MICROSATELLITES: SIMPLE SEQUENCES WITH COMPLEX EVOLUTION

Hans Ellegren

Few genetic markers, if any, have found such widespread use as microsatellites, or simple/short tandem repeats. Features such as hypervariability and ubiquitous occurrence explain their usefulness, but these features also pose several questions. For example, why are microsatellites so abundant, why are they so polymorphic and by what mechanism do they mutate? Most importantly, what governs the intricate balance between the frequent genesis and expansion of simple repetitive arrays, and the fact that microsatellite repeats rarely reach appreciable lengths? In other words, how do microsatellites evolve?

Assuming chance association of nucleotides, the probability of finding the sequence CACACACACACACA-CACACA more than once in the human genome is negligible. However, perfect or near-perfect tandem iterations of short sequence motifs of this kind are extremely common in eukaryotic genomes and, in the case of the human genome, they are found at hundreds of thousands of places along chromosomes. This particular genomic feature is not restricted to (CA)n repeats — every possible motif of mono-, di-, tri- and tetrancleotide repeats is vastly overrepresented in the genome. Ever since their discovery in the early 1980s, the ubiquitous occurrence of microsatellites — also referred to as short tandem repeats (STRs) or simple sequence repeats (SSRs) — has puzzled geneticists. Why are they so common? Do they fulfill some function or are they simply junk DNA sequences that should perhaps be viewed as 'selfish DNA'? Addressing these questions is important if we wish to understand how genomes are organized and why most genomes are filled with sequences other than genes.

Microsatellites are among the most variable types of DNA sequence in the genome. In contrast to unique DNA, microsatellite polymorphisms derive mainly from variability in length rather than in the primary sequence. Moreover, genetic variation at many microsatellite loci is characterized by high heterozygosity and the presence of multiple alleles, which is in sharp contrast to unique DNA. With the advent of PCR in the late 1980s, the analysis and genotyping of microsatellite polymorphisms became straightforward (see timeline). Microsatellites quickly became the marker of choice in genome mapping, and subsequently also in population genetics studies and related areas.

For a neutral marker, the degree of polymorphism is proportional to the underlying rate of mutation. Given the extensive polymorphism of microsatellites, it follows that mutations must occur frequently — an assumption that is supported by direct observations. The rate and direction of mutations constitute two basic factors in the estimation of genetic distance on the basis of microsatellite data. By applying theoretical models of microsatellite evolution to empirical data, population geneticists attempt to, for example, determine how long ago two populations diverged, or measure the amount of gene flow between populations. However, despite the extensive use of microsatellite markers over the past 15 years, it is clear that many theoretical models fail to accurately explain allele frequency distributions in natural populations. Importantly, it seems that microsatellite evolution is a far more complex process than was previously thought. A deeper understanding of the evolutionary and mutational properties of microsatellites is therefore needed, not...
REVIEWS

The early history of microsatellites


<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1962</td>
<td>Demonstration of extensive length variability of tandem repetitive DNA as revealed by DNA fingerprinting of minisatellites</td>
</tr>
<tr>
<td>1981</td>
<td>Development of PCR-based microsatellite genotyping</td>
</tr>
<tr>
<td>1982</td>
<td>Microsatellites used to derive the first detailed map of the human genome</td>
</tr>
<tr>
<td>1985</td>
<td>Fine-scale analysis of the genetic relationships among human populations made possible by microsatellites</td>
</tr>
<tr>
<td>1986</td>
<td>Large numbers of microsatellite mutations identified from pedigree analysis in humans</td>
</tr>
<tr>
<td>1989</td>
<td>Regions of cryptic simplicity identified as an important source of genetic variation</td>
</tr>
<tr>
<td>1991</td>
<td>Microsatellites introduced for studies of natural populations</td>
</tr>
<tr>
<td>1992</td>
<td>Microsatellites used to map the human genome</td>
</tr>
<tr>
<td>1993</td>
<td>Fine-scale analysis of the genetic relationships among human populations made possible by microsatellites</td>
</tr>
<tr>
<td>1994</td>
<td>Large numbers of microsatellite mutations identified from pedigree analysis in humans</td>
</tr>
</tbody>
</table>

