Mutation and migration in models of microsatellite evolution*

Marcus W. Feldman, Jochen Kumm, and Jonathan Pritchard

Department of Biological Sciences

Stanford University

Stanford, CA 94305

Version 2, revised January 26, 1999

*Contribution No. 2 from The Center for Computational Genetics and Biological Modeling, Stanford University. Research supported in part by NIH grants GM-28016 and GM-28428. J.P. is a Howard Hughes postdoctoral fellow.

Abstract

Dynamic and statistical properties of generalized stepwise mutation models are described and used to compare data on human di-, tri-, and tetranucleotide polymorphisms. The time-dependent behavior of an island model with stepwise mutation is analyzed and its equilibrium properties used to estimate the product Nm of the population size and the migration rate. Population statistics are derived from the complete equilibrium analysis of the model, and these are combined to give an estimate of Nm that is related to Slatkin's 1995 estimate which used R_{ST}. When this new statistic is applied to a set of 85 human microsatellite polymorphisms, the resulting population clusters match the tree of Bowcock et al. 1994 quite well.

Introduction

Because of their high level of polymorphism and ubiquity in eukaryotic genomes, microsatellites are widely preferred markers in evolutionary and ecological studies. Databases of microsatellites isolated for population-level analyses are under development. Statistical evaluation and evolutionary interpretation of microsatellite polymorphisms demand models for the origin and maintenance of this variability, and the models used so far are variants of the stepwise mutation model (SMM), which was originally introduced by Ohta and Kimura (1973) to model electrophoretically detectable enzyme variation in finite populations.

The first part of this paper describes these models and illustrates how stepwise mutation, in combination with genetic drift, affects important measures of within- and between-population variation. Using empirical estimates of overall mutation rates, these measures may then provide estimates of average population size or divergence times between populations. In the second part of the paper, the dynamics of a set of populations subject to migration according to Wright's (1943) island model are developed. From the equilibrium structure of this model, estimators of the extent of migration based on population statistics are suggested.

We shall assume throughout that the microsatellites are perfect; that is, each allele at a locus is completely specified by the number of times a motif is repeated. Generations are non-overlapping and each is produced from the previous by multinomial sampling from the parental generation's array of alleles. Our analysis follows that of Moran (1975), who developed recursions for the population central moments under the simplest stepwise mutation model permitting alleles of arbitrary (positive and negative) repeat number. The modifications to this model that we discuss below include asymmetric mutations of arbitrary size, a simple model of linear bias in which the rate of mutation depends on the number of repeats in the allele, and range constraints on the permitted repeat score.

Moran's Formulation: A Generalized SMM.

Consider a population of N diploid individuals and a locus at which each allele is characterized by its repeat score, which may take any positive or negative integer values. Let μ_c be the probability that mutation changes an allele by c repeat units irrespective of its original count. The total mutation rate is then $\mu = \sum_{c\neq 0} \mu_c$ and the expected change in repeat score for any allele is $\bar{c} = \sum c\mu_c$. The variance in repeat score change is $\sigma_m^2 = \sum c^2(\mu_c/\mu) - \bar{c}^2$. We write $w = \mu\sigma_m^2$. The standard one-step symmetric model has $\mu_{+1} = \mu_{-1} = \mu/2$, say, and $\mu_2 = 0$ otherwise, in which case $\bar{c} = 0$ and $w = \mu$. We shall refer to this special case as the one-step symmetric SMM.

Let p_i be the frequency of the allele carrying i repeats in the parental generation. Then, following mutation the frequency of this allele is

$$\tilde{p}_i = \sum_c \mu_{(i-c)} p_c \,. \tag{1}$$

The $2N$ gametes undergo multinomial sampling with probabilities $\{\tilde{p}_i\}$ to produce the $2N$ copies in the offspring generation. The process of mutation changes the population average repeat score $r = \sum_i ip_i$ to $\tilde{r} = \sum_i i\tilde{p}_i = r + \bar{c}$, while the variance changes from

$$V = \sum_i p_i(i-r)^2 \tag{2}$$

to

$$\tilde{V} = \sum_i \tilde{p}_i(i-\tilde{r})^2 = V + w. \tag{3}$$

Multinomial sampling has no effect on $\tilde{r}$ in the sense that $E_m(r') = r + \bar{c}$, where $E_m()$ is the expectation with respect to sampling, and the prime refers to the offspring generation. As shown by Moran, however, sampling reduces the variance by a proportion $\frac{1}{2N}$:

$$E_m(V') = \left(1 - \frac{1}{2N}\right)(V + w). \tag{4}$$

These results are minor modifications of Zhivotovsky and Feldman (1995), who also developed analogous recursions for the skew and kurtosis in the case $\bar{c} = 0$. From (4), taking expectations with respect to initial conditions, we obtain the equilibrium variance

$$\hat{V} = (2N - 1)w \tag{5}$$

Goldstein et al. (1995a) used a similar recursion in the one-step symmetric model for the quantity D_0, the average squared difference in repeat score between two alleles chosen randomly from a population of N dipoids. Note that $D_0 = 2V$.

