Mitochondrial Mismatch Analysis is Insensitive to the Mutational Process

Alan R. Rogers, Alexander E. Fraley,† Michael J. Bamshad,† W. Scott Watkins,† and Lynn B. Jorde†*

*Department of Anthropology and †Department of Human Genetics, University of Utah

Mismatch distributions are histograms showing the pattern of nucleotide (or restriction) site differences between pairs of individuals in a sample. They can be used to test hypotheses about the history of population size and subdivision (if selective neutrality is assumed) or about selection (if a constant population size is assumed). Previous work has assumed that mutations never strike the same site twice, an assumption that is called the model of infinite sites. Fortunately, the results are surprisingly robust even when this assumption is violated. We show here that (1) confidence regions inferred using the infinite-sites model differ little from those inferred using a model of finite sites with uniform site-specific mutation rates, and (2) even when site-specific mutation rates follow a gamma distribution, confidence regions are little changed until the gamma shape parameter falls well below its plausible range, to roughly 0.01. In addition, we evaluate and reject the proposition that mismatch waves are produced by pooling data from several subdivisions of a structured population.

Introduction

Mismatch distributions summarize information about genetic differences between pairs of subjects in a sample. They are built by counting the number of nucleotide (or restriction) site differences between each pair of subjects and using a histogram or scatter plot to display the relative frequencies of pairs that differ by zero sites, by one site, and so forth.

Similar distributions are sometimes constructed in which the horizontal axis estimates the number of substitutions per site rather than counting site differences. We use counts rather than estimated substitutions because this simplifies our statistical problem. With intraspecific human data, the pairwise differences are so small that the difference between the two methods is not important.

The open circles in figure 1 show a mismatch distribution calculated from 77 Asian individuals using nucleotide sequences comprising 630 sites within the mitochondrial D-loop. The distribution shows that many pairs of subjects differ at 9 sites and that very few differ at 0 sites or at 15 sites. Formulas are available for the expected value of such distributions under various hypotheses about history (Watterson 1975; Li 1877; Rogers and Harpending 1992). Since these formulas assume that recombination does not occur, the methods described below are appropriate only for genetic systems where recombination is either absent or very rare.

If our population were at mutation-drift equilibrium but had the same mean pairwise difference, the expected mismatch distribution would look like the dashed line shown in figure 1. The two distributions are very different, yet this in itself would not justify rejecting the hypothesis of mutation-drift equilibrium. To do that, one must show not only that the empirical and theoretical distributions differ, but also that such differences are rare in equilibrium populations. Together with several colleagues, we have developed methods to do this using computer simulation and have rejected the equilibrium hypothesis with various data sets (Harpending et al. 1993; Harpending 1994; Sherry et al. 1994; Rogers 1995). It is also possible, by fitting the theoretical curves to data, to estimate parameters describing population history and to place confidence regions about these estimates (Sherry et al. 1994; Rogers 1995). There is no guarantee that these methods are optimal. Indeed, they probably make less efficient use of the data than would analogous methods based on the principle of maximum likelihood. These methods are useful because they are simple and fast, and because they avoid the numerical and statistical difficulties associated with inferring trees.

Mismatch waves such as that in figure1 can be produced either by an episode of population growth or by the sweep to fixation of a favored mitochondrial allele (Rogers and Harpending 1992; Rogers 1995, 1996). Separating these hypotheses requires additional data. For example, one might use data from different loci. If the pattern reflects population growth, then all loci should look the same apart from sampling effects. If the pattern reflects selection, then unlinked loci should look different, since the selective histories of different loci are not identical. Elsewhere, we have argued that human mismatch waves are likely to reflect population history rather than selection (Harpending et al 1993; Rogers 1995). Here, we remain agnostic on this issue. To simplify the exposition, we will continue to speak of a population expansion, but we emphasize that if the selection hypothesis is correct then our "population" refers to the number of carriers of a favored mitochondrial allele.

Mismatch analysis has employed several unrealistic assumptions. First, much of this work assumes that the population mates at random—an assumption that is surely violated in any large widespread population. Several authors have shown that geographic population structure can have a major effect (Harpending et al 1993; Marjoram and Donnelly 1994). Nonetheless, the random-mating results are still accurate

Key words: mismatch distribution, pairwise differences, mitochondrial DNA, mutation, gamma distribution.

Address for correspondence and reprints: Alan R. Rogers, Department of Anthropology, University of Utah, Salt Lake City, Utah 84112. E-mail: rogers@anthro.utah.edu.

