Biased Mutations and Microsatellite Variation

Lev A. Zhivotovsky, Marcus W. Feldman,† and Sergei A. Grishechkin‡*

*Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia; †Department of Biological Sciences, Stanford University; and ‡Department of Mathematics, State Academy of Oil and Gas, Moscow, Russia

Mutation bias is one of the forces that may constrain the variation at microsatellite loci. Here, we study the dynamics of population statistics and the genetic distance between two populations under multiple stepwise mutations with linear bias and random drift. Expressions are derived for these statistics as functions of time, as well as at mutation–drift equilibrium. Applying these expressions to published data on humans and chimpanzees, the regression coefficient of mutation bias on allele size was estimated to be at least between -0.0064 and -0.013. The assumption of mutational bias produces larger estimates of divergence times than are obtained in its absence; in particular, the time of split between African and non-African human populations is estimated to be between 183,000 and 222,000 years, assuming one-step mutations and no selection. With multistep mutations, the divergence time is estimated to be lower.

Introduction

Short, tandemly repeated DNA sequences, microsatellites, occur widely throughout the genomes of higher organisms (Kashi et al. 1990). Although their abundance in plants is one fifth that in animals, the within-locus variation in plants is still high (Lagercrantz, Ellegren, and Anderson 1993). For evolutionary studies, microsatellite loci have several advantages over other genetic markers (allozymes, RFLP, RAPDs): they occur in all chromosomal regions; are extremely polymorphic, with up to several dozen alleles at each locus, and are easily scored using PCR. Thus, microsatellites are potentially very useful for the analysis of population structure, as well as the evolutionary history of closely related taxa, and have been used for the taxonomy of humans and higher primates (Bowcock et al. 1994; Deka et al. 1995; Goldstein et al. 1995*b*; Slatkin 1995; and others).

Analysis of microsatellite frequencies may be carried out with the statistical tools usually used for other markers (Nei 1987; Weir 1990). However, since each microsatellite allele may be characterized quantitatively by its size, i.e., the number of repeats of the DNA motif, we may also apply the techniques of quantitative genetics. Measures of population subdivision related to F statistics (Slatkin 1995; Michalakis and Excoffier 1996; Rousset 1996), distances between populations using the squared difference between the mean population values and variances (Goldstein et al. 1995*a,* 1995*b*), and higher order statistics (Zhivotovsky and Feldman 1995; Goldstein et al. 1996) have all been used for the study of microsatellite polymorphisms.

In most studies, the evolutionary dynamics under random drift and mutation have been investigated assuming no constraints on allele size. Although there is still very little known about the precise relationship between allele size and the rate of mutation, it seems reasonable that there are constraints that restrict microsatellite variation to bounded intervals (Bowcock et al. 1994; Goldstein et al. 1995*a*; Nauta and Weissing 1996; Feldman et al. 1997). One such constraint may be mutational bias such that alleles of large size mutate preferentially to alleles of smaller size, and vice versa for alleles of small size. Another possible constraint is selection acting against multiple repeats beyond some size threshold. In this paper we consider constraints in the context of mutation bias.

Garza, Slatkin, and Freimer (1995) proposed a model with a bias that increases in proportion with the deviation from some intermediate size. They used coalescence theory to obtain expressions for the ultimate within-population variance and the ultimate distance between two populations at mutation–drift equilibrium, and, using human population data, it was concluded that mutation bias is rather small, possibly even less than the mutation rate. However, in their application of the Ornstein-Uhlenbeck process, a factor μ, the mutation rate, appears to have been omitted from the formula for the mean of the derived normal distribution. It therefore remains to resolve whether mutation bias has a significant effect on evolutionary statistics. The dynamics of within- and between-population statistics in the presence of mutation bias also remain to be explicated. Feldman et al. (1997) analyzed a one-step mutation model with no bias, but with hard boundaries at two extremal alleles, i.e., those with the lowest (a single copy) and highest (R copies) repeat numbers. With this model, the within-population variance and the expected genetic distance between two populations as a function of time since they originally bifurcated were computed. In these two studies, only statistics of second order were calculated. However, statistics of higher order, such as the variance of within-population variances and the variance of the genetic distances, have been shown to be useful for estimation purposes (Zhivotovsky and Feldman 1995). In addition, how population statistics change over time can be important if the history of populations under study is not long enough for the attainment of equilibrium (Zhivotovsky et al. 1994; Goldstein et al. 1996).

In this paper, we consider linear mutation bias and derive expressions for population statistics from which estimators for the strength of bias (the regression coefficient of the mutation bias on allele size) and for the

Key words: microsatellites, stepwise mutation model, genetic drift, population statistics, genetic distance, divergence time, estimated bias.

Address for correspondence and reprints: Marcus W. Feldman, Department of Biological Sciences, Stanford University, Stanford, California 94305. E-mail: marc@charles.stanford.edu.

Mol. Biol. Evol. 14(9):926–933. 1997

divergence time since two populations split are suggested. Then, using published data, we show that our estimate of the strength of mutation bias is much higher than the mutation rate. We also show that the estimated divergence time between populations may depend greatly on the strength of mutational bias. Expressions are derived for the expected population mean, variance, skewness, and kurtosis in repeat number, as well for the expected distance between two populations produced as a bifurcation from an ancestral population and the variance of this distance.

