Microsatellite variation in North American populations of *Drosophila melanogaster*

David B. Goldstein* and Andrew G. Clark

Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

Received July 5, 1995; Revised and Accepted August 31, 1995

ABSTRACT

Computer database searching for microsatellites can be particularly effective for organisms like *Drosophila melanogaster* for which there are extensive sequence data. Here we demonstrate that 17 out of 18 such microsatellites are also highly polymorphic in natural populations of *Drosophila*, and that this variation is easily scorable with PCR followed by electrophoresis on high-resolution agarose. This form of variation is likely to be of great value in studies of the genomic distribution of polymorphism, population structure, the relation between intraspecific polymorphism and inter-specific divergence and the mutation rate and pattern of mutations of microsatellites. In this preliminary survey of 15 lines, we find that the variance in repeat count is most strongly correlated with the maximum count, that perfect repeats are significantly more variable than imperfect repeats and that repeats which are split by an imperfection have unexpectedly low variance given the size of the perfectly repeated portion.

INTRODUCTION

There are between 50 000 and 100 000 dinucleotide microsatellites in the human genome, most of which are likely to be polymorphic (1-3). While the density varies considerably among taxa (4), it seems likely that all eukaryotic genomes hold huge numbers of microsatellites and that many if not most are polymorphic (5). The consistently high levels of variation at these loci is certainly due to their high mutation rate, but almost all other aspects of the mutation process remain to be resolved: (i) What are the relations between repeat motif and number of repetitions and the rate and type of mutation? (ii) is polymerase slippage the predominant mechanism of mutation, and what role does repair play? (iii) are mutations generally stepwise? and (iv) is the mutation process alone sufficient to explain observed variation at microsatellite loci, or must selection be invoked?

A few studies have been published in which allele frequencies from natural populations were compared with predictions based on variations of the stepwise mutation model (6) or with the infinite alleles model (7). Most of these studies have suggested that microsatellite mutations fit a stepwise model (3,8), but at least one supported the infinite alleles model (9). Direct studies of mutations have been more clear. Studies based both on artificial constructs in yeast (10) and analyses of human pedigrees

(11) have shown that most mutations involve the addition or subtraction of a small number of repeat units.

With respect to the cause of mutation, an *in vitro* study demonstrated that polymerase slippage rates are sufficiently high to account for observed mutation rates (12) and at least some of the mutations observed in pedigrees are inconsistent with any kind of homologous exchange (13), further implicating polymerase slippage. Schloetterer and Tautz's (12) study of *in vitro* polymerase behavior also showed that the rate of slippage depends on the sequence of the repeated motif, but not on the number of repetitions of the motif (repeat count). This is in contrast to the phenomenon of 'dynamic mutation' (14,15) in which the probability of mutation evolves along with the repeat count. Such a dependence has been clearly demonstrated in the case of disease-causing trinucleotide repeats, but its generality is not known.

Surveys of population diversity have been similarly ambiguous on the question of the relation between repeat count and mutation rate. Several have found a weak, but positive correlation between some measure of diversity and repeat count (2,16,17), but the largest such study, based on 108 loci, showed no correlation (3).

Some of these contradictions probably result from a general lack of power to infer complexities of the mutation process. Those studies based on pedigrees involve a very small number of mutations and certainly reveal only the most common type of mutation. *In vitro* studies may be more powerful, but their relevance is unknown. Population studies which use fragment sizes directly (e.g. 3) may fail to uncover relationships between diversity and repeat count because the variability (from locus to locus) in the number of base pairs separating the primers from the microsatellite region is often as great or greater than the variability in average repeat count. This variation introduces a substantial amount of noise that may obscure underlying relationships. Another difficulty with the use of population variation is that variability will be directly related to the mutation rate only at mutation-drift equilibrium. Genome specific effects (selection for or against linked sites) will influence the variability of microsatellites, again obscuring the relationship between the mutation process and variability. Despite these complications, population variation is a powerful tool for inferring the basic properties of microsatellite loci and can be especially revealing in those cases where diversity levels can be compared directly with the microsatellite sequence. Here we report the analysis of 18 microsatellite loci in North American populations of *Drosophila melanogaster*. Because the loci were identified from GenBank sequences we are able to relate diversity levels directly to repeat count.