Mol. Biol. Evol. 13(7):895–902. 1996

© 1996 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

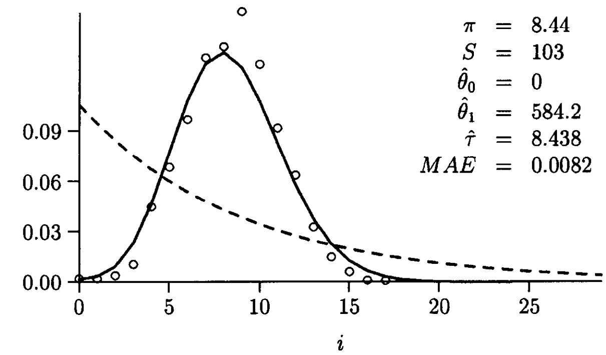

FIG. 1.—Mismatch distribution of Asian sample. On the horizontal axis, i is the number of nucleotide site differences between pairs of individuals. Open circles show the relative frequencies of pairs with i differences. The solid line is the theoretical mismatch distribution fit using equations (2) and (3) of Rogers (1995). π is the mean pairwise difference, S is the number of segregating sites, and MAE is the mean absolute error. The other symbols are defined in the text.

provided that one interprets them as applying to effective rather than actual population size (Rogers 1996).

Previous methods are also unrealistic in assuming that no site can mutate more than once. This has been called the model of "infinite sites," since it would hold if there were an infinity of sites, each with an infinitesimal mutation rate (Kimura 1971). Since this assumption is patently unrealistic, Rogers (1992) studied the error it introduces into the theoretical mismatch distribution. For intraspecific human data, this error turns out to be small. This provides some comfort, but not enough. Perhaps the error introduced into parameter estimates is much larger than that introduced into the theoretical mismatch distribution. Several authors have argued that this is the case (Bertorelle and Slatkin 1995; Aris-Brosou and Excoffier 1996). Thus far, however, only indirect evidence has been offered in support of this position. These authors show that two statistics—the mean pairwise difference and the number of segregating sites—are both sensitive to assumptions about the mutational process. No one has yet shown that the estimates provided by mismatch analysis are equally sensitive. We address that issue here. In addition, we evaluate the hypothesis that waves in the mismatch distribution are produced by pooling data from several parts of a subdivided population (Bertorelle and Slatkin 1995).

Methods

Confidence Regions from Computer Simulations

This paper uses a simple model of population history, which assumes that the population has been constant in size except for an episode of growth (or decline) t generations before the present. N_0 and N_1 denote the effective number of females in the population before and after the burst of growth. Although this model is unrealistically simple, it often provides a fair description of the mismatch distribution even when the true history is complex (Rogers and Harpending 1992; Rogers 1996). We cannot estimate the parameters of this model directly from genetic data and must content ourselves with estimating

$$\theta_0 = 2uN_0 \tag{1}$$

$$\theta_1 = 2uN_1 \tag{2}$$

$$\tau = 2ut \tag{3}$$

where u is the aggregate mutation rate over the region of DNA under study. Note that a pair of individuals whose last common maternal ancestor lived t generations ago will on average be separated by τ mutations. Thus, τ measures time on a mutational scale.

Previous papers introduced methods for estimating these parameters and for inferring confidence regions (Rogers 1995, 1996). A confidence region is the union of all parameter vectors (θ_0, θ_1, τ) that cannot be rejected at some specified level of significance (Kendall and Stuart 1979, p. 110). In constructing a confidence region, *any* statistical test may be used, provided only that it involves the data and parameters of interest. Some tests yield smaller confidence regions than others, but all of them are valid. Confidence regions are generated by applying some test to a large number of parameter vectors. Those parameter vectors that are rejected at the 0.05 significance level lie outside the 95% confidence region. Those not rejected lie within.

In previous papers (Rogers 1995, 1996), we tried to find a test that would make the confidence regions as small as possible. The resulting confidence regions were acceptably small but were not optimal. We introduce here a test that usually produces confidence regions no larger than the old one yet is far simpler.

The new test, like the old one, uses computer simulations to evaluate a variety of population histories. The following steps are performed for each history:

1. Calculate the theoretical mismatch distribution from the history.
2. Calculate the Mean Absolute Error (MAE) between the observed and theoretical mismatch distributions. Call this the "observed MAE." (The MAE is the mean absolute value of the differences between observed and theoretical mismatch distributions. The observed distribution is normalized so that it sums to unity. The theoretical distribution is truncated so that it equals zero beyond the last nonzero value of the observed distribution, and the final term is augmented so that the truncated theoretical distribution also sums to unity.)
3. Simulate 1,000 data sets using this same history. These simulations use the coalescent algorithm described in appendix A. (Each simulation assumes that the population history has two epochs, the first of which has infinite duration, and that the population is not subdivided in either epoch.)
4. For each simulated data set, calculate the MAE between the simulated mismatch distribution and the theoretical mismatch distribution. Call these "simulated MAEs."

