Evolution, 49(4), 1995, pp. 608–615

GENETIC EVIDENCE FOR A PLEISTOCENE POPULATION EXPLOSION

ALAN R. ROGERS
Department of Anthropology, University of Utah, Salt Lake City, Utah 84112
E-mail: rogers@anthro.utah.edu

Abstract.—Expansions of population size leave characteristic signatures in mitochondrial "mismatch distributions." Consequently, these distributions can inform us about the history of changes in population size. Here, I study a simple model of population history that assumes that, t generations before the present, a population grows (or shrinks) suddenly from female size N_0 to female size N_1. Although this model is simple, it often provides an accurate description of data generated by complex population histories. I develop statistical methods that estimate $\theta_0 = 2uN_0$, $\theta_1 = 2uN_1$, and $\tau = 2ut$ (where u is the mutation rate), and place a confidence region around these estimates. These estimators are well behaved, and insensitive to simplifying assumptions. Finally, I apply these methods to published mitochondrial data, and infer that a major expansion of the human population occurred during the late Pleistocene.

Key words.—Demography, human evolution, molecular anthropology, population genetics, Upper Paleolithic.

Received November 8, 1993. Accepted May 19, 1994.

It is remarkable that genetic data can inform us about demographic changes that occurred 100,000 yr ago. The possibility exists because genetic differences between individuals measure the genealogical distance between them, and genealogical distances tend to increase with population size. Two random individuals are more likely to be siblings (connected by a short genealogy) in a population of 10 than in one of 10 million. Consequently, a population's history is written in its genes.

The question is, How can this record best be deciphered? Geneticists have been relating various genetic statistics to population size for many years (Wright 1931), but the classical methods do not make adequate use of modern molecular data. In principle, the most powerful methods are those that base inference on the lengths of branches in a phylogenetic tree (Felsenstein 1992). Unfortunately, these methods pose challenging numerical problems, and have not yet been implemented for the case of a nonstationary population.

Instead of sorting through phylogenetic trees, one can also work with the relative frequencies of pairs of individuals in a sample who differ by i nucleotide (or restriction) sites, where $i = 0, 1, \ldots$ (Slatkin and Hudson 1991; Rogers and Harpending 1992). The frequency distribution of such differences has been called the "distribution of pairwise genetic differences" and the "mismatch distribution" (Hartl and Clark 1989; Harpending et al. 1993). For brevity, I adopt the latter term here. Analysis of the mismatch distribution may not be optimal, but it is fast and will be shown to have satisfactory statistical properties.

The sections that follow (1) introduce the model of population history that underlies the analysis, (2) develop methods of point and of interval estimation, (3) investigate their behavior with simulated data, (4) defend these results against various criticisms, and (5) discuss their implications for the debate about modern human origins.

THE MODEL OF SUDDEN EXPANSION

Analysis is based on a simplified model of population history that Harpending and I (Rogers and Harpending 1992) have called the model of "sudden expansion": An initial population of female size N_0 is at equilibrium between the effects of mutation and genetic drift, then grows (or shrinks) quickly to a new female size, N_1, and is observed t generations later. Only the female population sizes matter because the mitochondria of males are not transmitted to offspring. Strictly speaking, N_0 and N_1 refer not to the actual numbers of females but to their "effective number," defined as the reciprocal of the probability that two random individuals have the same mother.

This model of demographic history is unrealistically simple. Its value results from three features of the dynamics of the mismatch distribution (Rogers and Harpending 1992): First, after a population decreases to a small size, convergence to the new equilibrium is rapid. This implies that "bottlenecks," or temporary reductions in population size, amount to growth from an equilibrium population unless the bottleneck is very brief. Thus, it is often reasonable to assume that the pre-expansion population was at equilibrium. Second, after a population grows large, convergence to the new equilibrium is exceedingly slow. Third, an initial expansion will obscure the effects of later expansions (and even those of minor bottlenecks) for a very long time. Throughout this extended period, the signature of the original expansion dominates the mismatch distribution. Consequently, the model of sudden expansion provides a good fit to data even when the true story is one of continued exponential growth (Rogers and Harpending 1992).

This model reduces population history to three parameters: N_0, N_1, and t. Unfortunately, the effect of each is confounded with u, the sum of per-nucleotide mutation rates in the region of DNA under study. Thus, the mismatch distribution can inform us only about three composite parameters, $\theta_0 = 2uN_0$, $\theta_1 = 2uN_1$, and $\tau = 2ut$. These parameters measure female population size in units of $1/2u$ individuals and time in units of $1/2u$ generations.

ESTIMATION BY THE METHOD OF MOMENTS

The estimators proposed here are obtained by fitting the empirical mean and variance to their theoretical counterparts. This procedure, called the method of moments, is widely used and usually successful. However, this is no basis for confidence here. Method of moments estimators are ordinarily

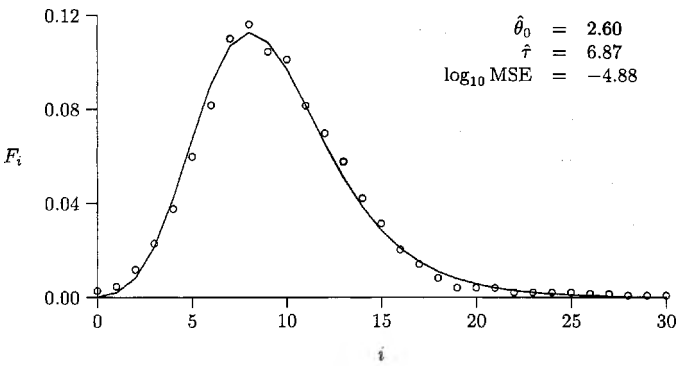

FIG. 1. Fit of the model to data. F_i is the relative frequency of pairs of individuals that differ by i restriction sites. The circles show the empirical distribution of Cann et al. (1987), based on their figure 1. The solid line is the theoretical distribution fit using equations (2) and (3).

