Statistical Properties of the Variation at Linked Microsatellite Loci: Implications for the History of Human Y Chromosomes

David B. Goldstein,*1 Lev A. Zhivotovsky,† Kavita Nayar,‡ Andres Ruiz Linares,§
Luigi L. Cavalli-Sforza,‡ and Marcus W. Feldman§

*Institute of Molecular Evolutionary Genetics, Department of Biology, The Pennsylvania State University; †Vavilov Institute of General Genetics, Russian Academy of Sciences; ‡Department of Genetics, Stanford University; and §Department of Biological Sciences, Stanford University

It has recently been suggested that observed levels of variation at microsatellite loci can be used to infer patterns of selection in genomes and to assess demographic history. In order to evaluate the feasibility of these suggestions it is necessary to know something about how levels of variation at microsatellite loci are expected to fluctuate due simply to stochasticity in the processes of mutation and inheritance (genetic sampling). Here we use recently derived properties of the stepwise mutation model to place confidence intervals around the variance in repeat score that is expected at mutation-drift equilibrium and outline a statistical test for whether an observed value differs significantly from expectation. We also develop confidence intervals for the time course of the buildup of variation following a complete elimination of variation, such as might be caused by a selective sweep or an extreme population bottleneck.

We apply these methods to the variation observed at human Y-specific microsatellites. Although a number of authors have suggested the possibility of a very recent sweep, our analyses suggest that a sweep or extreme bottleneck is unlikely to have occurred anytime during the last approximately 74,000 years. To generate this result we use a recently estimated mutation rate for microsatellite loci of 5.6 X 10^-4 along with the variation observed at autosomal microsatellite loci to estimate the human effective population size. This estimate is 18,000, implying an effective number of 4,500 Y chromosomes. One important general conclusion to emerge from this study is that in order to reject mutation-drift equilibrium at a set of linked microsatellite loci it is necessary to have an unreasonably large number of loci unless the observed variance is far below that expected at mutation-drift equilibrium.

Introduction

Because of their rapid rate of evolution, ease of scoring, and abundance, microsatellites are increasingly being used in evolutionary analyses of closely related groups. A number of recent studies suggest that they may prove especially useful in studies of modern human origins. One basic question to be resolved concerns the extent of the genetic contribution to modern humans of Homo erectus populations that were already distributed throughout the world more than one million years ago (Cann, Stoneking, and Wilson 1987; Stringer and Andrews 1988; Cavalli-Sforza, Menozzi, and Piazza 1994). The multiregional theory posits a significant contribution with modern humans evolving simultaneously throughout the world. The Out-of-Africa theory, on the other hand, posits a small group of fully modern humans arising first in Africa and spreading from there to the rest of the world in a short time, interbreeding little, if at all, with the Homo erectus populations already in residence. The emergence from Africa is assumed to have occurred about 100,000 years ago on the basis of archaeological evidence (Mellars and Stringer 1989).

Considered in terms of the rate of evolution of a typical DNA sequence the two theories differ only slightly in the predicted degree of differentiation among human populations. This is one of the reasons it has been difficult to discriminate between the theories. Nevertheless, genetic data have been largely interpreted as supporting an Out-of-Africa theory (Cavalli-Sforza, Menozzi, and Piazza 1994; Nei and Takezaki 1996).

In order to better distinguish between the predictions of the two models researchers have recently turned to faster-evolving microsatellites. Goldstein et al. (1995a, 1995b) developed a calibration-free method allowing “absolute” dating of population splitting times based on variation at microsatellite loci. They applied this approach to 30 autosomal microsatellites (Bowcock et al. 1994) and estimated that the deepest split in the human phylogeny occurred about 156,000 years ago. This result is in tantalizing agreement with paleoanthropologically based estimates of the time of emergence from Africa, and with the most recent and complete study of mtDNA coalescence times (Horai et al. 1995). Other explanations are possible, however. Templeton (1993), for example, has suggested that isolation by distance may be a more accurate model of recent human history than one involving bifurcations and subsequent reproductive isolation. It should be appreciated, however, that isolation by distance would be unlikely to result in the close correspondence between archaeologically based dates and genetic distances which has been recently reported (Mountain et al. 1992; Goldstein et al. 1995b). If this correspondence holds up we must conclude that bifurcation, with subsequent variable but low rates of gene flow, is the most accurate representation of human evolutionary history.