* To whom correspondence should be addressed

Table 1. Loci and oligonucleotide primers

Locus	GenBank name	Map site	Location	Primers
<i>zeste-white 3</i>	DMSGG3	1-1.3	3' UTR	TCCAGCAATCACAGCAAACCT TCTTTTGAAATTGCGGTGA
<i>zeste-white 3</i>	3DMZW3K25	1-1.3	3'	ATTGTCATTTTATTGCTGCCG AACGAAGAGAGTTGCGGAGA
<i>otu</i>	DROOTUA	1-22.7	3' UTR	TCACGAAATCAAAGAACCACC GTCTCTAGCCGAACAGCGAG
<i>sevenless</i>	DROSEV	1-33.4	intron	AATTTGTTATGGCCACTTCCC CATGGCCAATAAGTCACGC
<i>tenA</i>	DMTENA	1-38	3' UTR	CTCTTAGTGCGCAGGGATTCT GAGTCGCTCAATGGCAGG
<i>troponin</i>	DMTROPONI	1-57.6	intron	CAAGAGATCCCAGAGAGAGAGA ACGTGTGCGTGTGTTTCTC
<i>expanded</i>	DROEXPAND	2-0.1	expressed	GTGATCGATCCCGCTGTC TCCGGTTTCCAATTAGCTTG
<i>yan</i>	DROYANETSB	2-[4]	intron	TAATGGGGAATGGGTGAATG GCCGTGCTCTTCTCTTACG
<i>myosin heavy chain</i>	DROMHC	2-52	intron	AAACCCACACAACAACCTGCA GACATTACCGATATTGGATGCA
<i>mp20</i>	DMMP20	2-68	5' UTR	CATGCAAATGAGCAGTACTTTG TATTTTCACACATTTCCAATCG
<i>mastermind</i>	DMASTER	2-70.3	expressed	CAGCAGCAGATCCAAGTTCA GTTTGCAATTGTAGGGCGAGT
<i>elf1</i>	DMELF1	2-86	5' UTR	ACAGCAACAACGGAGCAAC TCTGCAACCTGGGAGTCTG
<i>z600</i>	DMZ60MEX1	3-41	intron	AAATCTGTTGCTCATACTGCCC AACCGGCGAAATGTTTCAG
<i>antennepedia</i>	DR015DC97Z	3-47.4	?	TGCATTGCTAATGATCGTGG GCTTTTCTTACACAATTCGCA
<i>nanos</i>	DRONANOS	3-66.2	3' UTR	CGCAAGTATTCAATTCAACACA TGCTGGCGGTGTTTCAT
<i>ula</i>	DMU1951	3-81	-	GGGTCTTTCTGCTTCAGTTACC GGAATACACGAATCCCCCTT
<i>Mic1</i>	DROMYALK	3-98	3'	GGGTAATCCCTTGCTAATATGG AATGGTTGTCGCTAAAAGTT

Primer sequences are written 5' to 3'. In noting the location of the microsatellite in the gene, 3' means 3' of the gene.

MATERIALS AND METHODS

The *findpatterns* routine of the Genetics Computer Group (GCG) was used to search GenBank for *D.melanogaster* loci containing di- or tri-nucleotide repeats. From the resulting set, large di- and tri-nucleotide microsatellites were chosen in order to provide as uniform as possible coverage of the *D.melanogaster* genetic map. We successfully amplified 18 of the 21 loci attempted, of which 17 were highly polymorphic. Table 1 shows the name and map position of each of the polymorphic loci. The location within the gene of the microsatellite (coding, intronic, etc) is also reported,

along with the primer sequences used in amplification. The primers were designed using Primer 0.5 (18) and made on a Beckman oligonucleotide synthesizer. Polymerase chain reaction amplifications were carried out in a 50 µl mixture containing 5 µl 10× reaction buffer, 2.5 U *Taq* polymerase, 14 mM MgCl₂, 200 µM dNTPs, 250 ng of each primer and 1 µg of template DNA. An annealing temperature of 55°C was used in all amplifications. Polymerase chain reaction products were electrophoresed in 4% Metaphor agarose gels (FMC Bioproducts) using FMC's rapid horizontal protocol, which allows resolution of 1-2% differences of fragment length. Since most of the primer pairs produced fragments

Figure 1. The distribution of fragment sizes observed at a dinucleotide repeat in *zeste-white 3* (GenBank locus name DMZW3K25).

of ~100 base pairs, Metaphor gels are sufficient for unambiguous typing of trinucleotides and moderately accurate typing of dinucleotides. pBR322 digested with *Hae*III was used as a length standard. The gels were stained with ethidium bromide and photographed with Alpha Innotech's gel documentation system. Fragment sizes were determined using the ProScan RFLP package.