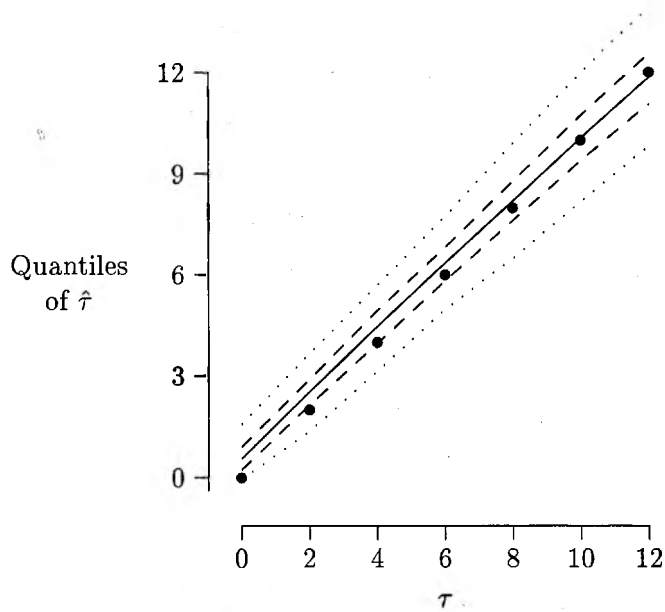

FIG. 2. Quantiles of $\hat{\tau}$. One-thousand data sets were simulated at each of several values of τ , and each was used to estimate the model's parameters. The bold dots indicate points at which $\hat{\tau} = \tau$. The solid line is the median, the dashed lines enclose the central 50% of the distribution, and the dotted lines the central 95%. Each simulated data set was generated using the coalescent algorithm with $\theta_0 = 1$, $\theta_1 = 500$, and $N = 147$.

applied to data with statistically independent observations. The observations that contribute to an empirical mismatch distribution, however, are far from independent: each pair of individuals is correlated to a greater or lesser degree with every other pair. Thus, later sections will use computer simulations to show that the statistics introduced here are in fact useful as estimators. In the meantime, the argument of this present section is intended to motivate these estimators, not to justify them.

The expectation of the rth power of a random variable is called its rth moment about zero. The method of moments estimates parameters by equating observed with theoretical moments, and solving the resulting equations. With three parameters to estimate, three equations are required. Thus, the straightforward approach would equate the first three theoretical moments with their empirical analogues. However, this approach requires numerical methods that often fail to converge. Better estimators are obtained from a reduced model obtained by letting $\theta_1 \to \infty$. This is a useful simplification because the case in which $\theta_1 \to \infty$ closely approximates that in which θ_1 is merely large (Rogers and Harpending 1992). It also applies exactly to pairs of individuals drawn from separate populations that have not exchanged migrants for τ generations.

Let $G_i(\tau)$ denote the probability that two such individuals differ by i nucleotide (or restriction) sites. Letting $\theta_1 \to \infty$ in Rogers and Harpending's (1992) equation (4) gives

$$G_i(\tau) = \frac{\theta_0^i}{(1 + \theta_0)^{i-1}} \sum_{j=0}^{i} \left(\frac{1 + \theta_0}{\theta_0} \right)^j \frac{\tau^j e^{-\tau}}{j!}. \quad (1)$$

The moment generating function, obtained from this expression or from Li's (1977, eq. 2) probability generating function, is

$$\phi(z, \tau) = \frac{ze^{-\tau(1-e^z)}}{1 + z\theta_0(1 - ze^z)}.$$

Standard methods (Kendall and Stuart 1977, eq. 3.18) provide the first two moments about zero:

$$\mu_1 = \theta_0 + \tau;$$

$$\mu_2 = \theta_0^2 + \theta_0 + \tau + \mu_1^2.$$

Setting the observed mean, m, and variance, v, equal to $m = \mu_1$ and $v = \mu_2 - \mu_1^2$ leads to two statistics,

$$\hat{\theta}_0 = \sqrt{v - m}, \quad (2)$$

$$\hat{\tau} = m - \hat{\theta}_0, \quad (3)$$

which I propose to interpret as estimators. In practice, I set $\hat{\theta}_0 = 0$ if $v < m$, and $\hat{\tau} = 0$ if $m < \hat{\theta}_0$.

To illustrate the method, I use the mitochondrial mismatch distribution from the world human sample of Cann et al. (1987, fig. 1). Figure 1 shows that the method provides an excellent description of the data. The estimates presented there are similar to the least squares estimates of Rogers and Harpending (1992). The fit of the theoretical curve should not, however, be interpreted as support for my proposal that $\hat{\theta}_0$ and $\hat{\tau}$ be interpreted as estimators—many other two-parameter functions would fit as well. The case in favor of these statistics is made in the section that follows.

Statistical Properties of Point Estimates

To determine the statistical properties of $\hat{\theta}_0$ and $\hat{\tau}$, I used the coalescent algorithm (Hudson 1990) to generate 1000 simulated data sets at each of a wide variety of parameter values. In order to allow for changes in population size, I used a modified version of the coalescent algorithm, which is described elsewhere (Rogers in press). I estimated θ_0 and τ from each simulated data set, thus obtaining an estimate of the sampling distribution of the estimators for each set of parameter values.

Figure 2 shows how the sampling distribution of $\hat{\tau}$

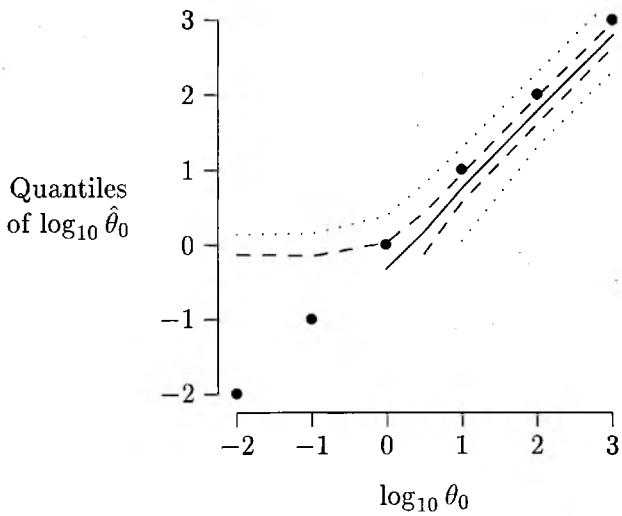

FIG. 3. Quantiles of θ_0. One-thousand data sets were simulated at each of several values of θ_0, and each was used to estimate the model's three parameters. In each run, $\theta_1 = 1000$, $\tau = 7$, and $N = 147$. The lines and bold dots are interpreted as in figure 2.