While multilocus studies are required for accurate inferences about population divergence, they average the evolutionary histories of distinct sets of linked loci and may thereby mask important variation among different genomic regions. The degree of variation in evolutionary history of different genomic regions is an im-
portant and largely untapped source of information about human evolutionary history. For example, the Out-of-Africa hypothesis would predict a fairly tight limitation on the amount of variability in coalescence times among loci. If a small population emerged from Africa, and that population were not extensively substructured, then most loci would have maximal coalescence times similar to or somewhat greater than the date of emergence from Africa, except in unusual cases involving selection. On the other hand, if the multiregional hypothesis were correct, some loci should have much older coalescence times.

A comparison of the coalescence histories of different loci therefore adds an important dimension to the genetic evidence on the origins of modern humans. In this paper we discuss how the variation observed at microsatellite loci might be used for this purpose, making particular reference to Y-specific variation. Dorit, Akashi, and Gilbert (1995) recently reported no sequence variation among 38 human Y chromosomes from which they estimated an expected maximal coalescent time of 270,000 years. The distribution underlying this estimate, however, is such that its expectation is not particularly informative (Donnelly and Tavaré 1996). The lack of statistical power is evidenced by the confidence interval on the date of a putative sweep, which ranges from 0 to 800,000 years. By using quickly evolving microsatellites, we achieve greater statistical reliability and show that a sweep is unlikely to have occurred in the recent past. Microsatellites suffer their own limitations, however. In particular, we shall see that the variance maintained at a microsatellite locus fluctuates greatly due simply to stochasticity in the processes of mutation and drift, and that this fluctuation makes it difficult to show positive evidence for a selective sweep or bottleneck unless a very large number of loci are available or the observed variance is very different from that expected at mutation-drift equilibrium. Thus, unless many more than five completely linked loci are available, Y-specific microsatellites cannot reliably date a bottleneck or sweep.

### Data and Analysis

Typings were performed with three pairs of primers that amplify regions of the nonrecombining portion of the Y chromosome. Two pairs of primers each resulted in two Y-specific bands, each presumably corresponding to a duplicated dinucleotide repeat. Another primer pair resulted in a single Y-specific band that included a tetranucleotide repeat. We determined fragment sizes for each of the primer pairs for 121 males from 13 worldwide populations. A complete description of the experimental methods as well as the raw data can be found in Ruiz Linares et al. (1996).

Although the products of the duplicated loci cannot be perfectly separated, a plot of the distributions of the two bands suggests that they are nearly distinct (data not shown). Here we assume that for each chromosome typed the smaller fragment of the pair is the product of one locus, and the larger fragment the product of the other locus. This may result in an underestimate of the variation, but, because our primary focus is to estimate the most recent possible bottleneck consistent with the observed variation, this bias results in conservative estimates. Furthermore, for the purposes of our analysis, the individuals were considered as if drawn from a single population. With these assumptions, the average variance in repeat score across the five Y chromosome microsatellites is 2.0.

Under the stepwise-mutation model introduced by Ohta and Kimura (1973), Moran (1975) showed that the expected variance in microsatellite repeat numbers at mutation-drift equilibrium is

\[ \hat{V} = (N - 1)\mu, \]

where the hat indicates mutation-drift equilibrium, \( N \) is the effective number of gametes, and mutation causes a change of +1 and -1 in repeat number each with probability \( \mu/2 \).