*The importance of these findings seemed to be insignificant, as laborious cloning and sequencing procedures were required to analyse the polymorphisms.*

The genome biology of microsatellites

**What is a microsatellite?** Genomes are scattered with simple repeats. Tandem repeats occur in the form of iterations of repeat units of almost anything from a single base pair to thousands of base pairs. Mono-, di-, tri- and tetranucleotide repeats are the main types of microsatellite, but repeats of five (penta-) or six (hexa-) nucleotides are usually classified as microsatellites as well. Repeats of longer units form minisatellites or, in the extreme case, satellite DNA. The term satellite DNA originates from the observation in the 1960s of a fraction of sheared DNA that showed a distinct buoyant density, detectable as a ‘satellite peak’ in *density gradient centrifugation*, and that was subsequently identified as large centromeric tandem repeats. When shorter (10–30 bp) tandem repeats were later identified, they came to be known as minisatellites. Finally, with the discovery of tandem iterations of simple sequence motifs, the term microsatellites was coined. The difference between the terms micro- and minisatellites might not be obvious *per se*, but it is motivated by the difference in the mutational mechanisms of repeats of just a few nucleotides and of ten or more (see below).

It is more complicated to define the minimum number of iterations needed for a repetitive sequence to be referred to as a microsatellite. For instance, the sequence CACA occurs frequently in the human genome: should it be seen as (CA), microsatellites or just as unique sequence? In practice, the threshold that is used when describing the occurrence of a microsatellite in a genomic sample data set must be specified. Unfortunately, no real consensus has been reached on this matter; whereas some use a minimum number of base pairs, others use a minimum number of repeat units, and in both cases, the numbers have varied. The issue is further complicated by the lack of agreement on how much degeneracy should be accepted for characterizing a slightly imperfect tandem repetitive sequence as a microsatellite. Mismatch considerations are particularly important when using algorithms (such as RepeatMasker, Sputnik and Tandem Repeats Finder; see online links box and BOX 1) to search large genomic sequences for repeats.

It is appropriate to further classify microsatellites according to their association with coding sequence as this is related to the mutational and selective forces that operate on different types of repeat. The bulk of simple repeats are embedded in non-coding DNA, either in the intergenic sequence or in the introns. Microsatellites that are used as genetic markers are usually of this type and are generally assumed to evolve neutrally. Their frequency and distribution should therefore reflect the underlying mutation process. In coding DNA, selection against frameshift mutations effectively hinders the expansion of everything other than trinucleotide repeats, for which there might be further length constraints related to protein function. Trinucleotide repeats associated with human disease comprise a special class of microsatellites in coding DNA. These loci undergo extensive repeat expansions, the mutational mechanism of which is thought to differ from that of most microsatellites in the genome. For instance, the establishment of hairpin structures with a relatively high amount of base-pair complementarities might stabilize loops that are generated during replication slippage.
Details of the evolution of expanded trinucleotide repeats have been described elsewhere and will not be considered further here.

**Microsatellite distribution.** The initial analysis of the draft sequence of the human genome concluded that microsatellites account for 3% of the genome1. There are more than one million microsatellite loci in the human genome, although the exact number greatly depends on the parameters of the search algorithm (for example, gap and mismatch penalties). This number also includes an appreciable proportion of interrupted microsatellites and many that are probably monomorphic. Dinucleotide repeats dominate, followed by mono- and tetrانucleotide repeats, and trinucleotide repeats are least dominant. Again, however, it is a matter of how microsatellites are defined. Among repeats that are at least 12 bp long, mononucleotide repeats outnumber dinucleotide repeats; the reverse situation is not valid until a higher threshold is used. Among dinucleotides, (CA)n repeats are most frequent, followed by (AT)n, (GA)n, and (GC)n, the last type of repeat being rare. Note that there are only four possible types of dinucleotide repeat, because CA = AC = GT = TG, GA = AG = CT = TC, AT = TA, and GC = CG.

