repeat loci with a mean allele length of 7 bp or more was such that in 37 of them the repeated nucleotide was A, in 43 it was T, in 2 it was G, and in 3 it was C. Thus, any conclusion regarding the level of variation applies primarily to the A and T repeat loci. The average and the maximum, respectively, for the variation among these 85 repeat loci were 1.95

and 8.0 for the number of alleles, 0.23 and 0.85 for the gene diversity, 1.0 and 32.4 for the variance in allele length, and 0.47 and 5.69 for the standard deviation of allele length.

Only nine of the repeat loci were di-, tri-, or tetranucleotide repeats and they were all located in non-coding DNA. The level of microsatellite variation in these loci was quantified in terms of the number of alleles and *H*. The results are shown in Table 4. All but one of the loci were variable, having up to four alleles. The level of variation was high, the highest value of *H* being 0.70 and the average value 0.36.

We searched for deviations from neutrality using a test of gene diversity excess described by Cornuet and Luikart (1996). Of 50 variable loci, assuming mutation-drift equilibrium, 29 are expected to have excess gene diversity. Of the 50 variable cp microsatellites (42 were mononucleotides and 8 were di- tri- and tetranucleotides), 29 (23 mononucleotides and 6 di- tri- and tetranucleotides) showed a deficit in gene diversity significantly larger than expected ($P = 0.014$, sign-test). This result is consistent with directional selection acting on the cp or with recent population expansion.

Empirically estimated mutation rate

The average slippage rate was estimated from the variation detected in 164 non-coding mononucleotide repeat loci among 15 accessions of *A. thaliana.* The repeat loci were placed in bins based on their average allele length, and the average slippage rates were computed for each length class. The average slippage rate varied between 0 and 3.7×10^{-6} per generation among the length classes (Fig. 2), and the standard deviation of the estimates was 1.2×10^{-6}. This empirically estimated slippage rate increased strongly with allele length. When a regression line was fitted to the data, it showed an increase of 3×10^{-7} per bp and $R^2 = 0.70$ (Fig. 2).

Table 4 The level of variation of chloroplast microsatellites with repeat units of two bases or more in 15 accessions of the *A. thaliana*

Position	Repeat units	Mean length repeat (in units)	Macro-region	Number of alleles	Gene diversity
206	TTA/AAT	13.13	LSC	2	0.13
4690	AT/AAT	16.17	LSC	4	0.70
4747	AT/ATT	8.18	LSC	2	0.18
31309	AT	8.57	LSC	2	0.13
63122	AT	8.13	LSC	4	0.64
77774	(T)TTAA[a]	5.75	LSC	1	0.00
112698	AT	5.33	SSC	2	0.48
113339	TA(A)[a]	5.27	SSC	2	0.25
119145	TAT	9.95	SSC	3	0.67

[a] Two forms of repeat units existed; the most frequent repeat unit excludes the base in parenthesis

The distribution of repeat lengths in the database sequence of the *A. thaliana* chloroplast

We analysed the mono- and dinucleotide microsatellites of the GenBank *A. thaliana* chloroplast sequence. The first step in our analysis was to investigate the occurrence of mononucleotides in non-coding DNA and the distribution of their repeat lengths. Table 5 shows the complete distributions for all four nucleotides in the entire non-coding part of the cpDNA. The expected distributions based on a random sequence of nucleotides (conditional on nucleotide frequency) are also included for comparison. One can clearly see that the observed and the expected distributions differ considerably, with the observed distributions showing a larger number of long repeats. This effect starts to appear at 5 bp and becomes very strong at ≥7 bp. A goodness-of-fit test shows the deviations to be highly significant ($P << 0.0001$). The excess of longer repeats is much more pronounced for A and T than for G and C.

The next step was to investigate the dinucleotide repeats in the non-coding cpDNA of *A. thaliana*. Altogether, 16 different repeats are possible if direction is considered. The 'homogenous' repeats such as AA were excluded since they are mononucleotides repeats. The remaining 12 repeats form six pairs in which there is a strong dependence within each pair, for example, a six-unit repeat of AT also representing at least five repeats of TA. The distributions of six dinucleotide classes together with the expected random sequence distributions are given in Table 6. As expected, the AT repeats are most abundant, due to A and T being the most common nucleotides. The effect of an overrepresentation of long repeats is also present for AT repeats. A tendency of such a discrepancy between observed and expected numbers of loci was also observed for AG and TC. For the AT repeats, the discrepancy becomes obvious at four repeat units. Recall that for the mononucleotide repeats the same effect was apparent at seven repeat units.