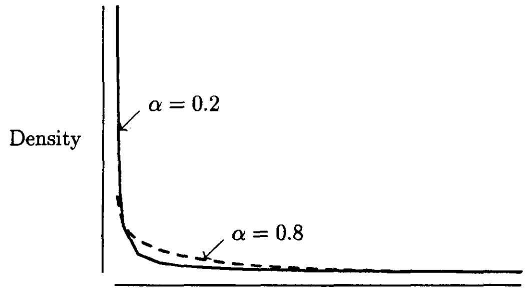

FIG. 2.—Gamma density functions.

5. Reject the history at the 0.05 significance level if the observed *MAE* is greater than 95% of the simulated *MAE*s.

Models of Mutation

In human data, transversions are extremely rare. Thus, the simulation model assumes that all mutations are transitions. The number of mutations along each branch of a gene genealogy is a Poisson random variable with parameter ut, where u is the mutation rate and t is the length of the branch in generations. In mutational time, branch lengths equal $\tau \equiv 2ut$ and the Poisson distribution has parameter $\tau/2$. Since our algorithm specifies the intervals of the population history in mutational time, there is no need to specify a mutation rate.

To relate the number of mutations to the number of substitutions observed in pairwise comparisons, something must be assumed both about the number of sites and also about the distribution of mutation rates across them. We compare three such assumptions:

Infinite sites. As discussed above, this model assumes that every mutation occurs at a distinct site.

Finite sites with uniform rates. This model takes the number of sites to be finite but assumes that all sites mutate at the same rate.

Finite sites with gamma-distributed rates. The number of sites is finite, and the mutation rate at each site is drawn independently from a gamma distribution with density

$$f(x) = \frac{(x/\beta)^{\alpha-1}e^{-x/\beta}}{\beta\Gamma(\alpha)}. \tag{4}$$

Here, β is a scale parameter that need not be specified because it is absorbed by the mutational time scale (see appendix B). α controls the distribution's shape as shown in figure 2. When α is near zero, the distribution is sharply L-shaped, with many sites having mutation rates near zero and a few having much higher rates. Consequently, the number of segregating sites tends to be low when α is small. Kocher and Wilson (1991) estimated that $\alpha = 0.11$ for the entire control region of the human mitochondrial genome, and Wakeley (1993) estimates that $\alpha = 0.47$ for a subset of this region—the first hypervariable region. We will consider a large range that includes these values. We generate gamma random deviates using the algorithm of Ahrens and Dieter (1974).

Sample

Our sample comprises 77 Asian subjects. We sequenced 630 nucleotides from both hypervariable regions of the mitochondrial control region (sites 15981–16410 and 71–270 of the human reference sequence [Anderson et al 1981]). Further details about the sample and about laboratory methods are published elsewhere (Jorde et al. 1995). The mismatch distribution of this sample is shown in figure 1.

Results

The Effect of Mutational Model on Confidence Regions

As a standard for comparison, panel A of figure 3 shows a confidence region calculated using the model of infinite sites. Each circle shows the result of a hypothesis test, open circles indicating rejected hypotheses and filled circles indicating hypotheses that could not be rejected. The open circles are thus outside of the confidence region, and the filled circles are within it. The confidence region rejects the hypothesis of no growth but not the hypotheses of 10-fold, 100-fold, or 1,000-fold growth. It also indicates that the episode of growth occurred more than 2 but less than 14 units of mutational time ago and that the initial population had fewer than $10/(2u)$ females.

We were interested to find out whether different mutational models would lead to different confidence regions. Accordingly, we estimated confidence regions from these same data under a variety of mutational models, each assuming that the number of sites is 630 (as in the data) rather than infinite. The first two models—those of uniform rates and of gamma-distributed rates with $\alpha = 0.9$—led to confidence regions identical to that shown in panel A of figure 3. Differences arose only in the gamma-distributed model, and then only when α fell to 0.1. Panel B shows this confidence region. It is nearly identical to the infinite-sites confidence region, differing in the addition of a single filled circle. Even when $\alpha = 0.05$—well below the published estimates of its value—the confidence region (shown in panel C) is still in fair agreement with that inferred under the model of infinite sites. This confidence region includes the same range of values of τ and the same range for growth (θ_1/θ_0). It differs only in allowing $\theta_0 = 10$, a value excluded by the infinite-sites confidence region. Only when α falls to 0.01 do we see a substantial effect. The resulting confidence region (shown in panel D) is very broad, implying that almost any parameter values are consistent with the data.