The Model

Mutation Scheme

Consider a randomly mating population with nonoverlapping generations, a constant effective number of gametes N, and an autosomal microsatellite locus with multiple alleles. We shall use i to refer to alleles with i repeats. Each allele may mutate to alleles of different sizes, with μ, the total mutation rate, assumed to be the same for all alleles. For an arbitrary allele i, define the bias, b_i, as the difference between the mean size $\bar{\imath}$ of the alleles to which the allele i mutates and the size of this allele:

$$b_i \stackrel{\text{df}}{=} \bar{\imath} - i. \tag{1}$$

This bias is positive if the average size of alleles among the mutants exceeds the parental allele size; otherwise it is negative. Denote by σ_m^2 the variance in repeat number among the mutants of an allele assumed to be independent of parental alleles (Slatkin 1995).

In the absence of any mutation bias, the allele distribution in the population may wander indefinitely under genetic drift (Moran 1975). If mutation bias is the only force that constraints the number of repeats, then alleles of large size should mutate more often to alleles with a smaller number of repeats, and vice versa, in which case alleles of some intermediate size might be nearly free of mutation bias. Here, we assume that the bias increases in absolute value with deviation from an intermediate level, r_m, as a linear function of allele size; $b_i = B(i - r_m)$ with $B < 0$, a mutation scheme introduced by Garza, Slatkin, and Freimer (1995). We call r_m the focal value of the bias. In other words, B is a (negative) regression coefficient of mutation bias on the size of parental alleles, and its absolute value can be considered as the strength of mutation bias. The results used here are derived in the appendix using differential equations for the dynamics of the population moments of repeat numbers under mutation and genetic drift under the assumption that the population size is sufficiently large (unpublished data).

Mean and Variance

Let r and V, respectively, be the mean and variance of repeat numbers calculated over all alleles weighted by their frequencies in the population. At mutation–drift equilibrium, the expected values of the statistics will be marked with a "hat"; for example, $\hat{r}$ denotes the equilibrium mean repeat number and $\hat{V}$ the equilibrium variance.

We show in the appendix that the expected value of the mean repeat number $r(\tau)$ in the population is an exponential function of time (in τ generations):

$$\mathcal{E}[r(\tau)] = r_m + (r_0 - r_m)e^{B\mu\tau}. \tag{2}$$

The second term approaches zero with increasing time, and the ultimate expected number of repeats approaches the focal value r_m, near which the mutation bias is close to zero. The rate of approach to the equilibrium does not depend on population size, so the expectation of the mean value of repeat numbers is close to its theoretical value if the population has been evolving for a sufficiently long time.

The dependence of the population variance on time is more complicated and is also derived in the appendix. Its value at mutation–drift equilibrium is

$$\hat{V} = \frac{2\sigma_m^2 N\mu}{(2 + B)(1 - 2BN\mu)}. \tag{3}$$

We use approximations that hold for large population size (unpublished data) so that in the case of no bias, $\hat{V}$ becomes $\sigma_m^2\mu N$ instead of the exact value, $\sigma_m^2\mu(N - 1)$ (see Zhivotovsky and Feldman 1995). This is unimportant unless the population size is very small.

Variance of Means

The expected number of repeats in a population is the average of the means of repeat numbers taken among all possible evolutionary trajectories (realizations). Any single realization may deviate far from the expected value due to random drift. We can calculate the variance of means among realizations, $\mathrm{Var}(r)$, as a measure of this deviation. It is shown in the appendix that, at mutation–drift equilibrium,

$$\widehat{\mathrm{Var}}(r) = \frac{\hat{V}}{2(-B)N\mu} \tag{4}$$

(recall that $-B$ is positive). Likewise, the within-population variance of repeat numbers in a single realization deviates from its expected value, $\hat{V}$. The variance of variances among realizations is provided in the appendix.

Genetic Distance

Consider two independently evolving populations that have the same set of parameters (N, μ, B). Goldstein et al. (1995*b*) defined a distance $(\delta\mu)^2$ between two populations, A and B, as the squared difference of their mean values, r_A and r_B. At generation τ since the populations diverged,

$$(\delta\mu)^2(\tau) \stackrel{\text{df}}{=} [r_A(\tau) - r_B(\tau)]^2. \tag{5}$$

The ultimate value $\widehat{(\delta\mu)^2}$ to which the expectation of $(\delta\mu)^2(\tau)$ converges under mutation bias is

$$\widehat{(\delta\mu)^2} = \frac{\hat{V}}{(-B)N\mu} \tag{6}$$

(see appendix).

With the standard assumption that both populations, as well as the ancestral population, are in mutation–drift equilibrium,

Table 1
Regression Coefficients of Mutation Bias on Allele Size

Mutation Variance	Relative Variation in r_m ($100Q\%$)			
σ_m^2	0	20	50	90
1.0	−0.0064	−0.0077	−0.0110	−0.0256
1.5	−0.0096	−0.0116	−0.0165	−0.0385
2.0	−0.0128	−0.0154	−0.0220	−0.0513
4.0	−0.0257	−0.0308	−0.0440	−0.1026

Note.—The regression coefficient is calculated as a function of the mutation variance, σ_m^2, and the percent of the observed between-locus variance of the mean repeat numbers due to variation in focal values, $100Q$, based on human and chimpanzee data (see text).

the expected distance between the daughter populations since their split is an exponential function of time,

$$(\widehat{\delta\mu})^2(1 - e^{2B\mu\tau}). \quad (7)$$

The variance of the distance is given in the appendix.