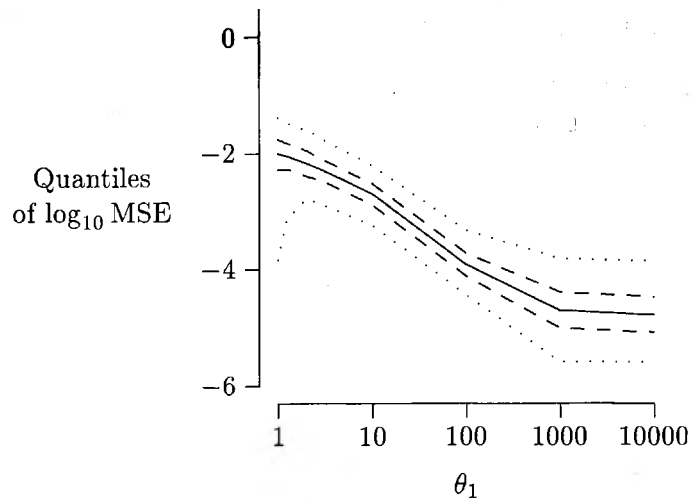

FIG. 4. Quantiles of $\log_{10}$MSE. Quartiles were estimated from 1000 data sets simulated at each of several values of θ_1. In each run, $\theta_0 = 1$, $\tau = 7$, and $N = 147$. The lines are interpreted as in figure 2.

changes in response to variation in the underlying parameter τ. If $\hat{\tau}$ is in fact an estimator of τ, we would expect the median of $\hat{\tau}$ (shown as a solid line in the figure) to increase in response to increases in τ. This is indeed the case. An ideal estimator should also have a relatively narrow distribution at each value of τ. The dashed and dotted lines show that $\hat{\tau}$ also satisfies this test. The dashed lines enclose the central 50% of the distribution, and the dotted lines the central 95%. Both sets of lines enclose a relatively narrow interval about the median. In all of these respects, $\hat{\tau}$ behaves as an estimator of τ.

Figure 3 performs a similar analysis on $\hat{\theta}_0$, and shows it to perform well as an estimator when $\theta_0 > 1$. The distribution is tightly centered about the bold dots, showing that $\hat{\theta}$ is rich in information and nearly unbiased when $\hat{\theta}_0 > 1$. But when $\theta_0 < 1$, the upper quantiles of $\log_{10} \hat{\theta}_0$ are horizontal, while the median and lower quantiles of $\hat{\theta}_0$ equal zero. Thus, an estimate of $\hat{\theta}_0 \approx 1$ is equally consistent with the hypotheses that $\theta_0 = 1$ and that $\theta_0 = 0$. Although $\hat{\theta}_0$ will always allow us to place an upper bound on θ_0, it can provide no lower bound unless $\hat{\theta}_0$ is much greater than one. This is no serious problem; it means only that when the estimate is near unity, the confidence interval will reach all the way to zero.

But what about θ_1? We have no estimate of this parameter, but Harpending has shown that empirical distributions tend to be "smooth" when θ_1 is large and θ_0 is much smaller than τ ; otherwise, they tend to be "rough" (Harpending et al. 1993; Harpending 1994). Thus, a measure of roughness may provide information about θ_1. Harpending et al. measure roughness by the sum of squared differences between successive entries of the empirical mismatch distribution. My own simulations suggest that this statistic is less informative than another measure of roughness, the mean squared error (MSE) between the observed and fitted mismatch distributions. Rather than calculating the fit using equation 1, which assumes that $\theta_1 \to \infty$, I use the full three-parameter equation (Rogers and Harpending 1992, eq. 4), with $\theta_1 = \hat{F}_0^{-1} - 1$, where $\hat{F}_0$ is the relative frequency in the data of pairs of individuals that differ by zero sites. This approach was suggested by Rogers and Harpending, and usually provides a better fit when $\hat{F}_0$ is far from zero. The quantiles of the sampling distribution of $\log_{10}$MSE are plotted against θ_1 in figure 4, and verify that this statistic contains information about θ_1.

This section has shown that the statistics presented above contain information about the parameters they are intended to estimate. I turn next to the task of constructing a confidence region.

CONFIDENCE REGIONS

In this section, I ask which parameter values can be rejected by the data, and which cannot. The set of parameter values that cannot be rejected will be interpreted as a confidence region. This procedure is justified by the very definition of a confidence region. A 95% confidence region is a set of parameter values constructed by any procedure that guarantees the following property (Kendall and Stuart 1979, p. 110): If, each time we construct a 95% confidence region, we assert that it includes the true parameter value, we will in the long run be correct 95% of the time (and incorrect 5% of the time). One way to construct such a region is to define some statistical test whose outcome depends only on the data and the parameters of interest. The set of parameter values that cannot be rejected at significance level α will constitute a $100 \times (1 - \alpha)$% confidence region.