In the symmetric case, the variance across loci of the equilibrium random variable V in (2) was given by Zhivotovsky and Feldman (1995), who showed that this variance will usually be quite large, of the order of $\hat{V}^2$. Pritchard and Feldman (1996) obtained the small sample analogue of this variance (see also Roe, 1992) and showed that as the sample size n increases, the two variances coincide. Goldstein et al. (1996) used properties of this variance to derive a confidence interval for the population variance. Consider $\bar{V}$, the mean over L loci of $\hat{V}$. We have recently obtained an expression for the variance $\mathrm{Var}_u(\bar{V})$ over evolutionary realizations of $\bar{V}$ for sampes of size n when the loci are unlinked:

$$\mathrm{Var}_u(\bar{V}) = \frac{N\mu\gamma(n^2+n) + 4N^2\mu^2(\sigma_m^2)^2(2n^2+3n+1)}{6L(n^2-n)}\,,$$

where $\gamma = \sum c^4(\mu_c/\mu)$. The corresponding result when the L loci are completely linked is

$$\mathrm{Var}_\ell(\bar{V}) = \mathrm{Var}_u(\bar{V}) + \frac{2N^2\mu^2(\sigma_m^2)^2(L-1)(n^2+n+3)}{9L(n^2-n)}\,.$$

An Application of $\hat{\mathbf{V}}$

The expectation (5) may be used to estimate the effective population size N when the mutation parameter w is known. Goldstein et al. (1996) used the variance obtained by averaging the variance in the pooled worldwide sample of human genotypes across the 30 dinucleotide loci studied by Bowcock et al. (1994). This variance was 10.1, and with Weber and Wong's (1993) estimate of 5.6×10^{-4} for μ, an estimate of about 9000 for

the effective human population size was obtained, under the assumption that $\sigma_m^2 = 1$ and $\bar{c} = 0$, the simplest symmetric SMM. Among the 30 loci in the study by Bowcock et al., one locus exhibited a variance some tenfold greater than the others. When this apparently aberrant gene is deleted, the variance drops to 6.827, giving a corresponding estimate for N of about 6100. In a situation such as this, it may pay to examine the molecular properties of the outlying locus in more detail.

Recent work in Cavalli-Sforza's laboratory has increased the number of dinucleotide loci studied to 85, with a corresponding average across loci of the pooled worldwide variance (using the same human sample as in the study by Bowcock et al.) of 8.652. Under the same assumptions as above, this corresponds to an effective population size of about 7700.

The assumption in these calculations that $\sigma_m^2 = 1$ is critical; it is clear that for a given observed $\hat{V}$, the estimate of N decreases by the factor $1/\sigma_m^2$. Dib et al. (1996) report on 5264 dinucleotide polymorphisms in CEPH families. The mean heterozygosity for these polymorphisms was 70% and the overall mutation rate was 6.2×10^{-4} while the average mutation resulted in an increase of about 0.39 repeat units. Although some 90% of the mutations produced repeat number changes of one or two, the mutation distribution had long tails on both positive and negative sides that resulted in an observed variance of about 4.5. A symmetrized geometric distribution with its parameter based on the average rate of one-step increases and decreases would produce a variance of about 2.5, considerably less than that observed. (The observed variance would, of course, be strongly influenced by a few atypical loci of large size that contribute many mutations.)

For the purposes of illustration, assume that the overall mutation rate is 6.0×10^{-4} and the variance in mutational changes is 2.5. Then, with a worldwide pooled average variance of 8.652, as observed for the 85 dinucleotide polymorphisms mentioned above, the estimate of N would be 2900. This would drop to about 1800 with $\sigma_m^2 = 4$. The reduction in estimated average effective human population size due to σ_m^2 is, therefore, substantial. It must be remembered, however, that in the model leading to these estimates of N, the probability of mutation is independent of the size of the allele in which the mutation occurs. We have yet to investigate the effect of this "stationarity" assumption on these estimates.

Tri- and Tetranucleotides

Data on a scale comparable to that discussed above for dinucleotide mutation and variation are not available for tri- and tetranucleotides. Weber and Wong (1993) directly observed mutations at 28 chromosome-19 loci and reported that the average mutation rate for tetranucleotides was nearly four times that for dinucleotides. The small number of loci and mutations suggests that it would be risky to take this factor very seriously. Heyer et al. (1997) directly estimated the mutation rate across seven Y-chromosome tetranucleotides as 0.002, in reasonable agreement with Weber and Wong (1993).

Chakraborty et al. (1997) used a relation for the population variance $\hat{V}$ similar to that of Zhivotovsky and Feldman (1995; see equn (5) above) to obtain indirect estimates of the relative mutation rates of di-, tri-, and tetranucleotides from polymorphisms reported in a number of different data sets. Dinucleotides had mutation rates 1.5–2.0 times higher than tetranucleotides, with non-disease-causing trinucleotides intermediate between the di- and tetranucleotides. In their analysis, they assumed that all three motifs had the same mutational variance, σ_m^2. (In fact, as mentioned above, Chakraborty et al. make the same approximation as Kimmel et al. (1996), namely $\bar{c}^2 = 0$; their assumption therefore amounts to all second moments about zero being the same.)

Using the same individuals sampled in the original study of dinucleotide polymorphisms by Bowcock et al. (1994), Bennett et al. (1998) have examined 22 trinucleotide and 21 tetranucleotide polymorphisms. For the trinucleotides, the pooled worldwide average variance was 3.994, and for the tetranucleotides it was 4.148. Recalling the estimate of 8.652 for the 85 dinucleotides, and since the same populations are involved, we may regard the ratios $8.652/3.994 = 2.166$ and $8.652/4.148 = 2.086$ as estimates of the ratios $w_{\mathrm{di}}/w_{\mathrm{tri}}$ and $w_{\mathrm{di}}/w_{\mathrm{tetra}}$, respectively. Of course, these ratios are products of the ratios of overall mutation rates times the ratios of mutational variances. If the latter are unity (i.e. the mutatational variances of di-, tri-, and tetranucleotides are all the same), then the above numbers are in good agreement with the estimates of Chakraborty et al. Until we have reliable direct estimates of the mutation rates and variances, interpretation of the observed variances will be difficult. If, for example, Weber and Wong (1993) and Heyer et al. (1997)

are correct, and the mutation rate for tetranucleotides is four times that of dinucleotides, then the values of the variance ratios above entail that the mutational variance of tri- and tetranucleotides should be about one eighth that of dinucleotides, but we estimated that σ_m^2 should be in the range 2-4 for dinucleotides, and σ_m^2 must be no less than 1. This contradiction suggests either that the factor four is incorrect or that the assumption of equilibrium under the SMM is incorrect. Also, we must remember that these calculations were predicated on the assumption that the size of mutational jumps does not depend on the size of the original allele, an assumption which has yet to be given a rigorous empirical test. Conceivably the constraints on the ranges of tri- and tetranucleotides could be much more stringent than on dinucleotides, and this would have an important effect on the relative variances (Feldman et al. 1997).