Results

Estimator of Mutation Bias

Using equations (3) and (4) and neglecting B^2, we suggest the following estimator $\mathbf{b}$ of B based on population data with multiple microsatellite loci:

$$\mathbf{b} = -\frac{\sigma_m^2}{2[\mathrm{Var}(\bar{r}) + \bar{V}]}. \quad (8)$$

Here, $\bar{r}$ is an estimate of the mean repeat number at a particular locus, $\mathrm{Var}(\bar{r})$ is the variance of the mean estimates among the loci, and $\bar{V}$ is the estimate of within-locus variances of repeat numbers averaged over loci.

Using equations (3) and (4), various estimators of mutation bias are possible. The estimator (8) is derived under the following assumptions:

(i) The within-population statistics of repeat numbers have already achieved mutation–drift equilibrium.
(ii) Mutation processes at all loci used in the estimate are the same. In particular, the expected values of the means, r_m, and the regression coefficient B are the same.
(iii) The means at all the loci evolve independently of each other.
(iv) B is small.

The validity of these assumptions is discussed later. Note, however, that if (ii) is violated, the between-locus variance, $\mathrm{Var}(\bar{r})$, consists of two components: evolutionary variation in the mean repeat number among loci, $\widehat{\mathrm{Var}}(r)$, and variation in the focal values r_m, $\mathrm{Var}(r_m)$. Therefore, $\mathbf{b}$ actually estimates

$$-\frac{\sigma_m^2}{2[\widehat{\mathrm{Var}}(r) + \mathrm{Var}(r_m) + \hat{V}]}, \quad (9)$$

which is less than B in absolute value. Thus, $\mathbf{b}$ gives a lower estimate of the numerical value of B when there is variation among loci due to differences in their biases.

The absolute value of $\mathbf{b}$ is an increasing function of σ_m^2 and a decreasing function of $\mathrm{Var}(r_m)$. Therefore, we suggest

$$\mathbf{b}_0 = -\frac{1}{2[\mathrm{Var}(\bar{r}) + \bar{V}]} \quad (10)$$

as a lower bound (in absolute value) for the estimate of B when variation occurs among loci in the extent of mutational bias, $\mathrm{Var}(r_m)$, and when there is unknown variance in the size of new mutations, σ_m^2.

Using the estimator $\mathbf{b}_0$, we may calculate the lower bound for the strength of mutational bias based on data for the eight microsatellite loci in humans and chimpanzees reported by Garza, Slatkin, and Freimer (1995, table 1). The within-locus variance averaged over the loci and the species was 13.02 and the between-locus variance (unbiased estimate) in the mean values averaged over the species (with equal weights) was 64.82. Thus, the lower estimate, $\mathbf{b}_0$ turns out to be −0.0064. Therefore, the strength of mutation bias is at least 10 times the average mutation rate at autosomal loci, estimated by Weber and Wong (1993) to be 5.6×10^{-4}.

Actually, the lower estimate given by equation (10) would be improved if we knew σ_m^2 and $\mathrm{Var}(r_m)$. Let Q be the fraction of the observed between-locus variance due to variation in focal values, r_m, i.e. $Q = \mathrm{Var}(r_m)/[\widehat{\mathrm{Var}}(r) + \mathrm{Var}(r_m)]$; recall that in applications, the denominator is estimated as $\mathrm{Var}(\bar{r})$. Then the following statistic, $\mathbf{b}_1$, would give an improved estimate of B:

$$\mathbf{b}_1 = -\frac{\sigma_m^2}{2[(1 - Q)\cdot\mathrm{Var}(\bar{r}) + \bar{V}]}. \quad (11)$$

Using loci with dinucleotide repeats in humans, Di Rienzo et al. (1994) have estimated σ_m^2 to be between 4 and 20. Such large values of the mutation variance are possible only if large changes in the number of repeats occur among new mutations. For example, if 90% of new mutations are one step and 10% are two steps, then σ_m^2 is about 1.3. However, with 90% for one-step mutations and 2% each for two- to six-step mutations, σ_m^2 would become 2.7. Di Rienzo et al. (1994) have shown that multistep mutations significantly influence the divergence rate and within-population variation. Zhivotovsky and Feldman (1995) have shown that higher order statistics are even more affected by σ_m^2. The same is true for the estimate of mutation bias.

Table 1 gives the estimates of B for the above-mentioned data from Garza, Slatkin, and Freimer (1995) using different values of the mutation variance and the relative variation in the focal value. Clearly, the esti-

mated bias becomes stronger as the variation σ_m^2 among new mutations and/or the variation in focal values r_m increases (see eq. 9).

Estimator of Divergence Time

From equation (7), we suggest the following estimator, **T**, of the divergence time since the populations bifurcated:

$$\mathbf{T} = \frac{1}{2B\mu} \ln\left[1 - \frac{(\delta\mu)^2}{(\widehat{\delta\mu})^2}\right], \qquad (12)$$

where $(\delta\mu)^2$ is the esimated distance between the two populations. We may use this esimator unless the expected distance $(\widehat{\delta\mu})^2$ exceeds the observed distance $(\delta\mu)^2$. Using equations (6) and (3) and replacing $\hat{V}$ with its estimate, $\bar{V}$, we have

$$\mathbf{T} = \frac{1}{2B\mu} \ln\left[1 + \frac{(\delta\mu)^2}{2} \cdot \frac{B(2 + B)}{\sigma_m^2 + B(2 + B)\bar{V}}\right].$$

Since **T** is a decreasing function of σ_m^2, the estimator

$$\mathbf{T}_0 = \frac{1}{2B\mu} \ln\left[1 + \frac{(\delta\mu)^2}{2} \cdot \frac{B(2 + B)}{1 + B(2 + B)\bar{V}}\right], \qquad (14)$$

obtained by substituting $\sigma_m^2 = 1$ from the one-step mutation model, gives an upper bound of the divergence time. Note that both the true (eq. 13) and upper (eq. 14) estimates increase with increasing bias strength as measured by the numerical value of B.