This broad confidence region probably reflects the fact, shown several years ago by R. Lundstrom (unpublished manuscript), that if a few sites mutate very fast, then waves appear in the mismatch distribution even if there has been no change in population size. The assumption that $\alpha = 0.01$ implies that most sites are invariant, while a few mutate very fast. As the figure shows, the result is that the mismatch distribution contains little information about history. The question is, are such small values of α plausible?

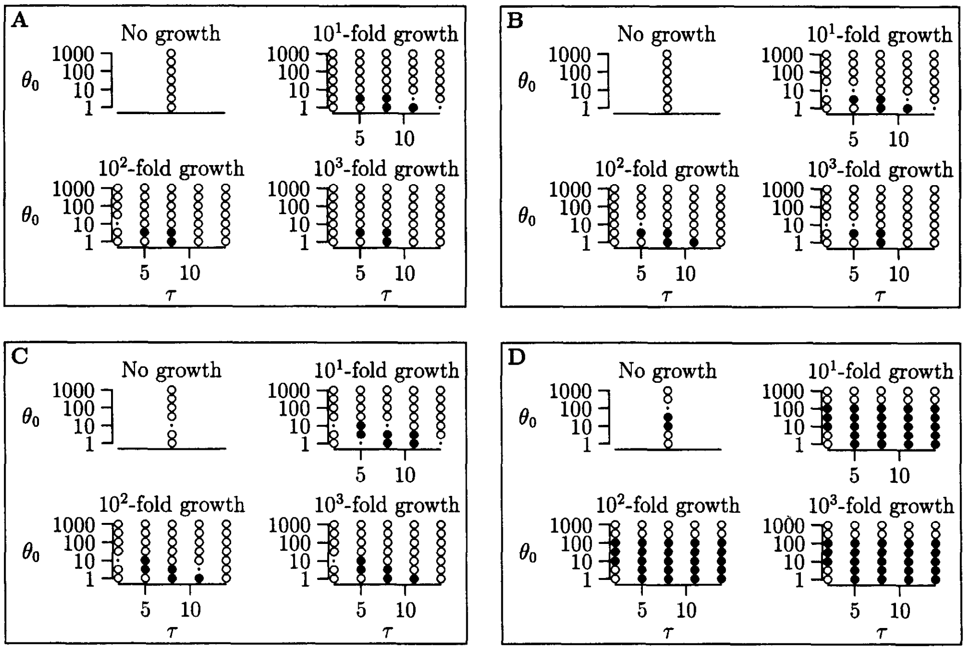

FIG. 3.—Confidence regions. Open circles indicate rejected hypotheses and are thus outside the confidence region. Closed circles indicate hypotheses that could not be rejected and are thus inside. Dots indicate hypotheses for which the algorithm failed to converge. Panel A: infinite sites; panels B–D: gamma-distributed mutations with α equal to 0.1, 0.05, and 0.01, respectively.

Testing Hypotheses About α

As discussed above, published estimates indicate that α = 0.11 for the human mitochondrial control region as a whole (Kocher and Wilson 1991) and 0.47 for the first hypervariable region (Wakeley 1993). However, no confidence interval is available for either estimate, and Wakeley's paper shows that both estimates are biased upward. Consequently, these results to not exclude the possibility that α is small.

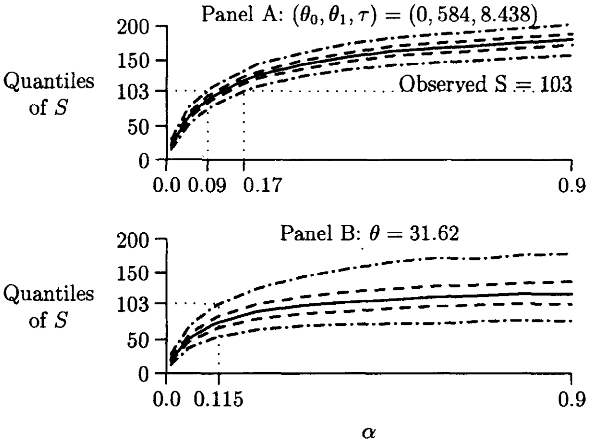

FIG. 4.—Quantiles of S under two models of history. In each panel, 1,000 data sets were simulated at each of several values of the gamma shape parameter α, drawing a fresh set of gamma-distributed mutation rates for each simulation. The number S of segregating sites was calculated from each simulated data set. The solid line is the median of S, the dashed lines enclose the central 50% of the distribution, and the dashed-and-dotted lines the central 95%. The dotted lines construct a confidence region for α. Panel A: Simulations use the population history suggested by the data. Panel B: Simulations assume a constant population size, with θ = 31.62.