There are innumerable ways to construct such a test, and each will lead to a valid confidence region. However, some are more useful than others. To make my confidence intervals small, I have tried to construct a test whose region of acceptance A is small, subject to the constraint that there be a fixed probability $1 - \alpha$ that an observation will fall within it. This requires that A be chosen so that all points along its boundary have equal probability density. In other words, the

region of acceptance should be defined by one of the contour lines of the density function. When the distribution is multivariate normal, points of equal density also have equal values of the Mahalanobis distance,

$$\mathbf{D(X)} \equiv (\mathbf{X} - \mathbf{M})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{X} - \mathbf{M})$$

where **X** is a vector of observations; **M**; the corresponding vector of mean values; **C**, the covariance matrix; and the superscript **T** indicates the matrix transpose. This suggests a procedure for constructing small confidence intervals from normal data: For each set of parameter values, the first step would estimate **M** and **C** from simulated data. The second would calculate D both from the real data and also from each simulated data set. The parameter values could be rejected at the 5% level if less than 5% of the simulated distances were as large as the observed distance.

Unfortunately, this test generates confidence intervals that are disappointingly large, apparently because the probability distribution is far from multivariate normal. Graphical analysis indicates that $\log_{10}\hat{\theta}_0$ and $\hat{\tau}$ are approximately bivariate normal, and that the marginal distribution of $\log_{10}$MSE is also approximately normal. But the distribution is far from normal when the three variables are considered together. Therefore, I use a modified procedure that exploits the bivariate normality of $\log \hat{\theta}_0$and $\hat{\tau}$ but does not assume full multivariate normality. The modified test is performed as follows:

1. Use 1000 simulated data sets to estimate **M** and **C** as above, but include only two variables, $\log_{10}\hat{\theta}_0$ and $\hat{\tau}$.
2. Define the Mahalanobis distance D using only these two variables. In this calculation, I use the algorithm described by Dongarra et al. (1979, pp. 8.8–8.9).
3. Count the number n of simulated data sets for which the simulated D is at least as large as the observed D, and the simulated MSE is at least as small as the observed MSE, and reject if $n/1000 \leq 0.05$.

This test uses the approximately normal distribution of $\log_{10}\hat{\theta}_0$ and $\hat{\tau}$ to define a relatively small region of acceptance, and then reduces that region still further by imposing an additional condition involving the MSE. As figure 4 shows, the MSE tends to be smallest in data from populations that have grown. Thus, the test is more appropriate for expanded than for equilbrium populations, producing a narrower confidence region in the former case than in the latter.

This choice is sensible because the behavior of the method with data from equilibrium populations is not very important. Equilibrium populations produce mismatch distributions that are extremely ragged, and not at all like the corresponding theoretical curves (Slatkin and Hudson 1991; Rogers and Harpending 1992). The estimators developed here should work poorly with such data anyway. Mismatch distributions from equilibrium populations can be recognized by their roughness (Harpending et al. 1993; Harpending 1994), and I don't expect the present methods to be applied to such data anyway. Thus, it makes sense to sacrifice precision with equilibrium populations in order to gain precision with expanded populations.

To evaluate this method for producing confidence regions,

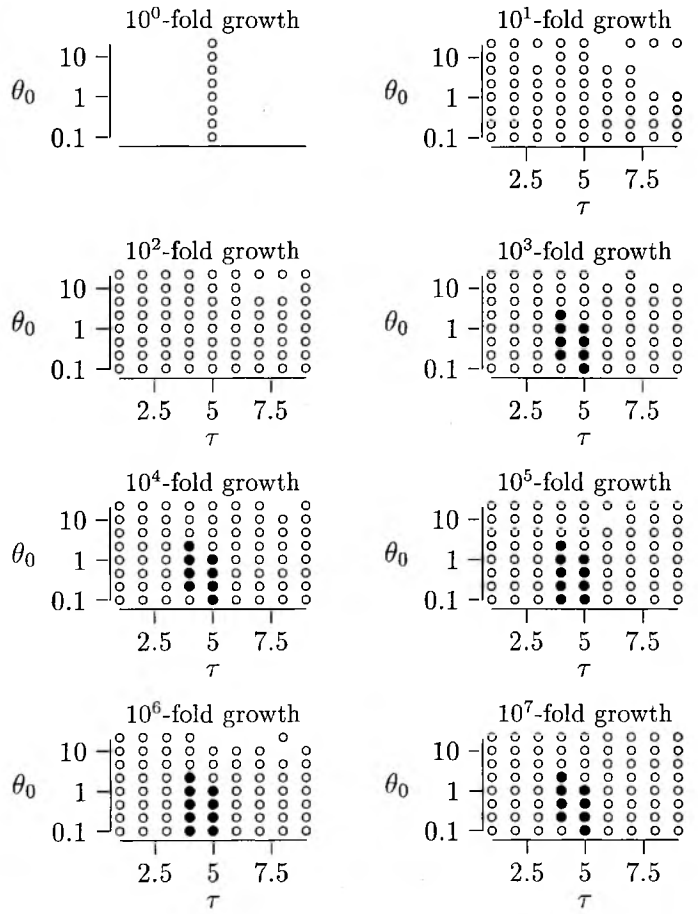

FIG. 5. Ninety-five percent confidence region for a simulated population with $\tau = 4$. A data set of size $N = 147$ was simulated assuming that $\theta_0 = 1$, $\theta_1 = 500$, and $\tau = 4$, and a confidence region was then generated as described in the text. Open circles represent points outside the 95% confidence region; filled circles represent points within.

I simulated data under several different assumptions and constructed confidence regions for each. The first of these, shown in figure 5, is based on data for which $\theta_0 = 1$, $\theta_1 = 500$, and $\tau = 4$. Each panel there considers a different hypothesis about the magnitude of the population expansion. At each point in the "no-growth" panel, $\theta_1 = \theta_0$, which leaves τ undefined. Thus, there is only one parameter to vary, θ_0. The eight open circles indicate that eight different values of θ_0 were considered and rejected. Thus, the method correctly rejects the hypothesis of no growth. The "tenfold growth" panel in figure 5 entertains the hypothesis that the population increased in size by a factor of ten, so that $\theta_1 = 10\theta_0$. Here, there are two free parameters, so a rectangular matrix of parameter values was considered. All were rejected. The hypothesis of 100-fold growth was also (correctly) rejected. In the "10^3-fold growth" panel, we see for the first time a new symbol, the filled circle, which indicates a set of parameter values that was not rejected. The 95% confidence region is defined by the filled circles in the various panels. Note that the confidence region is narrow, and includes the true parameter values.