Using the upper bound of divergence time given by equation (14), the strength of mutational bias is actually larger than the lower estimate, 0.0064. Indeed, from the same data of Garza, Slatkin, and Freimer (1995, table 1), we calculate the distance, $(\delta\mu)^2$, between humans and chimpanzees as the squared difference of their means averaged over the loci, which turns out to be 50.15. Using the above estimate of the within-locus variance, $\bar{V}$ (13.02), a mutation rate at autosomal microsatellite loci of 5.6×10^{-4} (Weber and Wong 1993), and a generation time of 20 years (used in Garza, Slatkin, and Freimer 1995), and taking the lower bound for B (-0.0064), we obtain the upper estimate $\mathbf{T}_0$ of the divergence time between chimpanzees and humans as 1.35×10^6 years, which is about three or four times lower than most estimates of their separation time. However, as seen from table 1, the value 0.0064 underestimates bias, and larger numerical values of B could be closer to the truth. For example, if B is -0.01, then $\mathbf{T}_0 \approx 2 \times 10^6$, while if B is -0.013, then $\mathbf{T}_0 \approx 5.1 \times 10^6$, which is much closer to the commonly accepted time of divergence between humans and chimpanzees (see, e.g., Horai et al. 1995).

Therefore, our calculations again suggest that mutational bias may be an important factor in evolution at microsatellite loci. Our results show that the linear bias could be in the range of B values from -0.0064 to -0.013 for humans. Of course, this argument assumes that selection is weaker than mutation bias, which seems reasonable in this case, since these data represent only dinucleotide repeats, on which selection is generally assumed to be most weak.

Table 2
Estimates of Divergence Time Between African and Non-African Human Populations

Regression of Mutation Bias on Allele Size, B	Mutation Variance, σ_m^2			
	1.0	1.5	2.0	4.0
-0.0064	183	115	84	40
-0.0100	202	122	88	41
-0.0200	288	149	100	44
-0.0500	*	444	178	53

NOTE.—**T** estimates of divergence time are given in thousands of years as a function of B and the variance in repeat number among new mutations, σ_m^2 (see text); * indicates that the expected equilibrium distance is smaller than the observed distance.

We apply our findings to estimate the divergence time between African and non-African human populations. Goldstein et al. (1995*b*) have estimated the distance between African and non-African modern human populations to be 6.47. Considering, as they did, a generation time for humans of 27 years, a one-step mutation process, within-population variance at the microsatellite loci studied of 10.1 (Bowcock et al. 1994), and $B = -0.0064$, the time of split between African and non-African populations is estimated to be 183,000 years. That our estimate is larger than that of Goldstein et al. (1995*b*), 156,000 years, is due to our use of the model with biased mutations; the earlier model had no mutational constraints.

Both estimates of divergence time are based on the assumption that $B = -0.0064$ and there is one-step mutation. However, as follows from the above discussion, greater numerical values of B and σ_m^2 might be more realistic. For example, with $B = -0.013$, which turned out to fit the genetic distance between humans and chimpanzees better than $B = -0.0064$ (see above), and holding the other parameters the same, the divergence times between African and non-African human populations would be even greater, about 222,000 years. The effect of mutation variation, σ_m^2, on **T** is opposite that of the strength of mutation bias: the larger the numerical value of B, the lower the estimated divergence time. Table 2 records estimates from equation (13) of the time of split between these populations for different values of σ_m^2 and B. The estimates are clearly sensitive to both strength of bias and mutation variance.

Discussion

We have attempted to develop an analytical approach to the analysis of population dynamics of microsatellite variability under random drift and mutation with constraints on the number of repeats caused by mutation bias. The bias we study is defined as the difference between the mean repeat number among mutants of a particular "parental" allele and the size of this allele. In general, bias is a function of the allele size, and, in order for the allele size to be constrained, the bias should be

positive for alleles with a small number of repeats and negative for alleles with a large repeat number. In other words, the slope of the bias function, i.e., the regression coefficient, B, of the bias on the allele size, must be negative.