Data from the mouse genome have confirmed the abundance of microsatellites but have also revealed impressive differences4. If identical search criteria are used, the mouse genome proves to be repeat-rich with two–threefold more microsatellites than humans. Moreover, microsatellites are longer in mice than in humans, and the same holds true for the rat–human comparison9. Preliminary data from other mammalian genomes indicate that rodent genomes have particularly high microsatellite numbers. This might be a general phenomenon—that microsatellite occurrence differs between related species. In fact, differences might even occur between such closely related species as humans and chimpanzees16, and within the genus Drosophila1,2,12.

Microsatellite density tends to positively correlate with genome size13–15. Among fully sequenced eukaryotic genomes, microsatellite density is highest in mammals. However, in plants, microsatellite frequency is negatively correlated with genome size16. This has been attributed to the fact that microsatellites are underrepresented in the repetitive parts of the plant genome that are involved in genome expansion, such as the long terminal repeats of RETROTRANSPOSONS16. Another peculiar feature of most plant genomes is that (AT)„ is the most common motif among dinucleotides17. Assuming that, on a genomic scale, microsatellite sequences are at equilibrium, the contrasting distributions of microsatellite motifs in different genomes strongly indicate that there is interspecific variation in the mechanisms of mutation or repair of specific motifs. Alternatively, there might be variation in the selective constraints that are associated with different microsatellite motifs.

Are microsatellites equally common everywhere in the genome? There seem to be no distinct differences in density between intergenic regions and introns18. Base composition influences microsatellite density, which is consistent with their neutral origin and random generation by mutation18. There is, however, evidence for regional variation in microsatellite frequency that cannot be explained by base composition18, and, in the human and mouse genomes, microsatellite density is nearly twofold higher near the ends of chromosome arms1. What accounts for this heterogeneity remains to be explained. In several species, the density and/or the length distribution of microsatellites on the X chromosome differs from that on the autosomes18. It might be a result of factors such as sex differences in the mutation rate, differences in EFFECTIVE POPULATION SIZE between the X chromosome and autosomes, and the efficiency of selection on hemizygous chromosomes.

Microsatellites are also frequently found in the proximity of interspersed repetitive elements such as short interspersed repeats (SINEs) and long interspersed elements (LINEs). For example, human Alu repeats often have a microsatellite-like structure at their 3′ ends19 that might arise from the introduction of poly(A) tails of reversed transcribed messages when element insertion takes place. This is consistent with the observation that mononucleotide arrays (A)n and other types of A-rich microsatellite dominate at these sites. Other examples include the intimate association of microsatellites with

---

**Box 1 | Informatics approaches to finding microsatellites in a genomic data set**

**Sputnik**

Sputnik uses a recursive algorithm to search for repeats of two–five nucleotides in sequence files in FASTA format. Insertions, mismatches and deletions are tolerated but affect the overall score. If the score falls below a cutoff threshold, the search is abandoned and begun again at the next nucleotide. Sputnik does not compute an entire identity matrix first and then pick the best of the hits; instead, it starts at the beginning and compares the patterns until the score falls below a cutoff threshold.

**RepeatMasker**

RepeatMasker does not use a recursive algorithm. It scans for di- to pentameric repeats and simple repeats that are shorter than 20 bp, and those with >10% divergence from a perfect repeat are ignored.

**Tandem Repeats Finder**

This program works without the need to specify either the pattern or the pattern size. It models tandem repeats according to the percentage identity and frequency of indels (insertions/deletions) between adjacent pattern copies and uses statistically-based recognition criteria. The program can return a copy of the original sequence with the tandem repeats masked out.
retrotransposon-like elements in barley, and (AT)_n microsatellites that are frequently found to be juxtaposed with miniature inverted repeat-transposable elements (Micropon-4) in rice. In some of these cases, microsatellites have evolved from internal A-rich structures, and it is also possible that insertion of an interspersed element might in itself be favoured at sites with a pre-existing microsatellite.