To generate an expectation for the variance of Y-specific microsatellites we can use the characterization of 30 autosomal microsatellites in Bowcock et al. (1994). Again treating all individuals as if drawn from a single population, the average variance is 10.1. To derive the expected variance at mutation-drift equilibrium for Y-specific microsatellites from that of the autosomal microsatellites note that the expectation of the variance is a linear function of \( N \) (eq. 1) and that there are four autosomes for every Y chromosome in human populations. The effective population size for the Y, therefore, is \( 1/2 \) that for the autosomes and by the linearity in \( N \) the expected variance is \( 1/2 \) that on the autosomes (10.1) or about 2.52. This expectation assumes a similar mutation rate for autosomal and Y microsatellites.

From the analysis of the stepwise mutation model in Moran (1975) the dynamic for the variance in repeat score can be shown to be

\[ V_t = V_0(1 - 1/N)^t + (N - 1)(1 - (1 - 1/N)^t)\mu \]

(2)

where \( V_0 \) is the initial variance in the population at time \( t = 0 \). Related formulations may be found in Tajima (1989) and Slatkin (1995). Under the assumption of an initial expansion from a single Y chromosome, \( V_0 = 0 \). In this case, equation (2) could be used to calculate the expected time that has passed since an assumed elimination of variation.

It is tempting to use the dynamic given in equation (2) to estimate the expected amount of time it would take to build up a variance of 2.0 (that is, the time that has passed since a complete elimination of variation of human Y chromosomes) and perhaps relate this time to the date of emergence from Africa. Unfortunately, this is not an appropriate calculation unless one first shows that the observed variance is less than that expected at mutation-drift equilibrium. Otherwise, the variance may have long since reached its asymptotic value at which point there is no relationship between time and the variability.

In order to determine whether the observed variance is significantly different from that expected at mutation-drift equilibrium we must consider the range of
values that the variance at a microsatellite locus is likely to assume. With respect to the processes of mutation and repeated sampling, the population variance in repeat numbers is a random variable, denoted \( V_r \), whose expectation, \( \hat{V} \), at mutation-drift equilibrium is given in equation (1). The actual variance will vary from one realization of the evolutionary process to another due to stochasticity in the mutation and sampling processes. We use \( V_g(V_r) \) to denote the variance of \( V_r \) across a large number of potential replicate populations, where \( g \) refers to the process of genetic sampling, i.e., random drift and mutation (Weir 1990).

A recursion for the expectation of \( V_g(V_r) \) has recently been derived and its value at mutation-drift equilibrium obtained (Zhivotovsky and Feldman 1995). We will also require \( V_g \), the variance about \( V_r \) as a function of time since an elimination of variation. Zhivotovsky and Feldman (1995) present discrete dynamics for the expected variance, the expected squared variance \( W \), and the expected kurtosis \( K \) in a population undergoing the stepwise-mutation model. Continuous time approximations to these are the differential equations

\[
\begin{align*}
NV &= \hat{V} - \bar{V} \\
NW &= 2\hat{V}V + K - 3W \\
NK &= 6\hat{V}V - 4K + \hat{V} + 6W
\end{align*}
\]

where \( \hat{V} = (N - 1)\mu \) is the equilibrium expected variance. From these we derive a solution \( V_g(t) = V_g(Nt^*) \) at generation \( t^* \), where \( t^* \) is now scaled by the population size:

\[
V_g(t) = \frac{1}{5} \left( \hat{V} + 12\hat{V}^2(1 - e^{-t}) - \frac{1}{6}(\hat{V} + 2\hat{V}^2)(1 - e^{-6t}) + \frac{2}{5}\hat{V}^2e^{-5t}(1 - e^{-5t}) - 12\hat{V}^2te^{-t} \right).
\]  (3)