Genetic Distances for Stepwise Mutation Models

In order to estimate times of separation between populations (of the same or different species), it is desirable that the measure of separation, in our case the genetic distance, have an explicit relationship with time. Thus, for the infinite alleles model (IAM), Nei's (1972) standard genetic distance D_s between populations that diverged t generations ago is expected to be close to $2\mu t$, where μ is the mutation rate per locus per generation. When mutation occurs according to the SMM, no explicit form for the dependence of D_s on time has been found.

In considering properties of the variances in repeat numbers within and between populations under the SMM, we were led to an expression for the genetic distance between two populations which we called $(\delta\mu)^2$:

$$(\delta\mu)^2 = (m_x - m_y)^2, \tag{6}$$

where m_x and m_y are the mean repeat numbers at a locus in populations x and y (Goldstein et al. 1995a, b). For samples of size n_x and n_y from the two populations, the appropriate unbiased estimate of the population value $(\delta\mu)^2$ is (Goldstein and Pollock, 1997)

$$(\widehat{\delta\mu})^2 = \sum_i \sum_j (i-j)^2 x_i y_j - \tilde{V}_x - \tilde{V}_y = D_1 - \tilde{V}_x - \tilde{V}_y, \tag{7}$$

where $D_1 = \sum_i \sum_j (i-j)^2 x_i y_j$, with x_i and y_j the frequencies of repeat lengths i and j in the samples from x and y, respectively, and $\tilde{V}_x$ and $\tilde{V}_y$ are the usual unbiased estimates of repeat length variance in populations x and y. Goldstein et al. (1995a) showed that if populations x and y diverged from a common ancestral population t generations ago, then

$$E(\widehat{\delta\mu})^2 = 2\mu t \tag{8a}$$

under the simple symmetric SMM, while Zhivotovsky and Feldman (1995) showed that in the symmetric SMM with mutational variance σ_m^2, (8a) is replaced by

$$E(\widehat{\delta\mu})^2 = 2\mu\sigma_m^2 t = 2wt. \tag{8b}$$

It is not difficult to see that (8b) also applies in the asymmetric case. In applications, the sample values of this distance are calculated for each microsatellite locus, and then the average is taken over loci, under the assumption that the correlations between pairs of loci are weak.

It is worth comparing our result (8b) with the corresponding expression in equn (19) of Slatkin (1995). He estimates the time of divergence as

$$T_R = 4R_{ST}/(1 - R_{ST}),$$

where R_{ST} (with equilibrium assumed in each population) can be expressed in terms of our equations (5) and (7) above as

$$R_{ST} = \frac{D_1 - V_x - V_y}{D_1 + V_x + V_y},$$

when sample sizes are large. Hence we may write

$$T_R = \frac{2(D_1 - V_x - V_y)}{V_x + V_y}.$$

The corresponding ratio of expectations (assuming each population is at equilibrium) is

$$T_R \cong \frac{E(\delta\mu)^2}{E(V)} = \frac{2wt}{2Nw} = t/N$$

as in Slatkin's equations (17)–(19). Again, in the coalescent argument he uses, the factor $2N-1$ is replaced by $2N$. An argument similar to this relating R_{ST} to our $(\delta\mu)^2$ through the use of the variances was made by Kimmel et al. (1996).

Shriver et al. (1995) suggested a distance that replaces the squared differences in (7) by the absolute value of differences. This measure, D_{sw}, is difficult to relate analytically to the time of divergence between the groups under comparison, although simulation studies indicate that it is almost linear with time (Takezaki and Nei, 1996).

Although the distances $(\delta\mu)^2$ and D_{sw} were developed specifically for the SMM, they have high sampling variances (Goldstein et al. 1995a,b; Takezaki and Nei 1996). Thus, when $(\delta\mu)^2$ was applied to the data of Bowcock et al. (1994), the resulting population tree was poorly supported, with bootstrap values much lower than those obtained using an allele-sharing distance or the chord distance of Cavalli-Sforza and Edwards (1967). On the other hand, $(\delta\mu)^2$ was able to separate African from non-African populations, indicating that for groups at this level of divergence it can be used for reconstruction of population relationships. In the simulation study of Takezaki and Nei (1996), the coefficients of variation for $(\delta\mu)^2$ (and D_{sw}) were consistently greater than for the traditional allele-frequency-based distances, although $(\delta\mu)^2$ was relatively insensitive to changes in population size.

Zhivotovsky and Feldman (1995) showed that with complete sampling, the variance of $(\delta\mu)^2$ between two populations t generations after their divergence from a common ancestral population, with each in mutation-drift equilibrium according to a symmetric SMM, is

$$\mathrm{Var}(\delta\mu)^2 = 8w^2t^2. \tag{9}$$

Thus, if $(\delta\mu)^2$ is computed as the average over L loci of the population values at each locus,

all satisfying (9), and if all loci have the same mutation rate and mutational variance, then the coefficient of variation

$$CV_{(\delta\mu)^2} = \sqrt{2/L}. \tag{10}$$

If the mutation parameter w (recall $w = \mu\sigma_m^2$) varies among the loci, then, according to Zhivotovsky and Feldman (1995, equn 25),

$$CV_{(\delta\mu)^2} = \{2(1 + CV_w)/L\}^{\frac{1}{2}}, \tag{11}$$

which suggests an empirical method of evaluating CV_w, the coefficient of variation among the w values for each locus, from an observed $CV_{(\delta\mu)^2}$ (Reich and Goldstein 1998).