Microsatellites are present in low numbers in prokaryotes. This is particularly true for longer repeats, for which the numbers are lower than would be expected on the basis of nucleotide composition, in sharp contrast to the situation in eukaryotic genomes. Even short prokaryotic microsatellites might still vary in length. Unusually long microsatellites are sometimes associated with virulence factors, in which case, they act as translation and transcriptional ‘switches’; therefore, their presence is maintained by positive selection.

**The mutation process**

**Mutation models.** A mutation model of microsatellite evolution is needed if allele frequency data from two groups of individuals (for example, populations or species) are to be used for estimating the genetic distance between them. A wide range of models of the evolutionary dynamics of microsatellites has been presented, most of which derive from the stepwise mutation model (SMM) (FIG. 1). Adopted for microsatellites, the original SMM postulates that a mutation alters the length of a repetitive array through the addition or removal of one repeat unit at a fixed rate (a symmetric forward–backward random walk that is independent of repeat length). However, it soon became apparent that a simple SMM does not lead to stationary microsatellite-length distributions. For example, the fact that microsatellites seem to show an upper size limit is incompatible with the SMM. Extensions of the SMM have therefore introduced an upper limit on allele sizes, or a mutational bias such that large alleles mutate preferentially to alleles of smaller sizes.

Other approaches have involved more complex stepwise models. The parameters of the mathematical models are tested against measures of variability (heterozygosity, variance of repeat counts, skewness) that are observed within populations, and, more recently, against microsatellite distributions in genomic data sets.

An attractive model of microsatellite evolution holds that a genome-wide distribution of microsatellite repeat length that is at equilibrium results from a balance between length and point mutations. According to this model, two opposing mutational forces operate on microsatellite sequences. Length mutations, the rate of which increases with increasing repeat count, favour loci to attain arbitrarily high values, whereas point mutations break long repeat arrays into smaller units. At equilibrium, there will be a steady-state distribution of repeat lengths governed by the rate of length mutation and the rate of point mutation. This model, or derivatives thereof, has been well received in recent years because it can explain differences in microsatellite

**Figure 1 | Microsatellite mutation models.** The magnitude and direction of mutation events according to different forms of the stepwise mutation model. Mutations are indicated by the change in the number of repeat units; for example, +1 is an expansion of one repeat unit. **a** | A simple model that only involves one-step changes. **b** | A model that involves multi-step changes. **c** | A model with directionality in favour of repeat expansions. **d** | A length-dependent model in which short alleles tend to increase in size, whereas long alleles show a bias towards contraction.

**SKENNESS** Deviation from the normal distribution.
distribution among species and provides an elegant solution to the problem of why microsatellites do not expand into enormous arrays. However, even with this model, evolutionary dating of divergence times is not necessarily trivial.45

**Mutation mechanism.** Length changes in microsatellite DNA are generally thought to arise from replication slippage — that is, transient dissociation of the replicating DNA strands followed by misaligned reassociation at the nascent strand (Fig. 2a). When the nascent strand realigns out of register, renewed replication will lead to the insertion or deletion of repeat units relative to the template strand. Most of these primary mutations are corrected by the MISMATCH-REPAIR SYSTEM, and only the small fraction that was not repaired ends up as microsatellite mutation events.46

*In vitro* experiments that use purified eukaryotic or prokaryotic enzymes confirm that DNA polymerase is the only enzymatic activity needed for slippage.48 Slippage involves DNA polymerase pausing, during which the polymerase dissociates from the DNA. On dissociation, only the terminal portion of the newly synthesized strand separates from the template and subsequently anneals to another repeat unit.49