Slippage models

The observed patterns for both the mono- and the dinucleotide repeats indicate that there is some process that acts on repeated sequences above a certain length. The obvious candidate for this process is slippage during replication. In order to investigate whether the

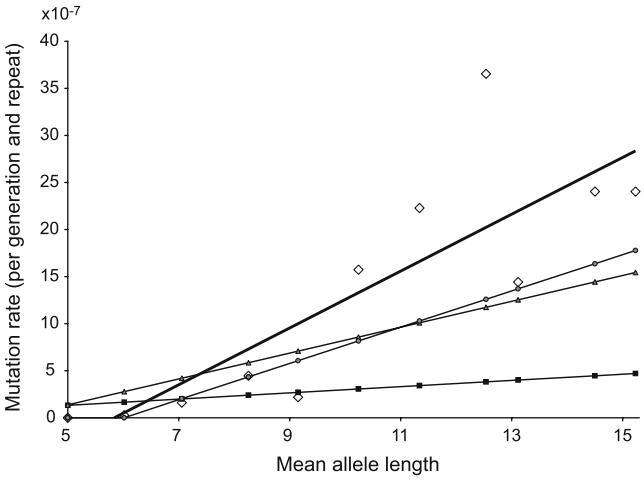

Fig. 2 Estimates of the slippage rate in relation to the allele length in the non-coding part of the single-copy regions (LSC and SSC) of the *A. thaliana* chloroplast genome. The empirically estimated slippage rates for each length class (see Materials and methods) are shown as *diamonds* and the fitted regression line is also shown (*solid black line*). The slippage rates predicted by the 'Kruglyak model' (cutoff = 5; *squares*), the 'Calabrese model' $\kappa = 3$; *triangles*) and the 'Calabrese model' ($\kappa = 6$; *circles*) are plotted in relation to the mean allele length

observed distributions could be explained by slippage, we fitted the length distributions to the model developed by Kruglyak and co-workers (Kruglyak et al. 1998; Durrett and Kruglyak 1999). The 'Kruglyak model' radically improved the fit to the observed distribution as compared with a distribution based on a random sequence. Figure 3a shows the results for the combined data of A and T, using five repeat units as the cutoff point. The ratio of the sum-of-squared differences of the 'Kruglyak model' (cutoff = 5) and the random sequence distribution was 0.14. In fitting the distribution based on the 'Kruglyak' model to the chloroplast data, a lower cutoff point was selected, as suggested by Kruglyak et al. (1998). A number of different cutoff points in the range of 5 to 10 repeat units were tested. The value of k varied only moderately, from 6.0, placing the cutoff point at 5 repeat units, to 3.5, placing it at 8 repeat units, corresponding to b-values in the range 1.0×10^{-8} to 1.8×10^{-8}.

We also used the 'Calabrese model', an extension of the 'Kruglyak model', to fit the distributions of nucleotide repeats and to estimate the slippage rate. The 'Calabrese model' assumes that slippage does not occur in sequences below or equal to a certain length, κ. The pattern of number of repeats shown in Table 5 indicates this modification to be justified and shows that κ should be approximately 5. The pattern of variation shown in Fig. 1 indicates that κ should be approximately 6. The model was fitted to the combined A and T data, and we tested values of κ ranging from 3 to 6. This model leads in all tested cases to much higher estimates of the slippage rate per unit than the

Table 5 The distributions for all four nucleotides in the non-coding part of the cpDNA of *A. thaliana* and the expected distributions of the nucleotides based on a random sequence conditional on the nucleotide frequency