We have compared the variance and mean of the corrected $(\widehat{\delta\mu})^2$ (i.e. equn 7) as a function of time since separation and the mutation rate, using a coalescent simulation of a simple symmetric SMM at 30 loci that were absolutely linked or completely unlinked, but all with the same mutation rate. Following separation of an initial population into two populations (identical to the progenitor population) that subsequently evolve independently, the value of $CV_{(\delta\mu)^2}$ is tracked over time. Table 1 reports the approximate time τ in units of $2N$ generations at which $CV_{(\delta\mu)^2}$ settles around the value $\sqrt{2/30} = 0.258$ predicted for unlinked loci. Linkage delays this convergence by retarding the convergence of the variance. As suggested by Takezaki and Nei (1996), smaller values of $N\mu$ produce faster approach of $CV_{(\delta\mu)^2}$ to the asymptote $\sqrt{2/L}$ with L loci, although in all cases convergence is slow.

[Table 1 here]

Application of $(\delta\mu)^2$

Removing the aberrant locus from the data set of Bowcock et al. (1994), the average variance across the remaining 29 dinucleotide loci (pooling all individuals from all 14 populations) was 6.827. The fourteen populations were grouped by Goldstein et al. (1995b) into seven clusters: I Australian, New Guinean. II Central African Republic Pygmies,

Lisongo. III Chinese, Japanese, Cambodian. IV Northern Italian, Northern European. V Karitiana, Surui, Mayan. VI Melanesian. VII Zaire Pygmies. In their analysis, Goldstein et al. included all 30 loci, giving an average variance of 10.1, and in their calculations of $(\widehat{\delta\mu})^2$ they did not correct for small sample size. Table 2 reports the 21 pairwise values of $(\delta\mu)^2$ corrected for small sample size and including 29 loci. Table 3 gives the same matrix for 85 dinucleotide loci.

[Table 2 here]

The average in Table 2 of the distances between African and non-African populations is 4.270, while from Table 3 it is 4.326. These values are remarkably similar, and differ substantially from the value 6.47 used originally by Goldstein et al., primarily because loci with extremely high variances have been omitted here. From equations (8), with knowledge of the mutation rate and the mutational variance, the time of divergence of the populations can be estimated from the observed $(\delta\mu)^2$ value. Using the value 4.326, a generation time of 27 years (Weiss, 1973), a mutation rate of 6.0×10^{-4} and mutational variance of 2.5, the estimated time of divergence of African and non-African populations is about 38,900 years.

[Table 3 here]

Table 4 reports the pairwise $(\widehat{\delta\mu})^2$ values for 22 tri- and 21 tetranucleotide polymorphisms studied by Bennett et al. (1998) in exactly the same individuals for which the distances for the 85 (and 29) dinucleotide loci are presented in Tables 3 (and 2). Under the assumptions of our analysis, the ratio $(\delta\mu)^2_{\rm di}/(\delta\mu)^2_{\rm tetra}$, for example, should be $w_{\rm di}/w_{\rm tetra}$. For the African-nonAfrican distance, the tetranucleotide value of $(\widehat{\delta\mu})^2$ is 2.137. The ratio $w_{\rm di}/w_{\rm tetra}$ is, therefore, estimated to be $4.326/2.137 = 2.024$. Recall that from the world (pooled) variances we estimated $w_{\rm di}/w_{\rm tetra}$ as 2.086. For the tetranucleotides, the variance and distance calculations thus produce consistent ratios of the compound mutation parameter. For the trinucleotides, however, the average African-nonAfrican distance in Table 4 is

only 1.044, giving an estimate for $w_{\rm di}/w_{\rm tri}$ of 4.144, much greater than the estimate 2.166 obtained from the variances.

[Table 4 here]

The assumption underlying the use of $(\delta\mu)^2$ entails that the two populations being compared have evolved independently since their formation from one ancestral population. An alternative view of the distance between two groups is an inverse measure of the extent of admixture or gene flow between them. We now discuss how statistics like $(\delta\mu)^2$ and the within population variances produce estimates of gene flow for a special migration model.

Dynamics of Microsatellites in an Island Model

Consider a set of d populations in each of which we follow one locus subject to a simple symmetric SMM. All populations have N diploid individuals, and the mutation rate μ is the same in all locations. The migration parameter m is defined as follows: After mutation each population is constructed by taking a fraction $1-m$ of that population and a fraction m made up of the average of all d populations after mutation. Thus, if $n_i^k(t)$ is the number of chromosomes with i repeats in population k at time t, the next generation in population k is produced by multinomial sampling of $2N$ chromosomes using the probabilities

$$\begin{aligned}\pi_i^k(t) = \frac{1-m}{2N}\left\{(1-\mu)n_i^k(t) + \frac{\mu}{2}\left[n_{i-1}^k(t) + n_{i+1}^k(t)\right]\right\} \\ + \frac{m}{2Nd}\sum_{l=1}^{d}\left\{(1-\mu)n_i^l(t) + \frac{\mu}{2}\left[n_{i-1}^l(t) + n_{i+1}^l(t)\right]\right\}.\end{aligned} \tag{12}$$

The new average repeat score in population k, $r_k(t+1)$, given those at time t, becomes

$$E\big[r_k(t+1 \mid t)\big] = (1-m)r_k(t) + \frac{m}{d}\sum_{l=1}^{d} r_l(t). \tag{13}$$

Hence

$$E\big[r_k(t+1) - r_l(t+1) \mid t\big] = (1-m)\big[r_k(t) - r_l(t)\big],$$

and all populations are expected eventually to have the same average repeat score.