We tested various hypotheses about α's value by comparing the number S of segregating sites in simulations with the number (103) in our data. Panel A of figure 4 shows the results from simulations in which the population history parameters (θ_0, θ_1, τ) are set in accordance with the estimates in figure 1. If this population history is correct, then we can reject values of α outside the interval [0.9, 0.17]. As we have seen, the infinite-sites model provides an excellent approximation to the confidence region when α is this large.

But what if panel A is based on the wrong model of population history? To investigate this possibility, we ran a second series of simulations, which assumed a population at equilibrium with θ equal to the observed mean pairwise difference, π = 8.44. These simulations are not shown since the simulated values of S were always much less than the observed value regardless of α's value. To maximize our chances of accepting a small value of α in an equilibrium model, we ran a third series of simulations with θ equal to the largest value not rejected by the confidence region in panel D of figure 3. The results, shown in panel B of figure 4, reject values of α less than 0.115. No larger value of θ is allowed by the confidence region in panel D of figure 3, and smaller values would lead to even larger estimates of α. Consequently, the equilibrium hypothesis implies that α > 0.1, which implies that the confidence region in panel A of figure 3 is correct, which in turn implies that we are not at equilibrium. Thus, the equilibrium hypothesis leads to a contradiction and must be incorrect.

We have not considered all population histories, but those considered all imply that α is large enough to make the infinite-sites confidence region a fair approximation to the true one. Although this does not guarantee that the infinite-sites confidence region is correct, it does justify a reasonable level of confidence.

This conclusion contradicts Bertorelle and Slatkin (1995) and Aris-Brosou and Excoffier (1996), who conclude that mismatch analysis should be regarded with considerable skepticism. They support this position with analyses showing that two statistics—the mean pairwise difference and the number of segregating sites—are sensitive to mutational assumptions. We have no quarrel with this latter claim; it is consistent with our own simulations. We disagree only with the additional claim that such results undermine the inferences of mismatch analysis. These inferences are robust because confidence regions inferred by mismatch analysis are remarkably insensitive to mutational assumptions. The model of infinite sites is patently false but nonetheless useful because it usually provides an excellent approximation to the confidence region.

Comparing π and S

R. Hudson (personal communication) has suggested a different method of evaluating the hypothesis that the pattern in these data was generated by a mutational process such as our gamma-distributed model. Hudson's argument involves Tajima's D (Tajima 1989), a statistic that is proportional to $\pi - S/A$, where π is the mean pairwise difference, S is the number of segregating sites, $A = \Sigma_{i=1}^{n-1} 1/i$, and n is the number of DNA sequences sampled. Hudson's point is that a mutational process that puts most mutations at a small number of sites should lead at equilibrium to positive values of D, yet in our data $D = -2.03$, which is not only negative but highly significant. And negative values are exactly what the hypothesis of population growth predicts.

It is not hard to understand why gamma-distributed mutations should lead to positive values of D: Under this mutational process, most mutations occur at a small number of sites, so S should be very small. This reduces the negative component of D, leading to positive values of D itself. To illustrate this effect, we ran 1,000 simulations of equilibrium populations with gamma-distributed mutations, each using 630 sites and 77 subjects and assuming that $\theta = 10$, $\alpha = 0.1$. Under the infinite-sites model, the mean values of π and of S/A should both have been close to 10. In our simulations they were 7.5 and 6.8, respectively. Since S/A is on average smaller than π, this model generates positive values of D.

To see why population growth leads to negative D, consider the most extreme form of population growth imaginable—that in which a population comprising just one individual grows suddenly to become infinite in size. This will give rise to a star phylogeny. If we sample n individuals from this population t generations after the episode of explosive growth, the expected values of π and of S under the infinite sites model will be $E[\pi] = 2ut$ and $E[S] = nut$. Thus, $E[S]/E[\pi] = n/2$, or 38.5, with a sample of 77. Under an equilibrium model, this same ratio equals A, or 4.91, with a sample of 77. A history of growth thus inflates S relative to π and would generate a negative value of Tajima's D.