Repeat units	Base							
	A		T		G		C	
	Obs[a]	Exp[b]	Obs	Exp	Obs	Exp	Obs	Exp
1	9,888	11,389.5	9,994	11,400.0	7,930	9,187.1	8,072	9,320.2
2	3,318	3,635.0	3,399	3,707.1	1,788	1,613.1	1,847	1,678.4
3	1,213	1,160.1	1,156	1,205.5	415	283.2	429	302.3
4	400	370.3	465	392.0	130	49.7	140	54.4
5	187	118.2	169	127.5	28	8.7	26	9.8
6	118	37.7	105	41.5	5	1.5	6	1.8
7	66	12.0	79	13.5	9	0.3	11	0.3
8	33	3.8	44	4.4	0	0.1	1	0.1
9	17	1.2	26	1.4	0	0.0	0	0.0
10	11	0.4	10	0.5	0	0.0	0	0.0
11	6	0.1	7	0.2	0	0.0	0	0.0
12	2	0.0	1	0.1	0	0.0	0	0.0
13	3	0.0	6	0.0	1	0.0	0	0.0
14	1	0.0	0	0.0	0	0.0	0	0.0
15	1	0.0	0	0.0	0	0.0	0	0.0
16	0	0.0	1	0.0	0	0.0	0	0.0
17	1	0.0	1	0.0	0	0.0	0	0.0

See Materials and methods for details

[a] Observed number of repeat units

[b] Expected number of repeat units

Table 6 Size distribution of dinucleotide repeats in the entire non-coding *A. thaliana* cpDNA

No. of repeats	Repeat unit											
	AT		GC		AG		TC		AC		TG	
	Obs[a]	Exp[b]	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
1	12,725	12,652.6	4,254	4,753.4	7,346	7,721.1	9,151	8,142.1	6,245	7,950.5	7502	7902.4
2	1,080	1,284.2	104	156.4	474	433.1	553	485.3	224	460.7	318	455.2
3	137	130.3	4	5.4	28	24.3	37	28.9	8	26.7	12	26.2
4	29	13.2	0	0.2	7	1.4	8	1.7	1	1.6	0	1.5
5	11	1.3	0	0.0	0	0.1	0	0.0	0	0.1	0	0.1
6	1	0.1	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
7	2	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
8	3	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Sum[c]	15,511	15,671.9	4,474	5,082.3	8,406	8,666.3	10,400	9,206.2	6,721	8,959.5	8174	8897.9

[a] Observed number of repeat units

[b] Expected number of repeat units conditional on nucleotide content (see Materials and methods for details)

[c] Total number of repeat units

'Kruglyak model', equalling 6.9×10^{-8} to 9.6×10^{-8}. In contrast, the distribution of mononucleotide repeats based on the 'Calabrese' model, shows a somewhat poorer fit than the distribution based on the 'Kruglyak model' (Fig. 3a), with the ratios of the sum-of-squared differences between the 'Calabrese model' distribution and the random sequence distribution equalling 0.37 for $\kappa = 3$ and 0.20 for $\kappa = 6$.

Chloroplast versus nucleus

Although several studies have considered the ability of microsatellite models to explain genome-wide distributions of microsatellite repeat sizes (Calabrese and Durrett 2003; Calabrese and Sainudiin 2005), we have found no other results in the literature that have applied the 'Kruglyak model' or the 'Calabrese model' to chloroplast data. Thus, no direct comparisons with other chloroplast findings can be made with respect to slippage rates or to the fit to the distributions. An important comparison, of course, is that with the nucleus of *A. thaliana*. We found no direct estimates of the mutation rates of mononucleotides in the literature. Instead, we compared slippage rates estimated by fitting the 'Kruglyak model' and the 'Calabrese model' to the observed distributions of mononucleotides in both the chloroplast and the nucleus. We assembled all mononucleotide and dinucleotide repeats in the non-coding DNA of chromosome 1, with the exception of a 3 Mb region around the centromere. The GC content of this data set was 31%. Figure 3b shows the results of fitting the 'Kruglyak model' (cutoff = 5), the 'Calabrese model' with $\kappa = 3$, and the 'Calabrese model' with $\kappa = 6$, respectively, to the combination of A and T mononucleotides. The distribution of mononucleotide repeats in chromosome 1 resembles the distribution in the chloroplast. However the number of repeats found for each length category declines more quickly in the nucleus than in the chloroplast. All the models tested improved the fit of the expected distribution dramatically in comparison with a random sequence distribution (Fig. 3b, the ratios of the sum-of-squared differences being 0.12, 0.31 and 0.26). The ratio of the slippage rate to the substitution rate, k, was 2–3 times as high in the chloroplast as in chromosome 1, for both the 'Kruglyak model' and the 'Calabrese model' with $\kappa = 3$, but not for the 'Calabrese model' with $\kappa = 6$. To compute the slippage rates we used substitution rates from the literature. Säll et al. (2003) have reported a substitution rate in the chloroplast of 2.9×10^{-9} and Koch et al. (2000) reported a substitution rate in the nucleus of 1.5×10^{-8}. Using the estimates above, we found the nucleus to have approximately twice as high a slippage rate (0.38×10^{-7} to 1.3×10^{-7} per generation and repeat unit) as the chloroplast.