**Recombination-like processes that involve unequal crossover** or **GENE CONVERSION introduce mutations in the larger minisatellite sequences.**53 There is little evidence that recombination would also contribute to microsatellite mutations. Genomic microsatellite distributions are associated with sites of recombination,54 most probably as a consequence of repetitive sequences being involved in recombination rather than being a consequence of it.55 Moreover, most tests for a correlation between recombination rate and microsatellite density or mutability have failed to demonstrate such an effect.18,56

**Character of observed mutations.** Although improved mutation models are now available, it can be difficult to assess to what extent they reflect true evolutionary processes. Fortunately, the high rate of mutation at microsatellite loci makes it possible to observe mutation events directly. Specifically, pedigree analysis offers a means for mutation detection (BOX 2), and data on *de novo* mutations have now been reported for a range of loci and organisms.53,54 The general pattern that emerges is compatible with replication slippage in which

---

**Fig. 2 | Replication slippage.** a | After the replication of a repeat tract has been initiated, the two strands might dissociate. If the nascent strand then realigns out of register, continued replication will lead to a different length from the template strand. If misalignment introduced a loop on the nascent strand, the end result would be an increase in repeat length. A loop that is formed in the template strand leads to a decrease in repeat length. b | Replication slippage also occurs during *in vitro* amplification of microsatellites, in this case mainly in the form of repeat contractions. These events can be recognized as minor peaks — known as stutter bands — that differ from each main product by multiples of the repeat unit length.
new variants differ from their progenitor alleles by integral numbers of repeats. The fact that mutations sometimes involve more than one repeat unit means that the single-step mutation model is not valid in most cases.

One important conclusion from observations of spontaneous mutations is that the mutation process seems to be heterogeneous with respect to loci, repeat types and organisms. For instance, most human studies find that <15% of mutation events are multi-step changes. However, the three largest human studies of this kind that have so far been presented reveal contradictory results. In agreement with other studies, Ellegren and Xu et al. found 11–14% multi-step mutations among 102 and 236 mutation events, respectively. By contrast, Huang et al., in an analysis of 97 mutation events, reported 63% multi-step changes. What accounts for this discrepancy remains unclear, but it might indicate that some loci are more prone to large changes than others. Analyses of individual loci in other organisms have revealed highly variable proportions of multi-step changes, in the range of 5–75% (REFS 62,63,71–73). A more extensive screening of zebrafish markers found 68% multi-step changes.

Heterogeneity is also seen in the propensity of mutation events to lead to different forms of alterations in microsatellite size. Directionality in the mutation process in favour of gains over losses has been observed for many human markers and for bird microsatellites. However, Xu et al. found no such bias, and Huang et al. found only a modest excess of contractions in their studies of human microsatellites. Whatever the cause of this heterogeneity, it will be interesting to see whether directionality is related to microsatellite length at the level of individual loci. Everything else being equal, we should expect loci that have a tendency to expand by mutation to grow more often than those that tend to contract. To add to the complexity, several studies have found evidence of a negative correlation between direction/magnitude of mutation and allele size — that is, long alleles being biased towards contraction. If generally true, this would offer a mechanistic explanation for the stationary genomic length distributions seen at microsatellite loci. Interestingly, mutations from three bacterial species show a downward bias, which can perhaps account for the rarity of microsatellites in prokaryotic genomes.

**Box 2 | Using pedigrees to detect microsatellite mutations**

The most straightforward approach for the study of microsatellite mutations is direct observations of allele transmissions in parent–child pairs (see figure). A mutant allele (asterisk) is identified as being incompatible with Mendelian inheritance because it is different in size from parents’ alleles. Data from other markers are needed to confirm that non-congruence between parental and offspring genotypes is not a result of incorrect parentage. One limitation of this approach is that detectable mutations are restricted to cases in which child genotypes cannot be generated by transmission from parents’ genotypes. A mutation to a character state that is already present in one of the parents can therefore remain undetected. However, likelihood-based estimates of mutation rates that take such events into account have been developed.