Following the analysis by Moran (1975, as used by Goldstein et al., 1995), write $V_k(t)$ for the variance in repeat scores in population k at time t. Then, after a considerable amount of algebra, we find

$$EV_k(t+1 \mid t) = \left(1 - \frac{1}{2N}\right)\left\{(1-m)V_k(t) + \frac{m}{d}\sum_{l=1}^{d} V_l(t) + \mu \right. \\ \left. + \frac{m}{d}\sum_{l=1}^{d}\big[r_k(t) - r_l(t)\big]^2 - \frac{m^2}{d^2}\left[\sum_{l=1}^{d}\big(r_k(t) - r_l(t)\big)\right]^2\right\} \tag{14}$$

Write $\tilde{V}(t) = \sum_{k=1}^{d} V_k(t)$. Then (14) may be rewritten as

$$E\tilde{V}(t+1 \mid t) = \left(1 - \frac{1}{2N}\right)\left\{\tilde{V}_k(t) + \mu d + \frac{m}{d}\left(1 - \frac{m}{d}\right)F(t) - \frac{m^2}{d^2}G(t)\right\}, \tag{15}$$

where

$$F(t) = \sum_k \sum_j \big[r_k(t) - r_j(t)\big]^2$$

and

$$G(t) = \sum_k \sum_{\substack{j \\ j \neq l}} \sum_l \big[r_k(t) - r_j(t)\big]\big[r_k(t) - r_l(t)\big].$$

The reader is now referred back to equn (7) where D_1 is defined as the average of the squared difference between pairs of alleles one chosen from each of two populations. For populations k and l we define analogously

$$D_1^{kl}(t) = \sum_i \sum_j \big[i^{(k)} - j^{(l)}\big]^2 n_i^k(t) n_j^l(t)/(2N)^2$$

where $n_i^k(t)$ and $n_j^l(t)$ are the numbers with length i and j, respectively, in populations k and l, respectively, and the sum

$$\tilde{D}_1(t) = \sum_{\substack{k \\ k \neq l}} \sum_{l} D_1^{kl}(t).$$

After a considerable amount of algebra we have the elegant simplification

$$E\tilde{D}_1(t+1 \mid t) = D_1(t) - \frac{2m}{d}\left(1 - \frac{m}{d}\right) F(t) + \frac{2m^2}{d^2} G(t) + 2\mu d(d-1). \tag{16}$$

To complete the iteration of (15) and (16) we need recursions for $F(t)$ and $G(t)$. These are the most algebraically demanding but lead ultimately to the following iterations

$$\begin{aligned} EF(t+1 \mid t) = F(t) &\left[(1-m)^2 - \frac{1}{2N}\left(1 - \frac{2m}{d} + \frac{2m^2(d-1)}{d^2}\right)\right] \\ &+ \frac{\tilde{D}_1(t)}{2N} - \frac{2m^2(d-1)G(t)}{2Nd^2} + \frac{2\mu d(d-1)}{N} \end{aligned} \tag{17}$$

and

$$\begin{aligned} EG(t+1 \mid t) = \tilde{V}(t)&\frac{(d-1)(d-2)}{2N} \\ &+ F(t)\left\{\frac{(d-2)(1-m)^2}{2} + \frac{(d-2)(d-1)}{2N}\frac{m}{d}\left(1 - \frac{m}{d}\right)\right\} \\ &- \frac{(d-2)(d-1)m^2}{2Nd^2} G(t) + \frac{\mu d(d-1)(d-2)}{2N}. \end{aligned} \tag{18}$$

The system (15), (16), (17) and (18) provides a complete linear recursion for the expectations of moments $\tilde{V}(t)$, $\tilde{D}_1(t)$, $F(t)$ and $G(t)$.

On inspection of (15) and (16), we see the first equilibrium result, namely

$$E\tilde{V} = (2N-1)\mu d^2.$$

Notice that with two populations (i.e., $d = 2$), $E(\tilde{V}) = E(V_1 + V_2) = 2E(V_1) = 2E(V_2) = 4\mu(2N-1)$ so that in each population, the expected variance is $\hat{V}_A = 2\mu(2N-1)$. In other words, with two populations exchanging immigrants at any positive rate, the equilibrium

repeat score variance in each is double that in the absence of gene flow (equn (5) above; Moran 1975; Goldstein et al. 1995*a*). Pritchard and Feldman (1996) also observed this phenomenon.

Equilibrium values for $E\tilde{D}_1$, EF and EG are obtained on inversion of the complete linear system (15), (16), (17), (18). These equilibria are given below:

$$E\tilde{V} = \mu(2N-1)d^2 \tag{19a}$$

$$E\tilde{D}_1 = \frac{2\mu d(d-1)\left\{d\left[1-m^2(2N-1)+4mN\right]-4m\right\}}{m(2-m)} \tag{19b}$$

$$EF = \frac{2\mu(d-1)d^2}{m(2-m)} \tag{19c}$$

$$EG = \frac{\mu d^2(d-1)(d-2)}{m(2-m)} \tag{19d}$$

It is worth noting that $E(\tilde{D}_1)$ may be expressed as a linear combination of $E(\tilde{D}_0)$ and $E(F)$, namely $E(\tilde{D}_1) = \beta_1 E(\tilde{V}) + \beta_2 E(F)$, where $\beta_1 = 2(d-1)$ and $\beta_2 = 1+2m(1-2/d)$. With two populations, $\beta_1 = 2$, $\beta_1 = 1$ and we have a result equivalent to equn (7) above.

In analyses of the island model, it is usual to ignore terms of order m^2 and smaller, to set $2N-1$ to $2N$ and to ignore m relative to Nm. Then, to this degree of approximation, from (19b)

$$E\tilde{D}_1 \simeq \frac{\mu d^2(d-1)(1+4Nm)}{m}, \tag{20}$$

so the average pairwise value of D_1 at equilibrium is

$$\overline{D}_1 = \frac{1}{d(d-1)} E\tilde{D}_1 = \frac{\mu d(1+4Nm)}{m} \tag{21}$$

Also, from (19a), the equilibrium average within population variance is

$$\overline{V} = E(\tilde{V}/d) \simeq 2N\mu d. \tag{22}$$

¿From (21) and (22), using the definition in equn (7), the average equilibrium value of $(\delta\mu)^2$ over all pairs of populations is

$$\overline{(\delta\mu)^2} = \overline{D}_1 - 2\overline{V} = \frac{\mu d(1+4Nm)}{m} - 4N\mu d = \mu d/m \tag{23}$$

Finally we have

$$L_m = \overline{V}/\overline{(\delta\mu)^2} = 2Nm. \tag{24}$$

In other words, at equilibrium, the ratio L_m of the average variance within populations to the average distance between populations provides an estimate of $2Nm$.