Fitting the 'Kruglyak model' and the 'Calabrese model' to dinucleotides from the chloroplast is simply not feasible because there are too few of them. When these models were fitted to the AT-, GA- and CT-repeats from the nucleus, we found higher slippage rates for these (0.78×10^{-7} to 6.7×10^{-7}per generation and repeat unit) than for the nuclear mononucleotides. The AT-repeats had a rate 5 to 18 times as high and the GA- and CT-repeats a rate about twice as high as the mononucleotides did. The fit of the model to the data on GA- and CT-repeats was exceptionally good, with the ratio of the sum-of-squared difference between the slippage models to the random sequence distribution being less than or equal to 0.017 for all versions of the model tested.

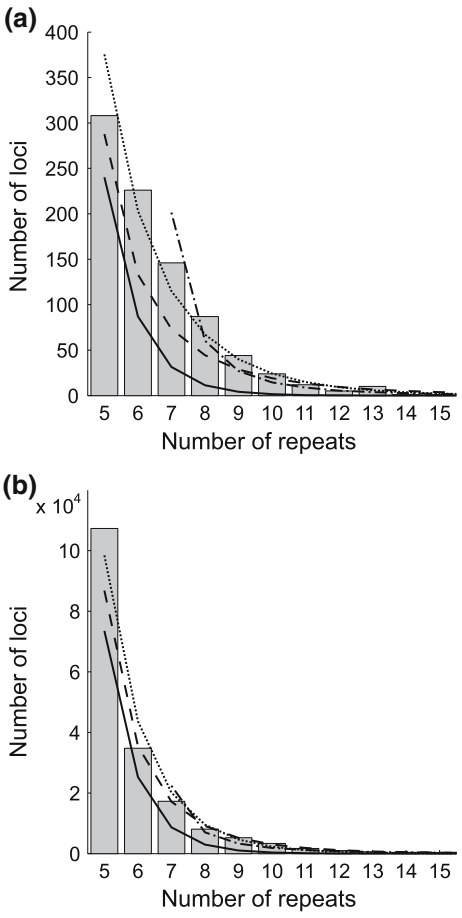

Fig. 3 Observed and expected distributions of A and T mononucleotide repeats from **a** the non-coding part of the single-copy regions (LSC and SSC) of the *A. thaliana* chloroplast genome and **b** the non-coding part of the *A. thaliana* chromosome 1. Observed numbers (*bars*), a random sequence distribution (*solid line*) and distributions based on the 'Kruglyak model' (cutoff = 5; *dotted line*), the 'Calabrese model' ($\kappa = 3$; *dashed line*) and the 'Calabrese model' ($\kappa = 6$; *dotted and dashed line*). For details of the models, see the text

Discussion

Level of variation in cpDNA

The level of variation in mononucleotide microsatellites found in the *A. thaliana* cpDNA is basically in line with other observations. Provan (2000) studied seven mononucleotide repeat loci in the *A. thaliana* chloroplast with repeat lengths that ranged from 13 to 16 repeat units. He found an average of 4.0 alleles per locus. Our own findings show an average of 3.3 alleles per locus in 13 loci with repeat lengths of 13 or larger. Provan et al. (1998) used a cutoff point at 10 bp when investigating mononucleotide chloroplast sequences in the genus *Pinus* and of 17 loci found in the non-coding DNA, 13 (76%) were variable. In wild soybean, Xu et al. (2002) found that all six tested cpDNA mononucleotide repeat loci were variable. These mononucleotide repeats had between 3 and 6 alleles and all alleles were at least 10 bp long. In cpDNA of *Aegilops tauschii*, Matsuoka et al. (2005) found that 8 of 11 mononucleotide repeats were variable. The invariable repeat loci were of length 8 and 9 bp. The fact that the overall level of variation is similar in as diverse genera as *Arabidopsis*, *Pinus*, and *Aegilops* suggests the generality of these observations, at least in plants.