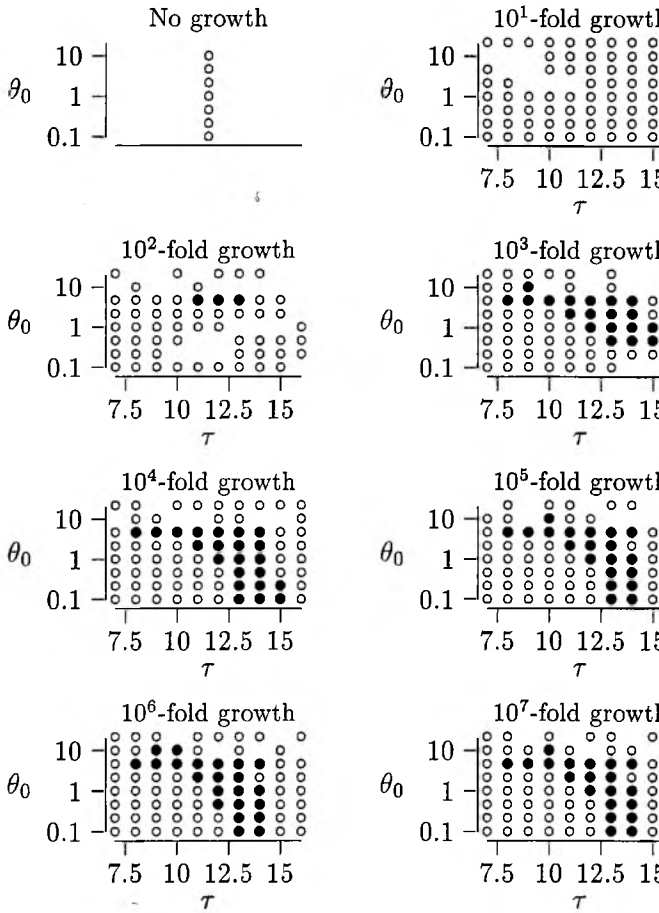

FIG. 6. Ninety-five percent confidence region for a simulated population with $\tau = 12$. A data set of size $N = 147$ was simulated assuming that $\theta_0 = 1$, $\theta_1 = 500$, and $\tau = 12$, and a confidence region was then generated as described in the text. Open circles represent points outside the 95% confidence region; filled circles represent points within.

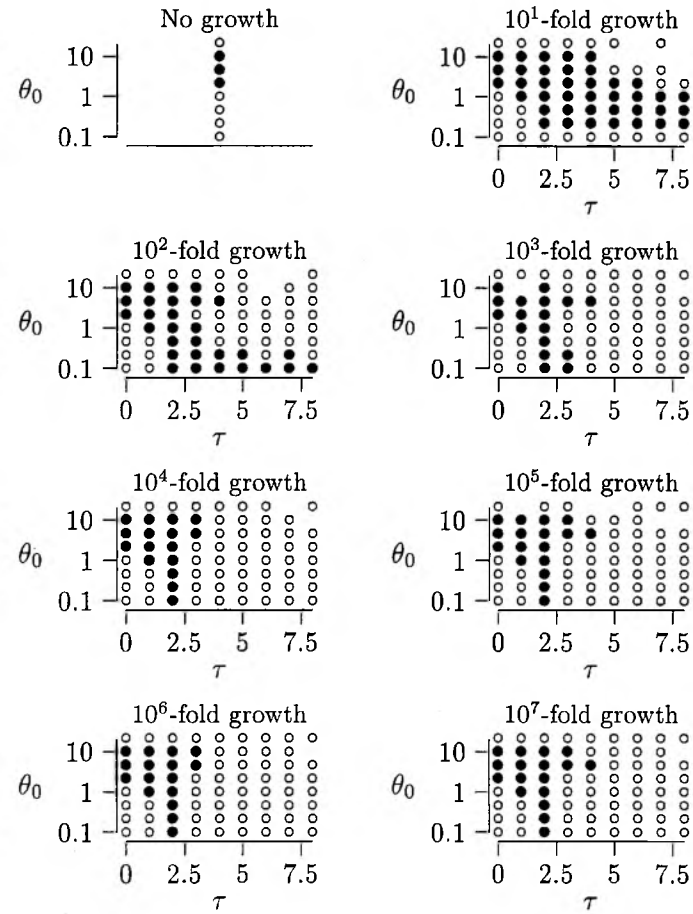

FIG. 7. Ninety-five percent confidence region from a simulated equilibrium population. A data set of size $N = 147$ was simulated assuming that the population was at equilibrium with $\theta = 3.1623$, and a confidence region was then estimated as described in the text. Open circles represent points outside the 95% confidence region; filled circles represent points within.

The confidence interval in figure 6 is based on data for which the true value of τ is 12. Once again, the confidence interval is small and includes the true parameter values.

Figure 7 shows a confidence region for a case in which I expect the method to work poorly—that of an equilibrium population. Note that there are closed circles in each panel, indicating that no value of growth (θ_1/θ_0) is excluded. Neither does the confidence region exclude any value of τ. Thus, it informs us neither about the amount of growth that has occurred, nor about the time of this growth. This poor performance agrees with my low expectations for equilibrium data. I was surprised, however, by the relatively narrow (and accurate) bound on θ_0. The method provides some useful information even with worst-case data.

In summary, it appears that the present method produces narrow confidence regions when applied to data from populations that have expanded. It does not misinform us even when applied to worst-case data.

A Confidence Region for Human Data

Figure 8 shows a confidence region calculated from the Cann-Stoneking-Wilson data shown in figure 1. The first three panels, corresponding to no growth, 10-fold growth, and 100-fold growth, contain only the open circles that indicate rejected hypotheses. Thus, the confidence region indicates that the human population expanded by more than 100-fold. It places no upper limit on the magnitude of growth, but does place rather narrow limits on the other parameters: $\theta_0 < 10$, and $4 < \tau < 9$.