In order to reconcile this approximation with Slatkin's (1995) use of R_{ST}, we will assume that the size n of the sample from each of the d sampled populations is large enough that we may write $(2n-1)/(2nd-1) \simeq d^{-1}$ and $2n(d-1)/(2nd-1) \simeq (d-1)/d$. Then, as indicated by Slatkin (1995), his S_W is equivalent to $2\overline{V}$ in our notation while his S_B is equivalent to our $\tilde{D}_1/d(d-1)$ and, according to his equn (15),

$$\left[\frac{1}{R_{ST}} - 1\right] = \frac{d}{d-1}\left[\frac{S_W}{S_B - S_W}\right] = \frac{d}{d-1}\left[\frac{2\overline{V}}{\overline{D}_1 - 2\overline{V}}\right].$$

His M_R, used to estimate Nm, is defined by

$$M_R = \frac{d-1}{4d}\left(\frac{1}{R_{ST}} - 1\right) = \frac{1}{2}\left(\frac{\overline{V}}{\overline{D}_1 - 2\overline{V}}\right) \tag{25}$$

in our terminology, and referring to equn (24), under the approximations we have made, $\overline{V}/(\overline{D}_1 - 2\overline{V}) = \overline{V}/\overline{(\delta\mu)^2}$ estimates $2Nm$. L_m, defined by $\overline{V}/\overline{(\delta\mu)^2}$, is suggested as an estimate of $2Nm$ when sample sizes are at least moderately large, but not necessarily equal. $2M_R$ (in (25)) was originally derived under the assumption of equal sample sizes, but in light of (25), for moderately large samples, the values L_m and $2M_R$ should be fairly close.

If the samples from each population are not really small, the estimate given by (24) of the average variance to the average pairwise genetic distance should be adequate to estimate $2Nm$ in this island model.

In the case of just two populations ($d = 2$), it is usual to express admixture in terms of a parameter ν, say, where population I in generation $t+1$ is produced by fractions $1-\nu$ from population I and ν from population II at time t. Thus, ν is the extent of admixture, and in the model the admixture continues at each generation during the evolution. In terms of the parameter m from the island model, we have $\nu = m/2$. For two populations, the dynamics under migration are simpler. Write H for $(r_1 - r_2)^2$. Then we have

$$E\big[D_1(t+1) \mid t\big] = D_1(t) - 2\nu(1-\nu)H(t) + 2\mu \tag{26a}$$

$$E\big[H(t+1) \mid t\big] = D_1(t)/2N + \delta H(t) + 2\mu/2N \tag{26b}$$

where $\delta = (1-2\nu)^2 - (1-2\nu+2\nu^2)/2N$. Hence there is convergence to

$$E(D_1) = 2\mu\left\{\frac{1}{2\nu(1-\nu)} + 2(2N-1)\right\} \tag{27a}$$

$$E(H) = \frac{\mu}{\nu(1-\nu)} \tag{27b}$$

Recalling that $\tilde{D}_1$ in (19b) is $2D_1$ in (27a) and F in (19c) is $2H$, these equations (27) are the same as the equilibria (19b) and (19c). Of course, as in (19a), $E(\tilde{V})$, which is the sum of the expected variances in populations I and II, comes to $4\mu(2N-1)$. Substituting this into (27a) and (27b), we obtain the two-population version of equn (24); $4N\nu = 2Nm$ is estimated by the average of the two populations' variances divided by the estimate of $(\delta\mu)^2$ between them. This ratio may be regarded as a measure of population affinity; the more migration (i.e. admixture) the greater is the ratio (and the smaller the standardized distance).

More on statistical measures of subdivision

Both L_m and $2M_R$ were shown in the previous section to estimate $2Nm$ for an island model of population subdivision with migration. Although the model of mutation considered here is specifically chosen to represent microsatellites, measures similar to L_m and R_{ST} have been used for many years to partition variation in gene frequencies into

within and between group components. The most widely used measures are F-statistics; for example the classical island model produces an equilibrium F_{ST} of $(1+4Nm)^{-1}$.

These F-statistics involve analysis of variance in allele frequencies with a rationale similar to the classical random effects that ANOVA models. The partition is described in some detail by Weir (1996, pp. 170–184) and has been adapted in a straightforward way by Michalakis and Excoffier (1996) to the case of microsatellites. Here, the repeat score of allele j from population i $(i = 1, 2, \cdots, d)$ is viewed as an observation on the random variable $Y_{ij} = \mu + A_i + \varepsilon_{ij}$, where μ is an overall mean, A_i is a random variable representing the random effect of population i, with zero expectation and variance σ_A^2, and ε_{ij} is an error random variable within population i, with zero mean and variance σ_w^2. In this framework, the intraclass correlation coefficient $\theta = \sigma_A^2/(\sigma_A^2 + \sigma_w^2)$ tells us the relative magnitude of between-group contribution to the total variation (Weir, 1996). The standard estimator $\hat{\theta}$ of θ is given by

$$\hat{\theta} = \frac{MSB - MSW}{MSB + (n_0 - 1)MSW}\,, \tag{28}$$

where MSB and MSW are the mean squared deviations between and within populations, respectively,

$$n_0 = \frac{1}{d-1}\left[2\mathcal{N} - \frac{\sum_{i=1}^{d}(2n_i)^2}{2\mathcal{N}}\right], \tag{28a}$$

with $2n_i$ the number of chromosomes sampled from population i, and $2\mathcal{N} = \sum_{i=1}^{d} 2n_i$ the total number of chromosomes sampled. When the samples from each population are the same size, i.e. $2n_i = 2n$ for $i = 1, 2, \cdots, d$, n_0 reduces to $2n$. Michalakis and Excoffier (1996) denote the ratio of the averages over loci of the numerators in (28) to the average over loci of the denominators in (28) as $\hat{\phi}_{ST}$ and claim that $\hat{\phi}_{ST} = (1-c)R_{ST}/(1-cR_{ST})$, where with equal n_i, $c = (2n-1)/(2nd-1)$.