We found that 8 of 9 di-, tri-, or tetranucleotide repeat loci were variable in the *A. thaliana* cpDNA. This indicates that whenever a di-, tri-, or tetranucleotide repeat comprising five or more repeat units is found in an *A. thaliana* accession, the nucleotide repeat is likely to represent a variable locus.

The observation of no variation among the coding sequences was not surprising due to the selective constraints on the coding DNA, as any variation, except in units of 3 bp, would result in nonsense mutations (Metzgar et al. 2000). The indication of directional selection or population expansion from the test of Cornuet and Luikart (1996) is interesting but should be interpreted with caution. The test assumes that the sample is taken from a panmictic population at mutation-drift equilibrium and also that the microsatellites follow the SMM model. Both these assumptions may well be violated. Despite these objections, it is interesting to note that we found the same tendency for the chloroplast as Symonds and Lloyd (2003) did for the nucleus. Säll et al. (2003) and Jakobsson et al. (2006a) obtained a similar although not significant result using SNPs from the chloroplast.

Factors affecting variation

Other investigators have used a locus-oriented approach to calculate the correlation coefficient between the average allele length and the level of variation. Both the Pearson (r_P) and the Spearman rank (r_S) correlation coefficients have been used. In cp DNA of *A. thaliana* we found $r_P = 0.66$ between H and average allele length, with the Spearman rank correlation coefficients being very similar. Provan et al. (1998) obtained a positive correlation between average allele length and variation for *Pinus sylvestris*. Using their data, we investigated the relationship between the numbers of alleles found in different accessions of *P. sylvestris* and allele lengths as listed in the cp database sequence for *Pinus thunbergi* and obtained $r_S = 0.78$. The fact that both the overall level of variation and the strength of the correlation between level of variation and allele length are similar in as diverse genera as

Arabidopsis and *Pinus* suggests that these observations are not confined to *Arabidopsis* alone.

For mononucleotide repeats there appears to be an allele length threshold, below which there is virtually no variation. This threshold was found to equal about 7 bp in *A. thaliana* cpDNA. In the yeast genome, a threshold of 9 bp was reported below which no apparent deviations from a random sequence distribution were observed (Rose and Falush 1998). The results of this study show that a detailed comparison of variation levels should take into account the average sequence length for each type of locus studied. Often, however, only a single average is taken over all investigated loci.

In the cpDNA of *A. thaliana*, we found no indications that imperfect repeat loci had higher levels of variation than the longest uninterrupted stretch of bases in the imperfect repeat. In humans, for microsatellites located in the nucleus, a similar result has been reported (Sibly et al. 2003). In conclusion, the critical property for the level of variation is the number of uninterrupted bases of the same type.

Slippage rate estimates

Provan et al. (1999) made an upper-limit (95%) estimate of the microsatellite mutation rate in *P. torreyana*. They obtained rates in the range of 3.2×10^{-5} to 7.9×10^{-5} per generation for repeats with an average of approximately 11 bp. Using a linear approximation, we estimated a mutation rate of 1.6×10^{-6} for 11 bp repeats in *A. thaliana*, which is more than one order of magnitude lower than the estimate of Provan et al (1999). Although the estimates of Provan et al. (1999) are upper-limit estimates, the true mutation rates probably being lower, a factor of 10 is a considerable difference. However, our estimates are per year, which corresponds to generation time in *A. thaliana*, whereas Provan et al. (1999) assume a generation time of 100 years, which means that their estimates are between one-fifth and one-half as large per year as our estimates. It would be of interest to know the average number of cell generations between the zygote and the flower in *A. thaliana* and in pine, since this would perhaps provide the most relevant comparison.