The most important indication of this study is that mutation bias could have an important effect on both within-population variability and phylogenetic properties inferred from microsatellite data. We have suggested the lower estimator, $\mathbf{b}_0$, of B (see eq. 10). Using data for humans and chimpanzees, B is estimated to be between -0.0064 and -0.013. This estimator is suggested under assumptions (i)–(iv). Assumption (ii) was discussed above, while among the other assumptions, the most questionable is that the within-population variation has achieved mutation–drift equilibrium. If a population has maintained a constant size during its history, we may assume that there was enough time to attain the equilibrium. However, if the population size has changed over time, the within-population variance, V, may be far from its equilibrium value. Fortunately, the ultimate value of Var(r) is almost independent of population size (see eq. 23). Moreover, as follows from equation (4), in which the product $2BN\mu$ is expected to be smaller than one, the value of Var(r) may be much greater than V. Indeed, in the above human and chimpanzee data, we found the between- and within-locus variances to be 64.82 and 13.02, respectively. Thus, we expect that the estimators should depend little on population size. Zhivotovsky and Feldman (1995) have shown that assumption (iii) holds for the case of no bias and thus should be valid for sufficiently small B. Assumption (iv), that B is small enough to neglect B^2, seems validated. Therefore, $\mathbf{b}$ seems to be a reliable estimator of the strength of mutational bias and $\mathbf{b}_0$ a lower bound of it, unless selection is important.

Mutation bias can significantly increase the estimated divergence time compared to estimates obtained assuming no bias. Using the same data as Goldstein et al. (1995*b*), whose estimate of the split time between African and non-African human populations was 156,000 years (assuming one-step mutations), and taking the bias regression coefficient, B, to be between -0.0064 and -0.013, our estimate lies between 183,000 and 222,000 years. As follows from equation (13), multiple-step mutations give smaller estimates of the divergence time which might be more appropriate.

Mutation bias may be difficult to assess in mutation experiments, since it is not expected to be numerically large. Obviously, in order to detect mutation bias, parental alleles of large (or small) size relative to the expected mean repeat number (i.e., to the focal value r_m) are more useful than those close to r_m. However, as table 3 shows, even for such extreme allele sizes, the deviation of the statistics from their expectations may be difficult to detect, unless the sample size of mutations is sufficiently large or mutation bias is much stronger than the lower estimate used in table 3.

In conclusion, it should be noted that selection could also restrict variation at microsatellite loci, at least in the case of trinucleotides. For example, large numbers of CAG repeats in gene IT15 are related to Huntington's chorea (Huntington's Disease Collaborative Research Group 1993) and would be therefore be selected against. It would be interesting to compare results obtained under the model of selection with those that assume constrained mutation.

Table 3
Mean Sizes and Proportions among Mutations Arising from Alleles of Large Size

Statistic	Size of Parental Allele		
	5	15	25
Mean mutation size	4.97	14.90	24.84
Proportion of -1 mutations . . .	0.516	0.548	0.580
Proportion of $+1$ mutations . . .	0.484	0.452	0.420

Note.—Allele sizes are given in deviations from the focal value (the expected mean repeat number), r_m (see text); the mean values and the proportions among mutations arising from a parental allele are derived assuming the one-step mutation model with $B = -0.0064$. Assuming no mutational bias, the null hypothesis for the first statistic is that the expected mean size of new mutations is equal to the size of the parental allele; that for the second statistic is that the proportion between mutations with size reduced and increased by 1 is 0.5:0.5, respectively.

Acknowledgments

Research was supported in part by National Institutes of Health grants GM 28428, GM 28016, and TW0-0491 and Russian Foundation of Basic Research grant 95-04-11445.

APPENDIX
Additional Notation

Introduce the skewness, s_m, and kurtosis, k_m, in repeat numbers among mutations arising from the allele i as the corresponding central moments. Also note that instead of the actual number of repeats, i, it is possible to use $i' = i - r_m$. This transformation has no effect on the subsequent results, and in order to make the algebra simpler, we set r_m to zero, giving a bias function $b_i = -\beta i$, where, for convenience, we set $\beta = -B$.

Let ε denote the expectation taken over replicates with respect to the initial generation. As usual in random drift models, the discrete time measured in generations, τ, can be replaced by continuous time scaled by the population size; $t = \tau/N$. Hereafter, we assume that the mutation rate, μ, is the same for all alleles, and the symbol v is used to denote $N\mu$.

We shall study the mean, r, the variance, V, the skewness, S, and the kurtosis, K, in repeat number observed in the population. All of these have been analyzed for the model without bias or range constraints (Zhivotovsky and Feldman 1995). These population statistics can be expressed via the noncentral moments in repeat numbers of up to fourth order, which, in turn, satisfy a linear system of differential equations

$$\frac{d}{dt}\mathbf{M}(t) = -\mathbf{A}\mathbf{M}(t) + \mathbf{b},$$

derived under standard assumptions usually used for dif-

fusion approximations (Karlin and Taylor 1981): N is large enough that terms of order $1/N^2$ may be neglected, and the mutation rates are $O(1/N)$ (unpublished data); here, $\mathbf{M}(t)$ is the column vector of the noncentral moments at time t, $\mathbf{A}$ is a matrix whose elements are functions of β, v, σ_m^2, k_m, etc. (triangular in this case), and $\mathbf{b}$ is a column vector of constants. In particular,

$$\frac{d}{dt}M_1(t) = -\beta v M_1(t), \tag{15}$$

$$\frac{d}{dt}M_2(t) = (-2\beta v + \beta^2 v)M_2(t) + v\sigma_m^2, \tag{16}$$

and

$$\frac{d}{dt}M_{11}(t) = M_2(t) - (1 + 2v\beta)M_{11}(t). \tag{17}$$

The population statistics are expressed in terms of these moments:

$$\varepsilon\{r\} = M_1, \qquad \varepsilon\{V\} = M_2 - M_{11}, \tag{18}$$

$$\mathrm{Var}\{r\} = M_{11} - M_1^2, \tag{19}$$

$$\varepsilon_A\varepsilon_B\{\Delta\} = M_{A,11} + M_{B,11} - 2M_{A,1}M_{B,1}. \tag{20}$$

Results

The Mean

It follows directly from equation (15) that the expected value of the mean repeat number $r(t) = \Sigma_i\, ip_i(t)$ at time $t > 0$ is

$$\varepsilon\{r(t)\} = r_0 e^{-\beta v t} + \hat{r}(1 - e^{-\beta v t}), \tag{21}$$

where $\hat{r}$ is the value of $\varepsilon\{r\}$ at mutation–drift equilibrium. From equations (18) and (15), $\hat{r} = r_m = 0$.