We use these expressions to generate confidence intervals for the variance, but first we must address two complications. First, the \( V_r \)'s at linked loci may covary. Zhivotovsky and Feldman (1995) have shown that the covariance is negligible if the recombination rate is large relative to the inverse of the population size. For smaller rates of recombination, however, the covariance may be larger. Since we are interested in gene histories we must consider the case of no recombination. Using computer simulations with parameters appropriate for the Y chromosome we have shown that the covariance among perfectly linked loci is approximately one order of magnitude less than \( V_r \) for the full time course of the buildup of variation. For example, we ran 500 independent simulations with two fully linked loci with a haploid population size of 4,000 and a mutation rate of 0.001. After 5,000 generations the variance in repeat score, averaged both over replications and over the two loci, was \( V_r = 2.8 \), while the average variance of \( V_r \) across the replicate simulations was \( V_g(V_r) = 5.2 \). The covariance between the two loci was 0.61, or 11% of \( V_g(V_r) \). After 18,000 generations \( V_r = 4.01 \) has reached its equilibrium value, \( V_g(V_r) = 21 \), and the covariance is 3.3, or about 15% of \( V_g(V_r) = 21 \). Thus, the covariance is substantially smaller than \( V_g(V_r) \) throughout the trajectory, and can be expected to have a relatively small impact on the confidence intervals. This expectation is confirmed by a computer simulation that directly estimates the confidence intervals, as described below. Although we will ignore the covariance in our analyses, we emphasize that its relative magnitude may depend on specific parameter values. In general, its magnitude should be assessed and incorporated as necessary.

A second complication in developing confidence intervals about \( V_r \) is that the distribution of \( V_r \) (over evolutionary outcomes) is not normal (fig. 1A). The dependence of \( V_g(V_r) \) on \( V_r \), however, suggests that a log
The equation for the expectation (solid curve) is given as equation (2) in the text, and expressions for the confidence intervals are given in the Appendix. The parameters are those assumed throughout the text.

transformation will make the distribution of \( V_r \) more normal (see Appendix), as illustrated in figure 1B. Therefore, we use the variance of the transformed variable to derive confidence intervals, and test the accuracy of this method by Monte Carlo simulations.

Our calculations will require estimates of both the effective population size and the mutation rate. The average mutation rate at autosomal dinucleotide microsatellites has recently been estimated as \( 5.6 \times 10^{-4} \) (Weber and Wong 1993). (The tetranucleotide repeat may have a different mutation rate, but we assume the same rate as for dinucleotide repeats since we do not have a reasonable estimate for the former.) The effective population size, \( N_e \), can be estimated using this mutation rate and the observed variation at the 30 autosomal loci using equation (1). We estimate \( N_e \) as about 18,000, a number consistent with some other estimates (Harpending et al. 1993). This implies an effective population size for Y chromosomes of 4,500, and an equilibrium variance in repeat score (\( \bar{V}_r \)) of about 2.52 (as discussed above).

We estimate (see Appendix) the confidence interval for the population variance at mutation-drift equilibrium as 0.76–6.31, which includes our observation of 2.0. In other words, we find no evidence that Y microsatellites are out of mutation-drift equilibrium. Therefore, there is no statistical basis for relating the expected time calculated from equation (1) to the coalescent time for human Y chromosomes.

Nevertheless, it remains appropriate to place a boundary on the most recent possible sweep consistent with the observed variation. The logic behind such a calculation is illustrated in figure 2, which plots the expectation of the variance as a function of time following an elimination of variation along with the 95% confidence intervals (see Appendix) for the variance. Notice that early in the evolution, for the first 2,750 generations, the 95% confidence interval is restricted to a band that is below 2.0. This implies that a variance of 2.0 would be very unlikely if a sweep or extreme bottleneck had occurred during the last 2,750 generations. We checked this conclusion with computer simulations as follows. We began with a population of 4,500 identical chromosomes and simulated stepwise mutations and drift until a variance of 2.0 was reached, at which time the number of generations was reported. Stepwise mutations occurred at a rate of \( 5.6 \times 10^{-4} \) at each of five completely linked loci. The average time at which the variance of 2.0 was reached was 5,600 generations. The empirically observed 95% confidence interval, based on 400 simulations, was 2,808–12,031. The lower value is in very good agreement with the predictions of our analytic confidence intervals, as represented in figure 2. This confirms that the tails of the transformed variable are sufficiently like that of a normal variate, that a confidence interval can be based on the standard deviation, and that the covariances of the \( V_r \) across linked loci are sufficiently small relative to \( \bar{V}_r \) that they can be ignored. The difference between the average time until 2.0 is first reached (5,800) and that predicted by equation (2) (7,100) is not unexpected. Equation (2) gives the average variance at 7,100 generations as 2.0, while the simulation reports the first time a population hits a value of 2.0. These times are obviously not the same.