A related concern derives from the fact that even when the mutant allele is different from parents’ alleles, it is not evident from which progenitor allele it originates. The standard assumption made is that the smallest mutational change is the most probable. However, this introduces circularity when such data are used to support the stepwise mutation model. Fortunately, simulations show that this assumption is valid in most cases, at least when compared with the alternative of random assignment to the progenitor allele.

An alternative approach for the identification of germline mutations is offered by the analysis of single sperm, or small pools of sperm. This gives virtually unlimited access to gametes from specific males to screen, allowing individual-specific estimates of the mutation rate to be made. However, a disadvantage with this approach is that it is technically demanding to amplify such small amounts of DNA, and that it therefore requires most stringent laboratory routines to avoid contamination. On the other hand, as it does not require access to pedigrees, it might be useful for applications such as genetic toxicology screenings.
Figure 3 | Microsatellite mutation rates in the human genome. Observed sex-average mutation rates (log scale, with 95% confidence intervals) for human microsatellites obtained from pedigree analysis. Markers are given in the following order: 1, D10S1214; 2, D12S1090; 3, ACTBP2; 4, D19S253; 5, D9S302; 6, D3S1744; 7, FGA; 8, HUMWGA31; 9, D22S883; 10, D18S51; 11, D8S1179; 12, D21S11; 13, CYP19; 14, D3S1358; 15, HUMCSF1PO; 16, D18S849; 17, D13S317; 18, D17S5; 19, D5S818; 20, D7S820; 21, Penta E; 22, D16S359; 23, HUMFESFPS; 24, HUMF13A01; 25, HUMILPOL; 26, D1S80; 27, HUMF13B; 28, D2S1338; 29, HUMTH01; and 30, HUMTPox. Data from The Annual Report Summary for 2000 from the US Parentage Testing Standards Committee.

Sequence data. Another empirical approach to studying microsatellite evolution involves characterizing the sequence structure of alleles within species, or comparing the sequence of orthologous loci in different species. Using this approach, the effect of mutations accumulated over evolutionary timescales can be readily studied, although it might be difficult to determine the precise order and character of individual mutation events if they have occurred at high rates during the time period being surveyed. Note that in contrast to base substitutions, for which an infinite allele model is applicable, microsatellite alleles are often identical in state (structure) but not by descent86. For example, two chromosomes that are drawn from a population with microsatellites being much shorter in flies than in vertebrates87. Within species, measures of repeat lengths correlate with mutability88. Furthermore, a positive correlation between allele size and mutation rate has been found in many organisms89,90. The single most important factor to affect mutation rate that has so far been discovered is microsatellite length — mutation rate increases with an increasing number of repeat units. Intuitively, this seems understandable — more repeat units give more opportunities for replication slippage. A length-dependent mutation rate explains part of the mutation-rate variation at several scales. The low mutation rate in Drosophila melanogaster is compatible with microsatellites being much shorter in flies than in vertebrates.87. Within species, measures of repeat lengths correlate with mutability88. Furthermore, a positive correlation between allele size and mutation rate has been seen in many organisms89,90. The precise character of the relationship between repeat count and mutation rate is less clear. Some studies have found a linear relationship89, whereas some recent data indicate a power or exponential relationship between size and rate89,90.

Mutation-rate variation

There is no uniform microsatellite mutation rate; the rates differ among loci and among alleles, and, perhaps as a consequence, also among species84. The single most important factor to affect mutation rate that has so far been discovered is microsatellite length — mutation rate increases with an increasing number of repeat units. Intuitively, this seems understandable — more repeat units give more opportunities for replication slippage. A length-dependent mutation rate explains part of the mutation-rate variation at several scales. The low mutation rate in Drosophila melanogaster is compatible with microsatellites being much shorter in flies than in vertebrates87. Within species, measures of repeat lengths correlate with mutability88. Furthermore, a positive correlation between allele size and mutation rate has been seen in many organisms89,90. The precise character of the relationship between repeat count and mutation rate is less clear. Some studies have found a linear relationship89, whereas some recent data indicate a power or exponential relationship between size and rate89,90.