Within-Population Variance

Solve equations (15) and (17), and neglect terms in β with power 2 and higher to obtain

$$\varepsilon\{V(t)\} \approx \hat{V} + (V_0 - \hat{V})e^{-(1+2\beta v)t} + \frac{\beta\hat{V}}{2}\left(e^{-(1+2\beta v)t} - e^{-(2-\beta)\beta v t}\right),$$

where V_0 is the variance of repeat numbers at time 0, and $\hat{V}$ is the expected variance in the population at mutation–drift equilibrium (eq. 3).

Variance of Means

Since the expectation of the within-population variance $V(t)$ is known (eq. 22), an expression for the between-realization variance of means can be obtained by integrating this linear scalar equation, given initial mean values and variance.

$\widehat{\mathrm{Var}}\{r\}$, the variance of the means among replicates at mutation–drift equilibrium is obtained from equation (19) to be equation (4). Using equation (3),

$$\widehat{\mathrm{Var}}\{r\} = \frac{\sigma_m^2}{\beta(2 - \beta)(1 + 2\beta v)}. \tag{23}$$

Higher Order Statistics

For higher order statistics, the corresponding differential equations cannot actually be solved directly. We have used two approaches to obtain approximate solutions of the complete system. First, note that its solution may be writen using an exponential matrix,

$$\mathbf{M}(t) = e^{-\mathbf{A}t}\mathbf{M}_0 + (\mathbf{I} - e^{-\mathbf{A}t})\hat{\mathbf{M}},$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{M}_0$ is the vector of moments at time $t = 0$, and $\hat{\mathbf{M}}$ is the vector of moments at mutation–drift equilibrium. The equilibrium values are expressed in terms of the inverse of matrix $\mathbf{A}$ as

$$\hat{\mathbf{M}} = \mathbf{A}^{-1}\mathbf{b}.$$

To find a solution as function of t, expand the matrix exponent in the previous expression:

$$e^{-\mathbf{A}t} = \mathbf{I} - \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 - \frac{1}{3!}\mathbf{A}^3t^3 + \ldots. \tag{24}$$

As a first approach, we substitute this expansion into the equation for $\mathbf{M}$ and then calculate the solution for the early generations of the process. The precision will depend on the number of terms retained in the expansion.

A second approach is to represent a particular solution using the matrix of fundamental solutions based on the eigenvalues and eigenvectors of the matrix $\mathbf{A}$ (e.g., Franklin 1968). This allows us to find a particular solution for arbitrary time. We have calculated the fundamental matrix, but some of the expressions are too cumbersome to be simplified, so these are omitted. *Mathematica* (Wolfram 1991) has been used to calculate the solutions for each of these approaches.

Variance of Variances

Now we consider the between-replicate variance of within-population variances, V, denoted as $\mathrm{Var}\{V\}$. Using the first approach above, approximate solutions for the early generations can be found. If r_0, V_0, and K_0 denote, respectively, the mean, variance, and kurtosis in a population at the initial time $t = 0$, then the variance of variances in the population at time t is

$$\begin{aligned}\mathrm{Var}\{V(t)\} &= t(K_0 - V_0^2) \\ &\quad + t^2[kv - (7 + 8\beta v)K_0 \\ &\qquad + 4\sigma^2 v V_0(1 - 3\beta) + (11 + 8\beta v)V_0^2 \\ &\qquad - 4\beta s_m v r_0 + O(\beta^2)]/2 + O(t^3).\end{aligned}$$

An approximation for this variance of variances at mutation–drift equilibrium is given by

$$\widehat{\mathrm{Var}}\{V\} \approx \frac{4}{3}\hat{V}^2 + \frac{k_m v}{6} - \beta\left(2\sigma_m^2 v^2 - \frac{\sigma_m^2 v}{2} + \frac{44\sigma_m^2 v^3}{9} + \frac{7k_m v^2}{9}\right) + O(\beta^2), \tag{25}$$

where $\hat{V}$ is given by equation (3).

Skewness and Kurtosis

The following are approximate solutions for the skewness and kurtosis in the early generations, with error of order t^2:

$$S(t) = S_0 + t[s_m v - 3S_0(1 + \beta v) - 3\beta v\sigma^2 r_0 + O(\beta^2)] + O(t^2),$$

$$K(t) = K_0 + t[k_m v - 4K_0(1 + \beta v) + 6V_0^2 + 6\sigma_m^2 v(1 - 2\beta)(V_0 - 4\beta s_m v r_0 + O(\beta^2)] + O(t^2),$$

Their ultimate values at mutation–drift equilibrium are:

$$\hat{S} \approx \frac{s_m v}{3}(1 - \beta v) + O(\beta^2),$$

$$\hat{K} \approx 5\hat{V}^2 + \frac{k_m v}{2} - \beta\left(6\sigma_m^2 v^2 - \frac{3\sigma_m^2 v}{2} + \frac{28\sigma_m^2 v^3}{3} + \frac{5k_m v^2}{3}\right) + O(\beta^2).$$

If there is no bias (i.e., $\beta = 0$), then these expressions, as well as equation (25), coincide with those derived previously (Zhivotovsky and Feldman 1995, eq. 8).