In summary we find that microsatellite variation on the Y does not provide any statistical basis for calculating an average date of a bottleneck or selective sweep. It is interesting to note that when the observed variance is not very different from the value expected at mutation-drift equilibrium a tremendous number of loci are required to reject mutation-drift equilibrium. For example, in order to show that 2.0 is significantly less 2.52 we would require about 110 loci. Despite the difficulties with rejecting mutation-drift equilibrium, microsatellites remain informative in the other direction: a very recent sweep or bottleneck can be rejected with only a moderate number of loci. With our current data set of only five loci we demonstrate that a sweep or complete bottleneck is unlikely to have occurred within the last 2,750 generations. To translate the figure into years we use a generation time calculated for contemporary populations with habits thought to be similar to paleolithic ones (Weiss 1973). This gives an estimated generation time of 27 years, which seems large and in any event must be considered very rough. This implies no sweep or bottleneck within the last 74,000 years (cf. Cooper and Schmitzke 1984; Malaspina et al. 1990; Dorit, Akashi, and Gilbert 1995).

The analytic methods employed here are relevant to recent suggestions that microsatellite loci can be used to identify genomic regions having undergone a recent selective sweep (Slatkin 1995). Although the details will depend on the estimated mutation rate and population size, it is instructive to consider the case of human Y-specific microsatellites as an example. Because of the magnitude of \( \bar{V}_r \), it is very difficult to reject mutation-drift equilibrium (rejection might indicate, for example, that a selective sweep has occurred) unless the number of loci is very large or the observed variation is far below \( \bar{V}_r \). As noted, with \( V_r = 2.0 \) (80% of \( \bar{V}_r \)), 110 loci are required to reject mutation-drift equilibrium. As \( V_r/\bar{V}_r \) decreases, however, the number of loci required decreases rapidly. For \( V_r = 1.25 \) (i.e., 50% of \( \bar{V}_r \)), 13
loci are required, while for \( V_r = 0.625 \), only 4 loci are required. This result has important implications for studies using microsatellites to infer patterns of selection. Because it will rarely (if ever) be possible to find more than a few tightly linked microsatellites, in searching for evidence of selective sweeps it is preferable to focus on loci with substantially reduced variances and ignore those with marginally reduced variances. Of course, it will be necessary to determine whether such loci have atypically low mutation rates, perhaps by evaluation of genetic distances (Goldstein et al. 1995b) or by evaluation of pedigrees (Weber and Wong 1993).

Discussion

We have developed a statistical framework for testing departures from mutation-drift equilibrium at microsatellite loci and a heuristic method for estimating boundaries on the most recent possible elimination of variation. We used these methods to estimate the most recent possible time at which there was no variation among human Y chromosomes (due to a bottleneck or sweep). Because the human population may be at or near mutation-drift equilibrium for the Y-specific microsatellites, however, the method is not suitable for calculating an expected date of an assumed elimination of variation. As it stands, therefore, the variation maintained at Y-specific microsatellites does not provide much information about human origins. If many more Y-specific loci were characterized it is possible that mutation-drift equilibrium could be rejected, at which point it would be appropriate to calculate the expected date of a bottleneck or sweep.

Lack of variation among human Y chromosomes (Cooper and Schmidtke 1984; Malaspina et al. 1990; Dorit, Akashi, and Gilbert 1995) has prohibited statistical reliability in inferring bottlenecks or sweeps. Using faster evolving markers, we have demonstrated that Y microsatellites are not unusually low in variation compared with autosomal microsatellites and that if a sweep or an extreme bottleneck occurred, it is not likely to have done so more recently than about 74,000 years ago.