Between 13 and 15 isofemale lines were scored to assess natural levels of variation. The isofemale lines were collected throughout North America between 5 and 15 years ago and have been maintained since collection in 95 mm vials. It is expected that some mutation has occurred in the laboratory, but the magnitude of the variation contributed by new mutations was assessed by typing 27 isogenic lines that have been maintained in the laboratory for about as long as the isofemale lines. At one of the typically polymorphic loci (*zeste-white* 3, GenBank locus name DMSGG3), no pair of isogenic lines differed by more than a single repeat unit and the majority were identical. This demonstrates that the total amount of variation introduced by new (laboratory) mutations is of far lower magnitude than the original variation in the population, in accord with theoretical expectation. Assuming a mutation rate on the order of 1×10^{-4} , the average difference between lines separated for ~200 generations would be less than one repeat unit (19).

From each of the isofemale lines, 20 flies were taken and pooled for DNA extraction using phenol/chloroform. Because the PCR was performed on the combined extract of numerous flies we must consider the possibility of multiple bands due to polymorphism in the isofemale line. Several lines showed evidence of polymorphism, as would be expected on theoretical grounds. Depending on the effective population size, the probability that polymorphism is maintained can be calculated using Kimura's solution to the Kolmogorov forward equation. For example, for an effective population size of 50 and a time of 175 generations, the probability that a line maintains polymorphism is ~10% (20). Because we cannot know the expected number of polymorphic lines without knowing the effective population size, it is not possible to determine exactly the number of alleles that are being scored unless we record only one allele from every isofemale line. This approach

Figure 2. The variance of repeat count versus maximum repeat count observed. (Top) Perfect dinucleotide microsatellites are represented by filled diamonds and imperfect (split) microsatellites are represented by unfilled diamonds. The regression line was fitted using data from both perfect and imperfect dinucleotide repeats. (Bottom) Perfect trinucleotide microsatellites are represented by filled triangles and imperfect trinucleotides by unfilled triangles. The regression line was fitted using both perfect and imperfect trinucleotides.

is also necessary because it is often impossible to distinguish multiple alleles from stuttering, in which bands of (often) weaker intensity appear due to replication errors during amplification. For these reasons, we always scored only the band of greatest intensity and in those rare cases (~1–2%) where there were two bands of the same intensity we chose one of the bands at random. In agarose gels stuttering sometimes resulted in a single, thick band. We determined the fragment size at the center of the band and recognized that the score might be off by as much as one repeat unit. Averaging of adjacent alleles (see below) was used to reduce the impact of inaccurate scores.

RESULTS AND DISCUSSION

Figure 1 shows observed allele frequencies at a dinucleotide repeat in *zeste-white 3* (GenBank locus name DMZW3K25), a representative locus. Notice that many alleles differ by a single repeat unit,

which is inconsistent with stepwise mutations. This is a common observation in PCR-based studies of microsatellites and is most likely due to a combination of improper PCR amplifications and imprecision in estimating fragment sizes from migration rates. For these reasons, we combined alleles separated by a single nucleotide and designated the combination a new allele whose size is the weighted average of the sizes of the constituent alleles. With runs of three consecutive alleles two were chosen at random for combination and the other left alone. The rationale for combining only two alleles in a run of three is that typings are probably accurate to within about a base pair or so and alleles differing by two base pairs should not be combined. None of the results will be very sensitive to the choice of which pair (out of three consecutive alleles) are combined. In general, the combining of adjacent alleles results in more conservative and probably more accurate estimates of diversity. As an example, *zeste-white 3* was scored as follows: The first two alleles (one of size 99, three of size 100) were combined into four observations of an allele of size 99.75. The next allele, 101, was left alone. Similarly, 103 and 104 were combined.

Table 2 shows the core microsatellite sequence along with diversity (the probability that two alleles, chosen at random, are different) and other characteristics for each of the polymorphic loci studied. A split dinucleotide microsatellite, with nine perfect repeats preceded by an imperfection and another six perfect repeats, located in *Sex lethal* (GenBank locus name DROSEXLS11), was apparently monomorphic. The variance in repeat count is also reported. Under the stepwise mutation model its expectation is $(2N-1)\mu$, with N the diploid population size and μ the mutation rate (21). The variance is therefore expected to be a more sensitive indication of the mutation rate than the diversity. The reported diversity level has been corrected for sample size (22).