Slatkin's original formulation of his R_{ST} was in terms of quantities S_B, S_W, and $\bar{S}$ with

$$R_{ST} = \frac{\bar{S} - S_W}{\bar{S}} = \frac{S_B - S_W}{S_B + \frac{2n-1}{2n(d-1)} S_W}. \tag{29}$$

As is remarked by Slatkin, $\bar{S}$ and S_W have interpretations in terms of the variance in the whole system (without partitioning), and the average variance within groups, respectively. Under this interpretation, if we use the notation standard in analysis of variance, and follow the suggestions of Slatkin after his equn (10), then

$$R_{ST} = \frac{MSB - MSW}{MSB + \frac{d(2n-1)}{d-1} MSW}. \tag{30}$$

Then, comparing (28) and (30) we see that

$$\hat{\phi}_{ST} = \frac{R_{ST}}{1 - c + cR_{ST}}. \tag{31}$$

¿From (31) it is obvious that

$$R_{ST} = \frac{(1-c)\hat{\phi}_{ST}}{1 - c\hat{\phi}_{ST}} \tag{32}$$

which is the same as equn (9) of Michalakis and Excoffier (1996) but with R and $\hat{\phi}$ interchanged. Note that the correct version (32) is given by Rousset (1996, equn 17). It is worth noting that when $n_i = n$, the expectation of R_{ST} is $[1 + 4Nmd/(d-1)]^{-1}$ while that of $\hat{\phi}_{ST}$ is $[1 + 4Nm(\frac{2N}{2N-1})]^{-1} \approx (1 + 4Nm)^{-1}$.

Both R_{ST} and $(\delta\mu)^2$ have been used as genetic distances from which to construct evolutionary trees of relationship among populations. From equns (24) and (25), we see that in an island model (and in particular, in a two-population model of admixture, as shown with equns (27)), M_R or $L_m/2$ both estimate Nm. Nm may be regarded as an index of closeness between two populations. By the same token, a matrix of pairwise values of $(M_R)^{-1}$ or $(L_m/2)^{-1}$ may be treated as a set of distances from which statistical clustering of the populations may be visualized. Figures 1 and 2 represent such trees computed using 85 dinucleotide loci (Fig. 2) from Li Jin et al. (1998). The 85 loci produce clusters that

match the earlier allele-sharing tree quite closely, with $(L_m/2)^{-1}$ apparently closer than $(M_R)^{-1}$ to the tree obtained by Bowcock et al. (1994) using an allele-sharing distance on 30 dinucleotide loci.

Models With Mutational Constraints

Two departures from Ohta and Kimura's infinte range SMM have been proposed. One sets a finite interval of length R in which repeat numbers vary and forces alleles at the boundaries to mutate only to the interior of the range. This might be termed a *hard boundary* model. The other model takes a focal repeat score r_m and assumes that repeat scores greater than r_m tend to mutate to lower scores while those lower than r_m mutate upwards. This introduces a bias b_i in the mutation from alleles with i repeats which is usually assumed to be linear: $b_i = B(i - r_m)$ with $B < 0$. This model might be described as having *soft* boundaries.

For the hard boundary model, Goldstein et al. conjectured that the expected value of D_1 (see equn 7) would approach the equilibrium $\frac{(R^2-1)}{6}$. This was shown to be the case by Nauta and Weissing (1996). Because D_1 converges, in order to construct a distance that is linear with time for a reasonable period, a correction must be made to $(\delta\mu)^2$. Feldman et al. (1997) suggest a distance

$$D_L = \log\left[1 - \sum_{i=1}^{L}(\delta\mu)_i^2/LM\right] \tag{33}$$

with L loci, where M is the average value of the distance at maximal divergence. D_L, which was derived in the special case where R and μ do not vary across loci, is linear for a reasonable period of time which increases as the number of loci, L, increases. Pollock et al. (1998) have improved on D_L by using a weighted least squares technique originally introduced by Goldstein and Pollock (1994). Interestingly, these corrected distances do not seem to be sensitive to moderate variation in the repeat range or the mutation rate.

The linear soft boundary model was introduced by Garza et al. (1995), who compared average allele sizes in humans and chimpanzees and found that these were sufficiently similar that some kind of evolutionary constraint on repeat lengths was likely to have been

in effect. They concluded from their analysis that the bias, measured by B, is likely to have been quite weak, substantially less than the mutation rate itself. A different mathematical approach to the same model was taken by Zhivotovsky et al. (1997), who developed an estimator for B of the form

$$\hat{B} = \frac{-\sigma_m^2}{2\big[\mathrm{Var}(\bar{r}) + \bar{V}\big]}\,, \tag{34}$$

where $\bar{r}$ estimates the mean repeat number at a locus, $\mathrm{Var}(\bar{r})$ is the variance across loci of these estimated means, $\bar{V}$ is the average over loci of the within-locus variances, and σ_m^2 is the usual variance in mutation sizes. The data of Garza et al., when inserted into (34), produced $\hat{B}$ between -0.0064 and -0.013, substantially higher than the mutation rate μ. Their analysis also allowed Zhivotovsky et al. to estimate the time since two populations subject to this biased mutation diverged from a common ancestral population. Assuming $\sigma_m^2 = 2.0$, for example, and $B = -0.02$, the data of Bowcock et al. (1994) gave an estimated divergence time for African and non-African populations of 100,000 years. With $B = -0.0064$, the estimate was 84,000 years. In both cases, a 27-year generation was used. Estimated divergence times are sensitive to σ_m^2 as well as to B.