But repeat length is not the sole cause of microsatellite mutation-rate variation. As can be seen in Fig. 3, the mutation rate of individual loci in a selected set of human markers varies within two orders of magnitude, which cannot be attributed to repeat length. Similar observations are made in other organisms90,91,92,93. One important consequence of this variation is that the mean mutation rate of a set of markers will vary a lot depending on which particular markers are used. Moreover, as most markers that are used in genetic studies are selected on the basis of being (highly) polymorphic and,
given the expected relationship between polymorphism and mutability, the observed rates might not provide a representative picture for the genome as a whole.

What else matters? One possibility is that sequences that flank the microsatellite affect the mutation rate. That inherent characteristics of individual loci are involved is indicated by covariation in levels of variability at orthologous loci in related species. However, in addition to a flanking-sequence effect, this observation could also be compatible with an effect of, for example, transcription-coupled repair, chromatin structure, regional sequence context and local point-mutation-rate variation. As for the last possibility, an extension of the balance model of microsatellite evolution states that not only will the equilibrium length distribution of simple repeats be dependent on the species-specific rate of point mutation, but the length of individual repeat loci will also depend on the local point-mutation rate, for which there is evidence for significant heterogeneity within genomes. In a study of orthologous microsatellite loci in the mouse and rat, a negative correlation between microsatellite length and substitution rate in nearby flanking sequence was found. This would indicate that when point mutations occur at high frequency, they seem to hinder further microsatellite expansion (mutability) by introducing interruptions or imperfections in the repeat array.

Assuming a replication origin of microsatellite mutations, we should expect the mutation rate to correlate with the number of germline cell divisions. By extension, mutations should be more frequent in males than in females, and in older males than in younger males. Although seemingly reasonable, these predictions are only partly supported by empirical data. Two of the large studies of human mutations find three–four times more mutations in men, which is close to recent data on the male-to-female mutation-rate ratio for point mutations. However, Huang et al. saw no sex bias in microsatellite mutation rate in their study. Moreover, significant variation in the mutational sex-bias has been documented for swallow microsatellites, with at least one locus showing a male-biased rate, whereas others have female-biased rates. Female-biased rates for individual loci have been reported for other organisms as well. Attempts to correlate human microsatellite mutation rates with the father’s age have either failed to find such an effect or only found a small effect.

**Experimental approaches**

The empirical data on microsatellite mutations described above all refer to spontaneous events observed after germline transmissions. An experimental approach to the study of microsatellite evolution is offered by the analysis of instability of artificial plasmid-borne microsatellite sequences introduced into bacterial or eukaryotic cells. By constructing plasmids with repeats that are associated with a resistance or a reporter gene, length mutations that disrupt or restore the reading frame of that gene can easily be monitored. Mutation profiles have in this way been particularly well characterized in the yeast *Saccharomyces cerevisiae*, in which repeats that are integrated into chromosomes have also been studied.

These studies confirm several observations from germline transmissions. Mismatch repair is identified as crucial to microsatellite stability as mutation rates in prokaryotic and eukaryotic cells that are deficient in mismatch repair are increased by several orders of magnitude compared with wild-type cells. Mutations in genes that encode proof-reading exonucleases and some DNA polymerases have also been implicated in repeat instability, although they have a more modest effect compared with mismatch-repair deficiency. In all systems, mutation rate increases with repeat length, but interruptions stabilize repeat tracts. The orientation of repeats—that is, whether a particular motif is on the coding or the complementary strand—does not seem to affect the mutation rate, with the exception of long trinucleotide repeat arrays. Observations of the destabilization of microsatellites by elevated levels of transcription would support a role for transcription-coupled repair.