Genetic Distance

Consider two independently evolving populations that have the same set of parameters (N, μ, β, etc.) and that were derived from an initial population whose repeat numbers had mean r_0 and variance V_0. The symbol Δ is used in the appendix, instead of $(\delta\mu)^2$, for simplicity. Solving equations (15) and (17) directly, substituting the solution into equation (20), averaging over ancestral populations assuming mutation–drift equilibrium, and using the equilibrium expected values of r_0, V_0 and r_0^2, namely 0, $\hat{V}$, and $\hat{M}_{11} = \hat{\Delta}/2$, respectively, we obtain the expectation given by equation (7), with the expected distance converging to the limit given in equation (6).

In the early generations of the independent divergence of two populations, the variance of the distance between them increases as twice the square of the distance. This corresponds to the result that the distance follows a χ^2 distribution, claimed for the case of no bias by Zhivotovsky and Feldman (1995, result 3), who considered the case where the within-population variance was $\hat{V}$ at the time of bifurcation ($t = 0$) and made the tacit assumption that the variance did not change over time. This assumption is not strictly correct and is approximately valid only if t is not large (i.e., early in the divergence process). In fact, the variance changes with time under random drift and this causes the variance of the distance to deviate from the χ^2 prediction when the terms of order t^3 become large. Indeed, averaging over ancestral populations assumed to be at equilibrium, we obtain

$$\mathrm{Var}\{\Delta(\tau)\} = 2(\varepsilon\{\Delta(\tau)\})^2 + 2(3\hat{K} - 7\widehat{V^2})\left(\frac{\tau}{N}\right)^3 + O\left(\frac{\tau^4}{N^4}\right),$$

where $\hat{K}$ and $\widehat{V^2} = \widehat{\mathrm{Var}}(V) + \widehat{V^2}$ follow from the previous sections.

As time increases, the variance of the distance approaches its mutation–drift equilibrium:

$$\widehat{\mathrm{Var}}\{\Delta\} \approx 2\hat{\Delta}^2\left[1 + \frac{\beta z}{4} + \beta^2\left(\frac{z}{8} + \frac{v(1 - z)}{2} + \frac{2v^2}{3}\right)\right],$$

where $z = (k/\sigma_m^4) - 3$ is the normalized kurtosis of the mutation distribution. If $\beta = 0$, that is, there is no mutation bias, the ultimate variance of the distance again becomes twice the squared distance, as for a χ^2 distribution.

The variance of the distance for arbitrary time $t = \tau/N$ can be calculated from the variance of variances averaged over ancestral populations assumed to be at mutation–drift equilibrium:

$$\begin{aligned}\mathrm{Var}\{\Delta(t)\} &= f_1(1 - e^{-2\beta v t}) + f_2(1 - e^{-(1+4\beta v - v\beta^2)t}) \\ &\quad + f_3(1 - e^{-v\beta t(4-3\beta+\beta^2)}) \\ &\quad + f_4(1 - e^{-t(3+4\beta v)}) + f_5(1 - e^{-2v\beta t(2-\beta)}) \\ &\quad + f_6(1 - e^{-2t(1+2\beta v)}) + f_7(1 - e^{-4\beta v t}) \\ &\quad + f_8(1 - e^{-\beta v t(4-\beta)}) + f_9(1 - e^{-t(1+4\beta v)}),\end{aligned}$$

with the following approximations for the f coefficients:

$$\begin{aligned}f_1 &\approx 4\sigma_m^4/\beta^2 + 4\sigma_m^4(1 - 4v)/\beta + 3\sigma_m^4(1 - 4v + 16v^2) \\ &\quad + \beta 2\sigma_m^4(3 - 12v + 32v^2 - 192v^3)/3, \\ f_2 &\approx 8v(3\sigma_m^4 - k + 5\sigma_m^4 v)/3 \\ &\quad + \beta 4v^2(-111\sigma_m^4 + 32k - 256\sigma_m^4 v)/9, \\ f_3 &\approx 2(-3\sigma_m^4 + k)/\beta + (3\sigma_m^4 - k)(-3 + 8v) \\ &\quad + \beta(-3\sigma_m^4 + k)(5 - 24v + 64v^2)/2, \\ f_4 &\approx -(8\sigma_m^4 v^2)/3 + \beta 4v^2(-3\sigma_m^4 + 2k + 48\sigma_m^4 v)/9, \\ f_5 &\approx -\sigma_m^4/\beta^2 + (\sigma_m^4 - k + 8\sigma_m^4 v)/(2\beta) \\ &\quad + (6\sigma_m^4 - 3k - 8\sigma_m^4 v + 8kv - 64\sigma_m^4 v^2)/4 \\ &\quad + \beta(11\sigma_m^4 - 5k - 48\sigma_m^4 v + 24kv + 128\sigma_m^4 v^2 \\ &\quad\quad - 64kv^2 + 512\sigma_m^4 v^3)/8, \\ f_6 &\approx -v(k + 8\sigma_m^4 v)/3 \\ &\quad + \beta v(-9\sigma_m^4 + 12\sigma_m^4 v + 14kv + 184\sigma_m^4 v^2)/9, \\ f_7 &\approx 3\sigma_m^4/\beta^2 + (k - 12\sigma_m^4 v)/\beta \\ &\quad + (-27\sigma_m^4 + 18k + 96\sigma_m^4 v - 72kv + 464\sigma_m^4 v^2)/12 \\ &\quad + \beta(-81\sigma_m^4 + 45k + 396\sigma_m^4 v - 216kv \\ &\quad\quad - 2{,}160\sigma_m^4 v^2 + 976kv^2 - 4{,}416\sigma_m^4 v^3)/36,\end{aligned}$$