References

Bennett, L., E. Minch, M. Feldman, T. Jenkins, and A. Bowcock. 1998. A set of independent tri- and tetranucleotide repeats for human evolutionary studies; additional evidence for an African origin for modern humans and implications for higher mutation rates in tetranucleotide repeats of the "GATA" type. In preparation.

Bowcock, A.M., A.R. Linares, J. Tomfohrde, E. Minch, and J.R. Kidd *et al.* 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455–457.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**: 152–154.

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

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.

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(4)**: 594–603.

Goldstein, D.B., and D.D. Pollock. 1994. Least squares estimation of molecular distance—noise abatement in phylogenetic reconstruction. *Theor. Pop. Biol.* **45**: 219–226.

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.

Goldstein, D.B., A.R. Linares, L.L. Cavalli-Sforza, and M.W. Feldman. 1995*b*. Microsatellite loci, genetic distances and human evolution. *Proc. Natl. Acad. Sci. USA* **92**: 6723–6727.

Goldstein, D.B., and D.D. Pollock. 1997. Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. *J. Heredity* **88**: 335–342.

Goldstein, D.B., L. Zhivotovsky, K. Nayar, A.R. Linares, L.L. Cavalli-Sforza, and M.W. Feldman. 1996. Statistical properties of variation at linked microsatellite loci: implications for the history of human Y chromosomes. *Mol. Biol. Evol.* **13**: 1213–1218.

Heyer, E., J. Puymirat, P. Dieltjes, E. Bakker, and P. de Knijff. 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* **6(5)**: 799–803.

Kimmel, M., R. Chakraborty, D.N. Stivers, and R. Deka. 1996. Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. *Genetics* **143**: 549–555.

Li Jin et al. 1998. Unpublished data.

Michalakis, Y., and L. Excoffier. 1996. A generic 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. Pop. 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. 1972. Genetic distance between populations. *Am. Nat.* **106**: 283–292.

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

Pollock, D.D., A. Bergman, M.W. Feldman, and D.B. Goldstein. 1998. Microsatellite behavior with range constraints: parameter estimation and improved distance estimation for use in phylogenetic reconstruction. *Theor. Pop. Biol.* To appear.

Pritchard, J.K., and M.W. Feldman. 1996. Statistics for microsatellite variation based on coalescence. 1996. *Theor. Pop. Biol.* **50**: 325–344.

Roe, A. 1992. Correlations and Interactions in Random Walks and Population Genetics. Ph.D. thesis, University of London, London, U.K.

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

Shriver, M.D., R. Jin, E. Boerwinkle, R. Deka, R.E. Ferrell, and R. Chakraborty. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol. Biol. Evol.* **12**: 914–920.

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

Takezaki, N., and M. Nei. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144**: 389–399.

Weber, J.L., and C. Wong. 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**: 1123–1128.

Weir, B.S. 1996. *Genetic Data Analysis II.* Sinauer Associates, Inc. Sunderland, Massachusetts.

Weiss, K. 1973. Demographic models for anthropology. *Am. Antiq.* **38**: 1–186.

Wright, S. 1943. Isolation by distance. *Genetics* **28**: 114–138.

Zhivotovsky, L.A., and M.W. Feldman. 1995. Microsatellite variability and genetic distances. *Proc. Natl. Acad. Sci. USA* **92**: 11,549-11,552.

Zhivotovsky, L.A., M.W. Feldman, and S.A. Grishechkin. 1997. Biased mutations and microsatellite variation. *Mol. Biol. Evol.* **14**: 926–933.

Table 1. Times to Convergence of $CV_{(\delta\mu)^2}$ to $\sqrt{2/L}$

	Linked	Unlinked
$\theta = 10$	$\tau \simeq 70$	$\tau \simeq 40$
$\theta = 0.1$	$\tau \simeq 20$	$\tau \simeq 10$

Table 2. Corrected $(\delta\mu)^2$ distances for 29 dinucleotide loci*

	I	II	III	IV	V	VI
I						
II	2.732					
III	0.702	2.292				
IV	1.491	3.102	1.343			
V	1.101	2.618	0.919	1.745		
VI	0.983	2.244	0.751	1.364	0.483	
VII	5.803	2.447	4.466	6.410	5.571	6.459

* Population groups are: I Australian, New Guinean. II Central African Republic Pygmies, Lisongo. III Chinese, Japanese, Cambodian. IV Northern Italian, Northern European. V Karitiana, Surui, Mayan. VI Melanesian. VII Zaire Pygmies.

Table 3. Corrected $(\delta\mu)^2$ distances for 85 dinucleotide loci

	I	II	III	IV	V	VI
I						
II	3.833					
III	1.602	2.507				
IV	2.033	2.721	1.471			
V	2.835	4.132	1.731	2.517		
VI	1.285	5.132	2.081	2.457	2.499	
VII	5.082	2.058	3.664	4.099	5.940	6.145

Table 4. Corrected $(\delta\mu)^2$ distances for 22 tri- and 21 tetranucleotides

	I		II		III		IV		V		VI	
	tri	tetra	tri	tetra	tri	tetra	tri	tetra	tri	tetra	tri	tet
I												
II	1.214	1.446										
III	0.414	0.493	0.939	1.109								
IV	1.147	0.858	0.442	0.608	0.532	0.919						
V	0.580	0.977	1.131	1.761	0.472	0.400	0.678	1.047				
VI	0.669	0.830	1.081	2.009	0.498	1.361	0.967	1.801	0.606	2.327		
VII	1.054	3.047	0.650	0.706	0.940	2.418	1.221	2.203	1.419	3.002	1.000	3.7

Figure Legends

Figure 1: UPGMA tree showing clusters of human populations, based on $\overline{V}/2(\delta\mu)^2$ as in equation 24.

Figure 2: UPGMA tree, showing clusters of human populations, based on $(M_R)^{-1}$ as in equation 25.