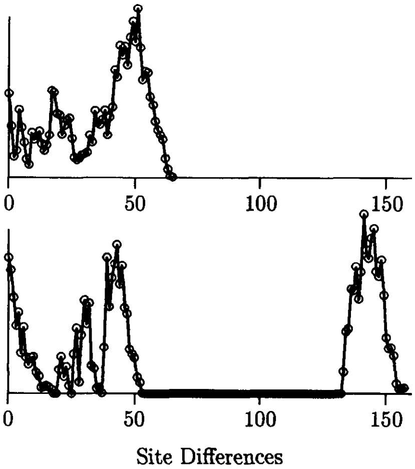

FIG. 5.—Test of the Excoffier-Bertorelle-Slatkin central-limit hypothesis. Mismatch distributions of samples from two simulated populations, each of aggregate size $\theta = 10$, and each divided into 50 groups each exchanging 0.1 migrants per generation according to the island model of population structure.

Clearly, the relationship between π and S in our data is not as predicted by the hypothesis of gamma-distributed mutations at equilibrium. It is, however, consistent with a model of gamma-distributed rates *with* population growth. For example, in 1,000 simulations with $(\theta_0, \theta_1, \tau) = (1, 100, 8)$ and $\alpha = 0.1$, the means $\bar{\pi} = 7.7$ and $\overline{S/A} = 15.0$ imply a negative D. Thus, models of population growth have no difficulty in accounting for strongly negative values of D. We see no way, however, to account for such values under any equilibrium model, even one with an uneven distribution of mutation rates across sites.

Do the Mismatch Waves Reflect the Central-Limit Theorem?

Bertorelle and Slatkin (1995, p. 891) cite L. Excoffier, who suggested that mismatch waves result not from any historical event but from the simple process of pooling data from different parts of a subdivided population. These authors suggest that unimodal waves occur because of the central-limit theorem.

We are unable to understand how the central-limit theorem applies. This theorem describes the result of averaging independent random variables, but no random variables are averaged here. When random variables are averaged, the resulting distribution is a convolution of the distributions of the original random variables. But a pooled mismatch distribution does not convolve the distributions of its subdivisions; it averages them (along with the between-group distributions) (Rogers 1995). The relevance of the central-limit theorem is therefore unclear.

Nonetheless, we ran simulations in order to test this hypothesis, using the geographically structured coalescent algorithm described in appendix A. The results are in figure 5, where each panel shows the mismatch distribution of a simulated population with an aggregate size of $\theta = 10$ containing 50 subdivisions that exchange migrants according to the island model of population structure. If the central-limit theorem applied here, both distributions should be bell-shaped. They are not. We ran a dozen similar simulations, none of which generated a bell-shaped mismatch distribution. We find no support for the proposition that mismatch waves are caused by pooling data from subdivided populations.

Discussion and Conclusions

Mismatch analysis has been criticized for its reliance on the model of infinite sites. The number of segregating sites in our data is indeed smaller than this model would imply. Nonetheless, it is large enough to ensure that this model will yield an excellent approximation to the confidence region of (θ_0, θ_1, τ). Mismatch analysis leads to essentially the same confidence region under a wide range of assumptions about the mutational process. Because of this insensitivity, it is useful even when the mutational process is poorly understood.

These findings echo earlier work showing that in comparisons within the human species, the mean pairwise difference is little affected by gamma-distributed mutation rates until α falls below 0.1 (Rogers 1992). The present results show that this insensitivity applies not only to the mean but also to confidence regions about parameters describing population history.

These results were obtained by studying a single data set and may not apply elsewhere. In particular, the mutational model may become more important when longer spans of time are considered. It would be necessary to repeat much of this analysis before analyzing data from a species with great mitochondrial diversity.

Mismatch analysis has also been criticized for pooling data from various parts of subdivided populations. We investigate and reject the hypothesis that this pooling generates the waves commonly seen in mismatch distributions.

Acknowledgments

We wish to thank Kenneth and Judith Kidd for providing DNA samples. Elisabeth Manderscheid helped collect blood samples. This research was supported by NSF grants BNS–8720330 and DBS–9310105, the Clinical Research Center of the University of Utah (NIH RR–00064), the Technology Access Center of the Utah Human Genome Project, a supplemental equipment grant from the University of Utah, and a CAP fellowship to Michael Bamshad (NIH RR–00064). We are grateful for comments from Elizabeth Cashdan, Henry Harpending, and Richard Hudson.