Sensitivity to Simplifying Assumptions

Before discussing what this confidence region implies, we should consider the possibility that it is unreliable. There are several causes for concern.

The Model of Sudden Expansion Is Not an Accurate Description of Population History.—The theoretical mismatch distribution is remarkably insensitive to violations of the model of sudden expansion. This was demonstrated by Rogers and Harpending (1992), whose results were summarized in the second section of the present paper. When an initial expansion is followed by later expansions or minor bottlenecks of population size, the theoretical mismatch distribution is affected only slightly. The empirical mismatch dis-

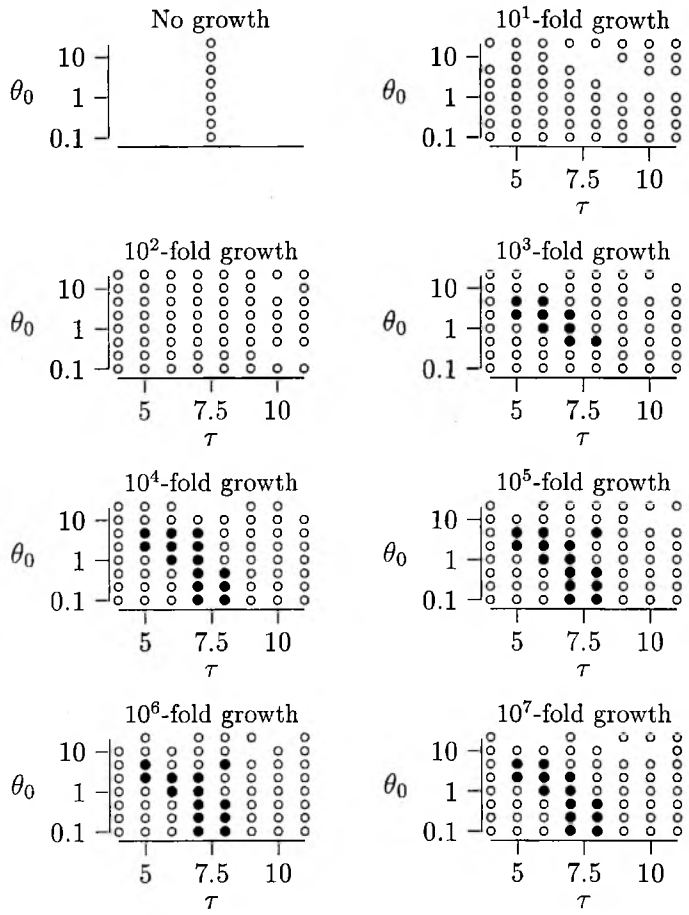

FIG. 8. Ninety-five percent confidence region for the CSW data. Large filled circles indicate points within the 95% confidence region, and open circles indicate points outside of the confidence region. 10^x-fold growth means that $\theta_1/\theta_0 = 10^x$. Data are from Cann et al. (1987).

tribution is also robust when the initial population is small, and is not subdivided (Rogers in press).

Mutation Rates Vary Across Nucleotide Sites.—Mutation is assumed to follow Kimura's (1971) model of "infinite sites," which implies that no nucleotide site mutates more than once. However, several of the sites studied have clearly mutated repeatedly (Kocher and Wilson 1991). This suggests that some sites may mutate faster than others, a possible problem since rate variation can generate signatures that mimic those produced by population growth (R. Lundstrom MS). However, I have shown elsewhere (Rogers 1992) that this probably introduces only a negligible error of about 3% in the expected number of site differences between pairs of individuals in human data. This suggests that little error is introduced into the theoretical curves, and possibly that the empirical curves will be similarly unaffected. Further work is needed on this point.

Real Populations Are Subdivided and Do Not Mate at Random.—The statistical methods assume random mating. Yet I apply them to the human population, which, far from mating at random, is divided into a large number of partially isolated subdivisions. This application can be defended only to the extent that the mismatch distribution is insensitive to subdivision. I treat this problem in detail elsewhere (Rogers in press) and deal here only with the effect of one form of subdivision on the theoretical mismatch distribution.

TABLE 1. Theoretical mismatch distributions in a subdivided and a randomly mating population. The subdivided and the randomly mating populations both began at time 0 as randomly mating populations at equilibrium with $\theta_0 = 1$, which then grew suddenly by a factor of 200. Both are observed at time $\tau = 8$. At time 0 the subdivided population split into two isolated subpopulations. Column 2 contains the distribution for pairs within subdivisions, column 3 that for pairs from different subdivisions, column 4 that for pairs drawn at random from the entire subdivided population, and column 5 that for pairs from the randomly mating population.

	Subdivided			Random
i	Within	Btw	Total	mating
0	0.0101	0.0002	0.0051	0.0051
1	0.0111	0.0014	0.0063	0.0063
2	0.0152	0.0061	0.0107	0.0106
3	0.0252	0.0174	0.0214	0.0213
4	0.0430	0.0373	0.0402	0.0402
5	0.0672	0.0645	0.0659	0.0658
6	0.0926	0.0933	0.0930	0.0930
7	0.1126	0.1164	0.1145	0.1145
8	0.1220	0.1280	0.1250	0.1250
9	0.1190	0.1260	0.1225	0.1225
10	0.1057	0.1127	0.1091	0.1092
11	0.0864	0.0924	0.0893	0.0894
12	0.0655	0.0703	0.0678	0.0679
13	0.0464	0.0499	0.0482	0.0482
14	0.0310	0.0334	0.0322	0.0322
15	0.0197	0.0212	0.0204	0.0205
16	0.0119	0.0129	0.0124	0.0124
17	0.0069	0.0075	0.0072	0.0072
18	0.0039	0.0042	0.0041	0.0041
19	0.0021	0.0023	0.0022	0.0022
20	0.0011	0.0012	0.0012	0.0012

Consider a population that initially mates at random and is at equilibrium with size θ_0, but then splits into K completely isolated populations of size θ_1/K which are observed τ units of mutational time later. Pairs of individuals drawn at random from the total population differ by i sites with probability

$$H_i(\tau) = F_i(\tau)/K + (1 - 1/K)G_i(\tau), \qquad (4)$$

where $F_i(\tau)$ is the mismatch distribution for pairs within a single randomly mating population of size θ_1/K (Rogers and Harpending 1992, eq 4), and $G_i(\tau)$ the mismatch distribution for pairs from separate, completely isolated populations (eq. 1).