An independent comparison of the slippage rates estimated by the 'Kruglyak model' can be made using the slippage rates that were estimated based on our empirical data. Figure 2 shows our estimates of the slippage rate in relation to allele length. The figure also presents a line indicating the slippage rate predicted by the 'Kruglyak model'. As can be seen, the predicted slippage rate is higher than the empirically estimated rate for repeat loci with average allele length of 5 or 6 bp, but is much lower for repeat loci with longer average allele lengths. A systematic error in the empirical estimates of the slippage rate, such as through use of an incorrect estimation of the branch lengths of the genealogy, would only affect the slope of the regression line of the empirically estimated slippage rate. Thus, the observed discrepancy between the empirically estimated slippage rate and the slippage rate estimated from the 'Kruglyak model' cannot be explained by a systematic error of this type. Instead, the graph points strongly to the discrepancy being due to a violation of one or more of the underlying assumptions of the model.

The 'Calabrese model' leads to approximately five times the size of the estimates of slippage rates derived from the 'Kruglyak model' and is less than half the empirically estimated rate. As a consequence, the model provides a much better fit to the empirical mutation rate shown in Fig. 2. In addition, in comparing the predicted slippage rates with empirically estimated rates, Kruglyak et al. (1998) found the former to be approximately half the latter in size, the same level of agreement and discrepancy as we obtained using the 'Calabrese model'.

Chloroplast versus nucleus

It is well established that the substitution rate in chloroplast DNA is lower than that of the nucleus. This is clearly the case for *Arabidopsis* as well. Säll et al. (2003) have reported a substitution rate in the chloroplast of 2.9×10^{-9}, and Koch et al. (2000) reported a substitution rate in the nucleus of 1.5×10^{-8}. The ratio of these two estimates is 1:5, which is rather close to the ratios (1:4) of the substitution rates in the chloroplast and the nucleus reported earlier from, for example, comparisons in maize and rice (e.g. Li 1997). Both these values are derived from interspecific comparisons. When comparing levels of variation within *A. thaliana* we also see a difference, with lower levels of SNP variation in the cpDNA (Säll et al. 2003).

Symonds and Lloyd (2003) typed 20 di- and trinucleotide repeats from the nucleus of *A. thaliana* and obtained an average H of 0.76, and Jakobsson et al. (2006b) found an average H of 0.80 across 52 microsatellites. We found much lower values of H in the *A. thaliana* cpDNA both for di- tri- and tetranucleotides and for mononucleotides. This indicates that H is lower for microsatellites located in the chloroplast than for microsatellites located in the nucleus. Thus, the well-established difference in SNP variation between the nucleus and the chloroplast appears to exist for microsatellites as well.

In applying the 'Kruglyak model' to nuclear dinucleotide repeats in mouse, yeast and drosophila, Kruglyak et al. (1998) found the estimated slippage rates per repeat unit to be in the range of 2.3×10^{-7} to 1.0×10^{-5}. The slippage rate of *A. thaliana* nuclear DNA is towards the lower end of this range. Our chloroplast slippage rate estimates from the slippage models are much lower, even for the 'Calabrese model', arguing strongly for the slippage rates being lower in the chloroplast than in the nucleus.

Conclusions

Our results show chloroplast mononucleotide repeats to be both common and variable, thus providing a useful tool for tracing maternal lineages within and between closely related plant species. The level of variation is highly dependent upon the average allele length at a locus; loci with longer alleles are more variable than loci with shorter alleles. The number of nucleotides in an unbroken chain is thus the crucial property of a repeat locus. No other factors (except location in coding or non-coding regions), such as the position in the genome, appeared to have an important effect on the level of variation. The dependence of the level of variation on allele length could, to a large extent, be explained by models of slippage during replication.

Acknowledgments We thank M. Nordborg and O. Savolainen for their help with the material, M. Sterner and L. Hall for their technical assistance, B. O. Bengtsson and N. A. Rosenberg for their comments on the manuscript. The work was supported by the Crafoord Foundation, the Erik-Philip Sörensen Foundation, the Trygger Foundation and the Magnus Bergvall Foundation.