We have also shown that it is very difficult to formally reject mutation-drift equilibrium unless many loci are available or the observed variance \( V_r \) is very low. Since it will not usually be possible to find more than a few tightly linked microsatellites, these considerations suggest that detection of sweeps that are not very recent will be difficult.

Caveats

We have ignored the obvious structure among human populations. The more structure, the higher will be the observed variation when all individuals are treated as drawn from a single population. If structure has a large impact on the variation, then the effective population size will be overestimated, but the ability to test for concordance among different genomic regions is retained.

The mutation rate for microsatellites is far from established, and may differ between autosomal and Y-specific microsatellites. The number of parent-offspring transfers available for direct study is continually increasing, however, and distances between species (e.g., primates) can also be used to estimate the mutation rate if the separation time for the species is known (Goldstein et al. 1995b). When this time is unknown, the relative mutation rates of different loci may still be estimated. To make more accurate inferences about the history of Y chromosomes, it will be especially important to compare the mutation rates of Y-specific and autosomal microsatellites.

Another important complication is our assumption of strictly stepwise mutations. While pedigree analyses, and other direct studies, clearly demonstrate the predominance of stepwise mutations, they do not preclude rare mutations of larger effect, which are suggested by some population data (Di Rienzo et al. 1994; Goldstein et al. 1995b). We tested the sensitivity of our calculations to such mutations by conducting simulations identical to those described above, except that 99% of all mutations were strictly stepwise, with the size of the remainder uniformly distributed on the interval [1, 9]. In this case the upper bound of the 95% confidence interval will reach the value 2.0 in only 1,000 generations, or about 27,000 years. In order to maximize the information provided by microsatellites it is clearly necessary to gain a better understanding of the full distribution of mutational effects.

Acknowledgments

Supported in part by NIH postdoctoral fellowship GM 16898 to D.B.G., NIH grant GM28428 to L.C.S. and M.W.F., and a grant from the John D. and Catherine T. MacArthur foundation. A. Clark and D. Pollock provided helpful comments on earlier drafts, and D.B.G. thanks A. Templeton for a conversation in which some of the ideas presented here took shape.

APPENDIX

To derive the confidence intervals we will need to determine the moments of a log-transformed variable in terms of the moments of the untransformed variable. This is done using the delta method (see, e.g., Stuart and Ord 1987, p. 324). Suppose that the random variable \( X \) has expectation \( m \) and variance \( v \) which is a function of \( m \). The transformed random variable \( Y = \ln X \) is frequently much closer to Gaussian than \( X \) and is commonly used in data analysis. In our case, we use results from Zhivotovsky and Feldman (1995) with the average variance (over loci) at time \( t \), \( V_r(t) \), taking the role of \( v \). The suggested 95% confidence interval is then \( (V_1, V_2) \) with

\[
V_1 = \exp \left\{ \ln \bar{V}_r(t) - \frac{V_r(t)}{2[\bar{V}_r(t)]^2L} - \frac{2}{\bar{V}_r(t)} \left( \frac{V_r(t)}{L} \right)^{1/2} \right\},
\]

\[
V_2 = \exp \left\{ \ln \bar{V}_r(t) - \frac{V_r(t)}{2[\bar{V}_r(t)]^2L} - \frac{2}{\bar{V}_r(t)} \left( \frac{V_r(t)}{L} \right)^{1/2} \right\}.
\]
As $t \to \infty$, $\hat{V}_r(t)$ approaches $\hat{V}_r = (N - 1)\mu$ and $V_g(t)$ approaches $\frac{4}{3} \hat{V}_r^2 + \frac{\hat{V}_r}{6}$, the values given in Zhivotovsky and Feldman (1995) at mutation-drift equilibrium.

**LITERATURE CITED**


Richard G. Harrison, reviewing editor

Accepted July 24, 1996