Note that $H_i = F_i$ when $K = 1$, and that $H_i \to G_i$ as $K \to \infty$. H_i falls between these limits when K takes intermediate values. But we already know that $G_i \approx F_i$ when θ_1/K is large and θ_0 small. This is illustrated below and is also illustrated by the close fit of the two-parameter model to the data in figure 1. If $G_i \approx F_i$, then equation (4) implies that $H_i \approx F_i$ whatever the value of K. Thus, the theory for a randomly mating population should hold approximately even when subdivisions are completely isolated. When subdivisions are incompletely isolated, the random mating approximation should be even better.

Table 1 illustrates this result for the case of a population with two completely isolated subdivisions. The subdivided

population is compared with one of equal size that mates at random. The two populations have identical demographic histories except that one has been subdivided for τ units of mutational time. The table shows that subdivision has a remarkably small effect. Indeed, the effect is entirely invisible when these distributions are displayed graphically. The similarity of these distributions is even more remarkable in view of my extreme assumption that there was no gene flow at all between subdivisions. With gene flow, the two distributions would be even more similar.

This shows that population structure has a negligible effect on the theoretical distribution in one important case: that in which θ_0 is small, θ_1/K is large, and the time of population growth coincides with the time of subdivision. Elsewhere (Rogers in press), I show that. in this case, the effect on the empirical distribution is similarly small. The effect is not so small when the initial population is subdivided: subdivision makes the upper bound on θ_0 even smaller.

The Sample Is Less Than Ideal.—My estimates are based on the sample of Cann et al. (1987), which has been criticized because its "African" component actually consists of American blacks (Spuhler 1988; Krüger and Vogel 1989). Yet, similar results are obtained from many other samples (Harpending et al. 1993; Sherry et al. 1994; Harpending 1994). Thus, the main conclusions of this analysis cannot be attributed to problems with this particular sample.

Pairs of Individuals in the Sample Are Not Independent.—Ideally, the estimators developed here should be applied to an empirical distribution based on statistically independent pairs of individuals. Unfortunately, this is impossible. The pairs of individuals studied here are correlated both because of genealogical relationships and because each individual participates in many different pairings. Consequently, there is no reason a priori to expect these estimators to perform well at all. Yet the simulations show that they do. The between-pair correlations are present not only in the CSW data, but also in the simulated data. Figures 2–4 show that the univariate estimators are useful, correlations notwithstanding, and figures 5–7 show that the confidence region is also useful.

In summary, the analysis makes several unrealistic simplifying assumptions, but for each there is reason to suppose that the violated assumption probably has no large effect on the estimates.

Modern Human Origins

In this final section, I consider what the results obtained above imply about the origin of modern humans. Figure 8 indicates that the lower bound on the confidence interval for τ is between 4 and 5, whereas the upper bound is between 8 and 9. On a conservative interpretation, we can conclude that the ancestors of the present human population expanded dramatically between 4 and 9 units of mutational time ago. The analysis places a lower bound, but no upper bound on the magnitude of the expansion: the increase must have been more than 100-fold. It should not be inferred that this increase occurred as the model assumes—all at once. Data like those observed could also have been produced by other trajectories of growth, including continued exponential growth beginning at around $\tau = 6$ (Slatkin and Hudson 1991; Rogers and Harpending 1992). The results imply that substantial population growth occurred in the neighborhood of $\tau = 6$, but say nothing about the later history of population growth.

To re-express τ in years, we must divide by $2u$ (twice the mutation rate) and multiply by the length of a generation, say, 25 yr. Unfortunately, the mutation rate is not known with great accuracy. The rate of human mitochondrial nucleotide divergence has been variously estimated at 2% and 4% per million years (Cann et al. 1987), but the confidence intervals around these estimates are unknown. The two estimates place u at 7.5×10^{-4} and 1.5×10^{-3}, respectively (Rogers and Harpending 1992). If we knew the larger estimate of u to be correct, then each unit of the mutational time scale would correspond to 8333 yr, and the confidence interval for τ would correspond to 33,000–75,000 yr B.P. The smaller estimate of u doubles these values, giving 66,000–150,000 B.P. Neither of these is a true confidence interval, because neither takes proper account of the sampling distribution of u. Calculation of a true confidence interval for t must await better information about the sampling distribution of u.

Similar comments apply to the estimates of θ_0. The confidence region says that $\theta_0 < 10$. With the smaller estimate of u, this gives approximately $N_0 < 7000$, in good agreement both with earlier estimates of our long-term effective population size and with earlier estimates of N_0 (reviewed by Rogers and Jorde 1995). The upper bound on N_0 is remarkably small and may be biased downward. If the wave in the empirical distribution resulted from a very brief bottleneck, the pre-expansion population may have been far from equilibrium. This could cause a downward bias in $\hat{\theta}_0$, and may account for the small estimates (Rogers and Harpending 1992). Further simulations are needed to check this conjecture. An opposite bias may be introduced by the assumption of random mating—if the initial population were structured, then the upper bound inferred here would be too high (Rogers in press).

Takahata (1993) used high genetic diversity at the HLA locus to argue that the human population has not passed through any small bottleneck. However, the bottlenecks he is excluding are smaller than the 7000 females in the initial population inferred here. Thus, Takahata's results are consistent with mine (Rogers and Jorde 1995).