APPENDIX A

Simulations Using the Coalescent Algorithm

This section describes a geographically structured coalescent algorithm similar to that of Hudson (1990). Our implementation was introduced by Rogers (1996), and the following description is modified only slightly from Rogers's.

The algorithm breaks the population history into an arbitrary number of "epochs," within each of which all parameters are constant. Within epoch i, the population is described by four parameters:

θ_i = $2uN_i$, where N_i is the effective female population size during epoch i;

M_i = the number of migrants per generation between each pair of groups during epoch i;

τ_i = $2ut_i$, where t_i is the length of epoch i in generations and τ_i is its length in mutational time;

K_i = the number of subdivisions during epoch i.

If $K_i = 1$, then M_i is undefined and the entire population mates at random. The earliest epoch is epoch 0 and has infinite duration, i.e., $\tau_0 = \infty$.

The algorithm begins with the final epoch, L. The n individuals of the sample are at first divided evenly among the K_L groups of epoch L. Thus, the algorithm requires that n be evenly divisible by K_L. (The allocation of individuals among groups in the simulation should match that in the data under study. Thus, the allocation used here is most appropriate when the real data include samples of equal size, drawn from several groups.)

As the algorithm moves backward into the past, two types of events occur. Migrations occur when an individual moves from one group to another, and "coalescent events" occur when two individuals have a common ancestor and therefore coalesce to become a single individual.

The hazard h at time τ is defined so that $h\,d\tau$ is the probability that an event of either type will occur between τ and $\tau + d\tau$, where τ measures mutational time looking backward into the past. The hazard depends on prevailing values of the population history parameters, on the number of individuals, and on how these are distributed among groups. At any given time, let s_j denote the number of individuals within group j, $Q \equiv \sum_j s_j$ (the total number of individuals), and $R \equiv \sum_j s_j^2$ (the sum of these numbers squared). Then the hazard of an event is

$$h = [QM_i + (R - Q)/2]/\gamma_i \qquad (5)$$

where $\gamma_i \equiv \theta_i/K_i$, and measures group size in epoch i.

This result can be derived as follows. Let m denote the migration rate per generation, g the group size, and $M \equiv mg$. The hazard per generation is

$$h^* \equiv \sum_j \left[s_j m + s_j(s_j - 1)/(2g)\right] = (1/g)\left[QM + (R - Q)/2\right].$$

The cumulative hazard in t generations is

$$h^*t = \frac{2ut}{2ug}\left[QM + (R - Q)/2\right] \equiv \frac{\tau}{\gamma}\left[QM + (R - Q)/2\right],$$

where $\tau \equiv 2ut$ and $\gamma \equiv 2ug$. Equation (5) follows from the observation that, by definition, the hazard h in mutational time obeys $h\tau \equiv h^*t$.

The algorithm first sets $Q = n$, $R = K_L(n/K_L)^2$, and then sets h using these values together with the parameters of the final epoch, L. It then enters a loop that is executed repeatedly. We describe the steps of this loop briefly before describing each step in detail.

Overview of Coalescent Loop

1. Find the time of the next event, changing epochs and recalculating h as necessary.
2. Determine whether the next event is a migration or a coalescent event.
3. Carry out the next event.

These steps are repeated until $Q = 1$. Mutations are then added along each branch.

Step 1: Let T_i denote the amount of time that we have already traveled (backward) into epoch i. To find the time of the next event, draw a random number x from an exponential distribution whose parameter equals unity. In a constant world, the time of the next event would be $T_i + x/h$. If this time lies within epoch i (i.e., if $T_i + x/h < \tau_i$), then we have found the time of the next event. Otherwise, change epochs as follows:

a. Subtract the portion of x that is "used up" by epoch i, i.e., subtract $h \cdot (\tau_i - T_i)$ from the value of x.
b. Reset population history parameters to those of epoch $i - 1$ and set T_i to zero. If $K_{i-1} < K_i$, join groups at random to diminish the number of groups. If $K_{i-1} > K_i$, increase the number of groups, but allocate no individuals to the new groups. Individuals will enter the new groups only through migration. (The assumption for $K_{i-1} > K_i$ implies that, in forward time, the number of groups has decreased because some groups have died out. Other assumptions are possible and the present one was chosen only for computational convenience.)
c. Reset R and h. Subtract 1 from the value of i.

This process repeats until $T_i + x/h < \tau_i$.