The effect of sequence composition on the relative instability of repeats is less clear. A study in *Escherichia coli* found no significant difference in the mutability of CA- and GA-repeats of similar length, in contrast to observations in human cells (see also Ref. 92). Conflicting observations are also made with respect to the effect of the length of the repeat unit, as seen, for example, in di- versus tetranucleotide repeats. However, in *E. coli* and in human and yeast cells, the mutability of G-monomonucleotide repeats is higher than that of A-repeats of the same length, potentially owing to stronger stacking interactions among Gs or Cs than among As and Ts.

Insertions generally outnumber deletions in eukaryotic cells, whereas the opposite is true in *E. coli*. In both cases, large deletions are frequently seen in long repeat tracts. In general, at least two explanations might account for observations of a directional bias in microsatellite mutation. The primary rate of slippage mutation might be higher for insertions than for deletions. Displaced loops might be more easily introduced in the newly synthesized strand (which results in an insertion) than in the template strand. Alternatively, mismatch repair might more easily recognize or more efficiently repair displaced loops on the template strand than on the nascent strand. That mismatch repair is involved in a directional bias is indicated by the fact that mutations in some mismatch-repair genes, such as yeast *MSH3*, differentially affect the rate of insertions and deletions. In *D. melanogaster*, mismatch repair preferentially recognizes and/or corrects primary expansion mutations to leave an excess of contractions and, generally, (AT) mutations are repaired more efficiently than (GT) changes. If the character of mismatch repair differs between groups of organisms, we might expect consequent differences in microsatellite frequency.
The future of microsatellites

The evolutionary process of simple repeats is far from simple. One important implication of the complexity of microsatellite evolution is, therefore, that care needs to be taken when using microsatellite data in population genetics studies. For instance, significant mutation-rate heterogeneity among loci means that it might be difficult to translate estimates of genetic distance into absolute timescales. Similarly, directional biases in the mutation process have important consequences for the interpretation of differences in allele size distributions among species, particularly if the character of the bias differs among species. Future mathematical models of microsatellite evolution should therefore aim to incorporate as many of the different forms of mutational heterogeneity as possible. Those who use microsatellites in population genetics studies should select only the markers that are well characterized in terms of mutational properties (mutation rates, directionality, whether all alleles of equal length are identical in sequence), and, preferably, use markers that show uniform rates and patterns. Alternatively, but in many species less realistically, the use of many markers might compensate for heterogeneity in mutual properties among loci.

Microsatellites continue to find their application in areas such as linkage mapping, paternity testing, forensics and for the inference of demographic processes. More recently, they have found most use in linkage-disequilibrium mapping studies, in which associations between markers and trait loci are searched for in population samples, and in hitchhiking mapping, in which genome-wide screens for regions that show signs of selection are made. But there are also prospects for new applications. Given their high mutation rate, microsatellites offer a realistic means to study how the overall genomic mutation rate is affected by environmental factors (genetic toxicology). Elevated rates of microsatellite mutations in the germline have been seen in animals and plants that are exposed to ionizing radiation, and similar observations have been made for minisatellites in humans. Estimating microsatellite mutation rates in samples that are exposed to different forms of radiation or toxic compounds could, when properly set in relation to data from control groups, help to make risk assessments.

Offers a suggestion for how variation in microsatellite length might relate to point-mutation-rate heterogeneity.


This paper, and references 128 and 129, introduce the use of PCR for genotyping microsatellites.


Acknowledgements
The author would like to acknowledge two particularly helpful reviewers who provided useful comments on the manuscript. The author’s work was supported in part by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning.

Competing interests statement
The author declares that he has no competing financial interests.

Online links

DATABASES
The following terms in this article are linked online to: Entrez: http://www.ncbi.nlm.nih.gov/Entrez

FURTHER INFORMATION
RepeatMasker: http://www.repeatmasker.org
Sputnik: http://espressosoftware.com/pages/sputnik.jsp
Tandem Repeats Finder: http://c3.biomath.mssm.edu/trf.html
Access to this links box is available online.