$$f_8 \approx -4\sigma_m^4/\beta^2 + 2(\sigma_m^4 - k + 8\sigma_m^4 v)/\beta + (18\sigma_m^4 - 9k - 48\sigma_m^4 v + 32kv - 160\sigma_m^4 v^2)/3 + \beta(99\sigma_m^4 - 45k - 432\sigma_m^4 v + 216kv + 1{,}824\sigma_m^4 v^2 - 832kv^2 + 3{,}200\sigma_m^4 v^3)/18,$$

$$f_9 \approx 4v(-6\sigma_m^4 + k)/3 + \beta 4v(-9\sigma^4 + 96\sigma_m^4 v - 18kv + 8\sigma_m^4 v^2)/9.$$

LITERATURE CITED

BOWCOCK, A. M., A. RUIZ-LINARES, J. TOMFOHRDE, E. MINCH, J. R. KIDD, and L. L. CAVALLI-SFORZA. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368**:455–457.

DEKA, R., L. JIN, M. D. SCHRIVER, L. M. YU, S. DECROO, J. HUNDRIESER, C. H. BUNKER, and R. E. FERRELL. 1995. Population-genetics of dinucleotide (dC - dA)n · (dG - dT)n polymorphisms in world populations. Am. J. Hum. Genet. **56**:461–474.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN, and N. B. FREIMER. 1994. Mutational processes of simple-sequence loci in human populations. Proc. Natl. Acad. Sci. USA **91**:3166–3170.

FELDMAN, M. W., A. BERGMAN, D. D. POLLOCK, and D. B. GOLDSTEIN. 1997. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics **145**:207–216.

FRANKLIN, J. N. 1968. Matrix theory. Prentice-Hall, London–Tokyo.

GARZA, J. C., M. SLATKIN, and N. B. FREIMER. 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. **12**:594–603.

GOLDSTEIN, D. B., A. R. LINARES, L. L. CAVALLI-SFORZA, and M. W. FELDMAN. 1995*a*. An evaluation of genetic distances for use with microsatellite loci. Genetics **139**:463–471.

———. 1995*b*. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92**:6723–6727.

GOLDSTEIN, D. B., L. A. ZHIVOTOVSKY, K. NAYAR, A. RUIZ LINARES, L. L. CAVALLI-SFORZA, and M. W. FELDMAN. 1996. Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol. Biol. Evol. **13**:1213–1218.

HORAI, S., K. HAYASAKA, R. KONDO, K. TSUGANE, and N. TAKAHATA. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. USA **92**:532–536.

HUNTINGTON'S DISEASE COLLABORATIVE RESEARCH GROUP. 1993. A novel gene containing trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell **72**:971–983.

KARLIN, S., and H. M. TAYLOR. 1981. A second course in stochastic processes. Academic Press, New York, N.Y.

KASHI, Y., Y. TIKOCHINSKY, E. GENISLAV, F. IRAQI, A. NAVE, J. S. BECKMAN, Y. GRUENBAUM, and M. SOLLER. 1990. Large restriction fragments containing poly-TG are highly polymorphic in a variety of vertebrates. Nucleic Acids Res. **18**:1129–1132.

LAGERCRANTZ, U., H. ELLEGREN, and L. ANDERSON. 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res. **21**: 1111–1115.

MICHALAKIS, Y., and L. EXCOFFIER. 1996. A genetic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics **142**: 1061–1064.

MORAN, P. A. P. 1975. Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. **8**:318–330.

NAUTA, M. J., and F. J. WEISSING. 1996. Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics **143**:1021–1032.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York, N.Y.

ROUSSET, F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation process. Genetics **142**:1357–1362.

SLATKIN, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

WEBER, A. O. M., and C. WONG. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2**:1123–1128.

WEIR, B. 1990. Genetic data analysis. Sinauer, Sunderland, Mass.

WOLFRAM, S. 1991. Mathematica. A system for doing mathematics by computer. 2nd edition. Addison-Wesley, Reading, Mass.

ZHIVOTOVSKY, L. A., and M. W. FELDMAN. 1995. Microsatellite variability and genetic distances. Proc. Natl. Acad. Sci. USA **92**:11549–11552.

ZHIVOTOVSKY, L. A., A. J. GHARRETT, A. J. MCGREGOR, M. K. GLUBOKOVSKY, and M. W. FELDMAN. 1994. Gene differentiation in Pacific salmon (*Oncorhynchus* sp.): facts and models with reference to pink salmon (*O. gorbuscha*). Can. J. Aquat. Sci. **51**(Suppl. 1):223–232.

DANIEL L. HARTL, reviewing editor

Accepted May 28, 1997