The diversity is not significantly correlated with the minimum, maximum or mean repeat count. With both the di- and tri-nucleotide

repeats, however, there is a correlation between the mean repeat count and the variance ($r = 0.7$, $P < 0.02$; and $r = 0.69$, $P < 0.08$, respectively). Interestingly, for both the di- and the tri-nucleotide repeats, the correlation between the variance and the maximum repeat count is stronger ($r = 0.84$, $P < 0.002$; and $r = 0.85$, $P < 0.014$, respectively) than that between the variance and the mean (see Fig. 2). Although based on very few observations, the suggestion of a stronger correlation between the maximum repeat count and variance than between the mean and variance suggests that the mutation rate does indeed increase with repeat count, but not smoothly. If confirmed, this may explain why some studies detect a positive correlation between repeat count and the mutation rate and others do not.

To assess the effect of imperfections on diversity, we compared the variance of dinucleotide repeats with imperfections to those without using an analysis of covariance. The number of repeats in the perfect microsatellites and the number in the larger half of the split microsatellites was considered a covariate, since we have shown that the variance goes up with repeat count. After adjusting for the covariate, the mean variance for the perfect microsatellites is significantly greater than that for the imperfect microsatellites (16.08 versus 8.4, respectively; $P < 0.05$). It is important to appreciate that only the larger portion of a split microsatellite contributes to the value of the covariate. Therefore, this analysis asks whether the additional repeats on the other side of the imperfection contribute anything to the variance. The answer is that they do not contribute, but moreover imperfections result in a lower variance than would be expected based on the size of the repeat string on the other side of the imperfection. The implication is that when imperfections split microsatellites, the mutation rate in the larger repeat string on one side of the imperfection is suppressed by the presence of the smaller string of repeats on the other side. This is an unexpected result and is not easy to understand.

Table 2. Summary of microsatellite sequences and statistics for population variability

Name	Sequence	Alleles	Min	Max	Mean	Var	Diversity
DMSGG3	(CAG) ₁₁	4(6)	6	13	10	5	0.80
DMZW3K25	(AT) ₁₄	8(14)	12.5	31	16	23.7	0.85
DROOTUA	(TA) ₅ CA(TA) ₉	4(15)	5.5	10.5	8	1.85	0.76
DROSEV	(TTG) ₉	3(13)	6	10	6.5	1.45	0.64
DMTENA	(TA) ₄ CC(AT) ₁₄	5(8)	9	16	13	4.54	0.86
DMTROPONI	(CA) ₁₁ CG(CA) ₄	5(10)	4.5	12	10	6.27	0.75
DROEXPAND	(CAG) ₈	5(12)	5	9	7	1.73	0.81
DROYANETSB	(TG) ₁₉	9(15)	8	28.5	17	43	0.91
DROMHC	(CA) ₁₃ GA(CA) ₂	4(6)	9	13.5	11	3.3	0.8
DMMP20	(CA) ₁₀	4(12)	4	10.5	9	3.26	0.77
DMMASTER	(CAG) ₈ (CAA) ₂ (CAG) ₅	5(14)	5	8	6	.98	0.76
DMELF1	(CAG) ₇	4(5)	6	14	9	12.7	0.97
DMZ60MEX1	(TTC) ₈	6(15)	3	9	7	1.72	0.79
DR015DC97Z	(CT) ₁₂ GT(CT) ₃ TT(CT) ₂	4(13)	11	17	13	1.78	0.59
DRONANOS	(TA) ₁₈	6(10)	7.5	22.5	12.5	24.4	0.88
DMU1951	(TA) ₁₆	8(12)	13.5	21	16	10.3	0.94
DROMYALK	(CAA) ₅ CAC(CAA) ₂	6(11)	3	7	5	1.37	0.91

Two trinucleotide repeats were found in coding regions, within *expanded* (DROEXPAND) and *mastermind* (DMMMASTER). While their variances are less than the average for the trinucleotides, they are still highly variable. This result may have been expected based on the observation that trinucleotides are found in coding regions while dinucleotides are not. Stepwise mutations at dinucleotide-repeat loci result in frameshifts, which would destroy the function of the gene. They must be prohibited from coding regions precisely because stepwise mutations cannot be suppressed. This suggests that when trinucleotides are in coding regions they will undergo stepwise mutations.

The monomorphic microsatellite *Sxl* has a nearly identical structure to the one in *otu* (DROOTUA), which was polymorphic. This suggests the possibility that the one in *Sxl* is monomorphic due to a recent selective sweep, and that, more generally, it may be possible to identify regions that have undergone a sweep by identifying large but monomorphic microsatellites (cf. 23). This possibility can be checked by determining the variability of microsatellites linked to those with reduced variation.