References

Calabrese P, Durrett R (2003) Dinucleotide repeats in the Drosophila and human genomes have complex, length-dependent mutation processes. Mol Biol Evol 20:715–725

Calabrese P, Sainudiin R (2005) Models of microsatellite evolution. In: Nielsen R (eds) Statistical methods in molecular evolution. Springer, Berlin Heidelberg New York

Calabrese PP, Durrett RT, Aquadro CF (2001) Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. Genetics 159:839–852

Chen X, Cho YG, McCouch SR (2002) Sequence divergence of rice microsatellites in Oryza and other plant species. Mol Gen Genome 268:331–341

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

Di Renzo A, Peterson AC, Garca JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91:3166–3170

Durrett R, Kruglyak S (1999) A new stochastic model of microsatellite evolution. J Appl Prob 36:621–631

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445

Feldman MW, Bergman A, Pollock DD, Goldstein DB (1997) Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics 145:207–216

Garza JC, Slatkin M, Freimer NB (1995) Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol Biol Evol 12:594–603

Ishii T, Mori N, Ogihara Y (2001) Evaluation of allelic diversity at chloroplast microsatellite loci among common wheat and its ancestral species. Theor Appl Genet 103:896–904

Jakobsson M, Säll T, Lind-Halldén C, Halldén C (2006a) The evolutionary history of the common chloroplast genome of *Arabidopsis thaliana* and *A. suecica*. J Evol Biol. DOI 10.1111/j.1420-9101.2006.01217.x

Jakobsson M, Hagenblad J, Tavaré S, Säll T, Halldén C, Lind-Halldén C, Nordborg M (2006b) A unique recent origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. Mol Biol Evol 23:1217–1231

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

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

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (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

Li W-H (1997) Molecular evolution. Sinauer Associates Inc. Publishers, Sunderland

Matsuoka Y, Mori N, Kawahara T (2005) Genealogical use of chloroplast DNA variation for intraspecific studies of *Aegilops taushcii* Coss. Theor Appl Genet 111:265–271

Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10:72–80

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

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

Palmer JD (1987) Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. Am Nat 130:S6–S29

Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA (1995) Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. Proc Natl Acad Sci USA 92:7759–7763

Provan J (2000) Novel chloroplast microsatellites reveal cytoplasmic variation in *Arabidopsis thaliana*. Mol Ecol 9:2183–2185

Provan J, Corbett G, Waugh R, McNicol JW, Morgante M, Powell W (1996) DNA fingerprints of rice (Oryza sativa) obtained from hypervariable chloroplast simple sequence repeats. Proc R Soc Lond B 263:1275–1281

Provan J, Soranzo N, Wilson NJ, McNicol JW, Forrest GI, Cottrell J, Powell W (1998) Gene-pool variation in Caledonian and Scots pine (*Pinus sylvestris* L.) revealed by chloroplast simple sequence repeats. Proc R Soc Lond B 265:1697–1705

Provan J, Soranzo N, Wilson NJ, Goldstein DB, Powell W (1999) A low mutation rate for chloroplast microsatellites. Genetics 153:943–946

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142–147

Rice P, Longden I, Bleasby A (2000) Emboss: the European molecular biology open software unit. Trends Genet 16:276–277

Rose O, Falush D (1998) A threshold size for microsatellite expansion. Mol Biol Evol 15:613–615

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

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res 6:283–290

Sibly RM, Whittaker JC, Talbot M (2001) A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. Mol Biol Evol 8:413–417

Sibly RM, Meade A, Boxall N, Wilkinson MJ, Corne DW, Whittaker JC (2003) The structure of interrupted human AC microsatellites. Mol Biol Evol 20:453–459

Sokal RR, Rohlf FJ (1995) Biometry, 3rd edn. Freeman, New York

Symonds VV, Lloyd AM (2003) An analysis of microsatellite loci in *Arabidopsis thaliana*: mutational dynamics and application. Genetics 165:1475–1488

Vendramin GG, Degen B, Petit RJ, Anzidel M, Madaghiele A, Ziegenhagen B (1999) High levels of variation at *Abies alba* chloroplast microsatellite loci in Europe. Mol Ecol 8:1117–1112

Weber JL (1990) Informativeness of human $(dC\text{-}dA)_n$ $(dG\text{-}dT)_n$ polymorphisms. Genomics 7:524–530

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 8:1123–1128

Xu DH, Abe J, Gai JH, Shimamoto Y (2002) Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. Theor Appl Genet 105:645–653