The wave in the mismatch distribution might also reflect natural selection rather than a population expansion. If a favorable mutation occurred in a mitochondrion, the carriers of that mutation might increase in number until the new allele was fixed. Thus, our "population" might consist of the female carriers of a new allele. However, Harpending et al. (1993) argued that this interpretation is inconsistent with results from between-population mismatch distributions.

Furthermore, the archeological record provides some support for the view that an expansion did occur. Throughout much of the Pleistocene, stone tools were relatively uniform over vast distances and spans of time. But at around 40,000 B.P. new types of stone tools appear throughout most of the Old World, and thereafter technological change is faster. When skeletal remains are found at these later sites, they are almost invariably those of anatomically modern humans

(Klein 1992). These observations have led some prehistorians to propose the "replacement model" of modern human origins, which holds that modern humans evolved in Africa some 50,000–100,000 yr ago, and then spread throughout the world, replacing earlier peoples as they went (Stringer and Andrews 1988). The expansion that this model proposes occurs at approximately the same time as that implied by the mitochondrial data.

The competing "multiregional model" (Wolpoff 1989) holds that modern humans evolved in a widespread population that inhabited much of Europe, Africa, and Asia. Favorable mutations arising in one place spread throughout the world by gene flow, not by the replacement of whole populations. This hypothesis does not require a population expansion, but neither does it preclude one. It is possible that the origin of modern humans involved some adaptation that allowed our ancestors to inhabit the landscape more densely. If so, a population expansion could have occurred even under the multiregional model. However, it is hard to imagine that this in-place expansion could have been as large as the several-hundred-fold expansion inferred here. To this extent, evidence for a population expansion weighs against the multiregional model. Furthermore, the multiregional model implies that modern humans evolved in a population that spanned several continents, yet the present results imply that this population contained fewer than 7000 females. And this number becomes even smaller if population structure is introduced into the analysis (Rogers in press). It is implausible that a population this small could have spanned three continents and still been connected by gene flow. Thus, the small estimate of N_0 also weighs against the multiregional model.

Finally, the wave the the mismatch distribution could also have been produced by a separation of the human population into several relatively isolated subpopulations. Analysis of between-group mismatch distributions indicates that this may well be the case (Harpending et al. 1993; Gibbons 1993). This interpretation of the data is also inconsistent with the multiregional model since the date of the event inferred here is much later than the original expansion of *Homo erectus* populations throughout Europe and Asia.

Acknowledgments

All calculations were done with Mismatch, a package of computer programs that is available via anonymous ftp from anthro.utah.edu. I thank E. Cashdan, H. Harpending, K. Hawkes, L. Jorde, R. Klein, S. Sherry and A. Templeton for comments. This work was supported in part by grants from the U.S. Department of Health and Human Services (MGN 1 R29 GM39593), and from the National Science Foundation (DBS-9211255 and DBS-9310105).

Literature Cited

Cann, R. L., M. Stoneking, and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Dongarra, J., C. Moler, J. Bunch, and G. Stewart. 1979. LINPACK users' guide. Society for Industrial and Applied Mathematics, Philadelphia.

Felsenstein, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genetical Research 59:139–147.

Gibbons, A. 1993. Pleistocene population explosions. Science 262: 27–28.

Harpending, H. 1994. Signature of ancient population growth in a low resolution mitochondrial DNA mismatch distribution. Human Biology 66:591–600.

Harpending, H. C., S. T. Sherry, A. R. Rogers, and M. Stoneking. 1993. The genetic structure of ancient human populations. Current Anthropology 34:483–496.

Hartl, D. L., and A. G. Clark. 1989. Principles of population genetics, 2d ed. Sinauer, Sunderland, Mass.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. Futuyma and J. Antonovics, eds. Oxford Surveys in Evolutionary Biology, Vol. 7. Oxford University Press, Oxford.

Kendall, M., and A. Stuart. 1977. The advanced theory of statistics. I. Distribution theory, 4th ed. Macmillan, New York

———. 1979. The advanced theory of statistics. II. Inference and relationship, 4th ed. Macmillan, New York,

Kimura, M. 1971. Theoretical foundation of population genetics at the molecular level. Theoretical Population Biology 2:174–208.

Klein, R. G. 1992. The archeology of modern human origins. Evolutionary Anthropology 1:5–14.

Kocher, T., and A. Wilson. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding region. Pp. 391–413 *in* S. Osawa and T. Honjo, eds. Evolution of life: fossils, molecules, and culture. Springer, New York.

Krüger, J., and F. Vogel. 1989. The problem of our common mitochondrial mother. Human Genetics 82:308–312.

Li, W.-H. 1977. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics 85: 331–337.

Rogers, A. R. 1992. Error introduced by the infinite sites model. Molecular Biology and Evolution 9:1181–1184.

———. In press. Population structure and modern human origins. *In* P. J. Donnelly and S. Tavare, eds. Mathematical population genetics. Springer, New York.

Rogers, A. R., and H. C. Harpending. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Molecular Biology and Evolution 9:552–569.

Rogers, A. R., and L. B. Jorde. 1995. Genetic evidence on modern human origins. Human Biology 67:1–36.

Sherry, S., A. R. Rogers, H. C. Harpending, H. Soodyall, T. Jenkins, and M. Stoneking. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. Human Biology 66: 761–776.

Slatkin, M., and R. R. Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562.

Spuhler, J. N. 1988. Evolution of mitochondrial DNA in monkeys, apes, and humans. Yearbook of Physical Anthropology 31:15–48.

Stringer, C. B., and P. Andrews. 1988. Genetic and fossil evidence for the origin of modern humans. Science 239:1263–1268.

Takahata, N. 1993. Allelic genealogy and human evolution. Molecular Biology and Evolution 10:2–22.

Wolpoff, M. H. 1989. Multiregional evolution: the fossil alternative to Eden. Pp. 62–108 *in* P. Mellars and C. Stringer, eds. The human revolution: behavioural and biological perspectives on the origins of modern humans. Princeton University Press, Princeton, N.J.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16: 97–159.

Corresponding Editor: W. Eanes