The average diversities of the di- and tri-nucleotide repeats are nearly identical at 0.80 and 0.81, respectively. The average variances, however, are significantly different at 10 and 3.57, respectively (Kruskal-Wallis test, $H = 5.04$, $P < 0.025$). The mutation rate for dinucleotide microsatellites in humans has been estimated at 5.6×10^{-4} (11). If *D.melanogaster* also had a mutation rate on the order of 10^{-4} for microsatellites, this would imply that the effective population size would be $\sim 10\,000$, well below the more typically estimated value of 10^6 (24). The discrepancy could be caused by a number of factors, such as: (i) the mutation rate in humans may be substantially different from that in *D.melanogaster*; (ii) many of these loci may not be at mutation-drift equilibrium; or (iii) the markers may not be evolving neutrally.

It is important to appreciate that we made no attempt to optimize PCR conditions, yet, 17 out of 18 loci that we scored proved to be highly polymorphic. We also tested the primers for five of the loci on three closely related species (*D.simulans*, *D.mauritiana* and *D.sechellia*) and found that amplification was about as strong as for *D.melanogaster*. These results confirm the

utility of microsatellites for both population studies and genetic mapping in *Drosophila*.

ACKNOWLEDGEMENTS

This research was supported by NIH fellowship GM 16898 to DBG and NSF grant DEB-9419631 to AGC.

REFERENCES

- 1 Weber, J.L. and May, P.E. (1989) *Am. J. Hum. Genet.*, **44**, 388–396.
- 2 Beckmann, J.S. and Weber, J.L. (1992) *Genomics*, **12**, 627–631.
- 3 Valdes, A.-M., Slatkin, M. and Freimer, N.B. (1993) *Genetics*, **133**, 737–749.
- 4 Lagercrantz, U., Ellegren, H. and Anderson, L. (1993) *Nucleic Acids Res.*, **21**, 1111–1115.
- 5 Kashi, Y., Tikochinsky, Y., Genislaw, E., Iraqi, F., Nave, A., Beckman, J.S., Grunebaum, Y. and Soller, M. (1990) *Nucleic Acids Res.*, **18**, 1129–1132.
- 6 Ohta, T. and Kimura, M. (1973) *Genet. Res.*, **22**, 201–204.
- 7 Kimura, M. and Crow, J.F. (1964) *Genetics*, **49**, 725–738.
- 8 Di Rienzo, A., Peterson, A.C., Garza, J.C., Valdes, A.-M., Slatkin, M. and Freimer, N.B. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 3166–3170.
- 9 Estoup, A., Garney, L., Solignac, M. and Cornuet, J. (1995) *Genetics*, **140**, 679–695.
- 10 Henderson, S. and Petes, T. (1992) *Mol. Cell. Biol.*, **12**, 2749–2757.
- 11 Weber, J.L. and Wong, C. (1993) *Hum. Mol. Genet.*, **2**, 1123–1128.
- 12 Schloetter, C. and Tautz, D. (1992) *Nucleic Acids Res.*, **20**, 211–215.
- 13 Mahtani, M.M. and Willard, H.F. (1993) *Hum. Mol. Genet.*, **2**, 431–437.
- 14 Richards, R.I. and Sutherland, G.R. (1994) *Nature Genet.*, **6**, 114–116.
- 15 Wooster, R., Cleton-Jansen, A.-M., Collins, N., Mangion, J., Cornellis, R.S., Cooper, C.S., Gusterson, B.A., Ponder, B.A.J., von Deimling, A., Wiestler, O.D., Cornelisse, C.J., Devilee, P. and Stratton, M.R. (1994) *Nature Genet.*, **6**, 152–156.
- 16 Buchanan, F.C. and Crawford, A.M. (1993) *Animal. Genet.*, **24**, 145.
- 17 Ostrander, E.A., Sprague, G.F. and Rine, J. (1993) *Genomics*, **16**, 207–213.
- 18 Daly, M.J., Lincoln, S.E. and Lander, E.S. (1991) Unpublished software, Whitehead Institute.
- 19 Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. and Feldman, M.W. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 6723–6727.
- 20 Kimura, M. (1964) *J. App. Prob.*, **1**, 177–232.
- 21 Moran, P. (1975) *Theor. Pop. Biol.*, **8**, 318–330.
- 22 Nei, M. (1987) *Molecular Evolutionary Genetics*, Columbia University Press, OH.
- 23 Slatkin, M. (1992) *Mol. Biol. Evol.*, **12**, 473–480.
- 24 Singh, R.S. (1989) *Ann. Rev. Genet.*, **23**, 425–453.