Step 2: Once the time of the next event has been established, step 2 classifies the event as either a migration or a coalescent event. Equation (5) implies that the event is a migration with probability

$$P_M = \frac{QM_i}{QM_i + (R - Q)/2}.$$

Thus, step 2 calls the next event a migration with probability P_M and a coalescent event with probability $1 - P_M$.

Step 3: If the next event is a migration, then move a random individual into a new, randomly chosen group. Then reset R and h.

Otherwise, we have a coalescent event and the procedure is as follows. First choose a group at random, weighting each group by the number of pairs of individuals within it. Then choose a random pair of individuals from within the chosen group, replace the two individuals with a single individual (their common ancestor), reduce Q by 1, and reset R and h.

Finally, mutations are added to the gene genealogy using one of the models discussed above. The number of mutations along a branch of length τ (in mutational time) is a Poisson random variable with mean $\tau/2$. Under the model of infinite sites, the number of mutations equals the number of substitutions. Under either of the finite-site models, the number of substitutions depends also on how mutations are allocated to sites.

To execute this algorithm, it is necessary to specify the sample size n and the parameters (θ_i, M_i, τ_i, and K_i) that describe the population's history. There is no need to specify the mutation rate, the number of individuals in the population, or the number of generations in each epoch.

APPENDIX B

Generating a Vector of Gamma Deviates that Sums to Unity

The mutational time scale requires that the sum of mutation rates across sites equal unity so that there will on average be one mutation per unit of mutational time. To accomplish this goal, we first define $y \equiv x/z$, where x obeys the density in equation (4) and z is an arbitrary scale parameter. The variable y is also gamma-distributed, with density $y^{\alpha-1}e^{-y}/\Gamma(\alpha)$. We first use this density to generate a vector $(y_1, y_2, \ldots, y_K)$. These variates are each equal to x/z. To satisfy the constraint imposed by the mutational time scale, we set $z = 1/\Sigma_i y_i$. In other words, we set $x_i = y_i/\Sigma_i y_i$. This provides a vector of gamma-distributed mutation rates that sums to unity, as required.

LITERATURE CITED

AHRENS, J., and U. DIETER. 1974. Computer methods for sampling from gamma, beta, Poisson, and binomial distributions. Computing **12**:223–246.

ANDERSON, S., A. T. BANKIER, B. G. BARRELL et al. (14 co-authors). 1981. Sequence and organization of the human mitochondrial genome. Nature **290**:457–465.

ARIS-BROSOU, S., and L. EXCOFFIER. 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. **13**:494–504.

BERTORELLE, G., and M. SLATKIN. 1995. The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. Mol. Biol. Evol. **12**:887–892.

HARPENDING, H. 1994. Signature of ancient population growth in a low resolution mitochondrial DNA mismatch distribution. Hum. Biol. **66**:591–600.

HARPENDING, H. C., S. T. SHERRY, A. R. ROGERS, and M. STONEKING. 1993. The genetic structure of ancient human populations. Curr. Anthropol. **34**:483–496.

HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. FUTUYMA and J. ANTONOVICS, eds. Gene genealogies and the coalescent process. Vol. 7. Oxford University Press, Oxford.

JORDE, L. B., M. J. BAMSHAD, W. S. WATKINS, R. ZENGER, A. E. FRALEY, P. A. KRAKOWIAK, K. D. CARPENTER, H. SOODYALL, T. JENKINS, and A. R. ROGERS. 1995. Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. **57**:523–538.

KENDALL, M., and A. STUART. 1979. The advanced theory of statistics. II. Inference and relationship. 4th edition. Macmillan, New York.

KIMURA, M. 1971. Theoretical foundation of population genetics at the molecular level. Theor. Popul. Biol. **2**:174–208.

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. Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding region. Springer-Verlag, New York.

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

MARJORAM, P., and P. DONNELLY. 1994. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136**: 673–683.

ROGERS, A. R. 1992. Error introduced by the infinite sites model. Mol. Biol. Evol. **9**:1181–1184.

———. 1995. Genetic evidence for a Pleistocene population explosion. Evolution **49**:608–615.

———. 1996. Population structure and modern human origins. *In* P. J. DONNELLY and S. TAVARÉ, eds. Population structure and modern human origins. Springer-Verlag, New York (in press).

ROGERS, A. R., and H. C. HARPENDING. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9**:552–569.

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. Hum. Biol. **66**:761–775.

TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37**:613–623.

WATTERSON, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**:256–276.

SIMON EASTEAL, reviewing editor

Accepted April 10, 1996