Population Structure and Modern Human Origins

Alan R. Rogers*†

1997

Abstract

This paper reviews statistical methods for inferring population history from mitochondrial mismatch distributions and extends them to the case of geographically structured populations. Inference is based on a geographically structured version of the coalescent algorithm that allows for temporal variation in population size, in the number of subdivisions, and in the rate of migration between subdivisions. Confidence regions are inferred under several models of population history. If the pattern in mitochondrial DNA reflects population growth rather than selection, then the confidence regions reject the multiregional hypothesis of modern human origins more strongly than has previously been possible. They do not reject the replacement hypothesis.

Keywords coalescent, mitochondrial DNA, modern human origins, mismatch distribution, population structure

1 Introduction

A mitochondrial mismatch distribution is a histogram that describes variation in the amount of genetic difference between pairs of individuals in a sample. In several recent articles, my coauthors and I have suggested that the mismatch distribution is rich in information about population history [19, 17, 6, 21, 5, 18]. This work suggests that the human population experienced a population explosion during the late Pleistocene, some 30,000 to 130,000 years ago.

This work has encountered two kinds of criticism. Some have objected to our use of a theory describing pairs of individuals when our data consist not of pairs but of much larger samples. Others have objected that when populations are subdivided, the mismatch distribution may not contain much information about population history, and methods such as ours should not work [15].

In response to the first criticism, sections 2–5 of this paper will emphasize that the theory in question has never been used as a basis for inference but is used instead as a basis for intuition. It will review the reasons why this approach seems plausible and what it has accomplished. Later sections will emphasize

*Dept. of Anthropology, University of Utah, Salt Lake City, UT 84112. This research was supported in part by National Science Foundation grant DBS–9310105.

†Published in *Progress in Population Genetics and Human Evolution*, edited by Peter J. Donnelly and Simon Tavaré. Springer-Verlag, New York, pp. 55–97, 1997.

that statistical inference has been based on computer simulation, not on the theoretical mismatch distribution.

The second criticism would not appear relevant to the work of Harpending et al [6], whose model incorporates the effect of genetic population structure. It may however bear on other work that assumes a randomly mating population. The final sections of the paper will therefore explore the effect of population structure on statistical methods that I develop elsewhere [18].

2 What is a mismatch distribution, and how can it inform us about history?

Genetic data provide a record of population history that stretches back tens or even hundreds of thousands of years. This record exists for two reasons. First, genetic differences between individuals measure the length of the genealogy that connects them. Second, genealogical distances tend to be longer in large populations than in small ones. For example, a random pair of individuals are more likely to be brothers, and thus connected by a short genealogy, in a population of 100 than in one of 100 million.

To get a feel for this effect, consider Figure 1. The upper panel there shows the history of a hypothetical population, with time measured in units of $1/(2u)$ generations before the present. Here, u is the aggregate mutation rate over the region of DNA under study. This scale of measurement is useful because it makes the time separating two individuals equal to the expected genetic difference between them.[1] For concreteness, I assume that $u = 0.0015$.[2] If each generation lasts 25 years, this mutation rate makes each unit of mutational time equal to 8333 years. In the figure, N_F denotes the effective female population size. The hypothetical population expanded by 500-fold at 7 units of mutational time (58,000 years) before the present.

The middle panel shows a simulated mitochondrial genealogy of 50 individuals, which was generated from this population history using the "coalescent" algorithm that I describe below. The 50 individuals in this sample are represented by 50 horizontal lines at the left edge of the genealogy, which corresponds to the present. The vertical lines in the genealogy mark places where two lineages have a common ancestor and "coalesce" into a single lineage. In this genealogy, coalescent events occur only rarely during the period from the expansion to the present. 35 of the 49 coalescent events are compressed into a relatively brief interval just prior to (to the right of) the expansion. This reflects the history of population size. After the expansion, the population was large and a random pair of individuals was unlikely to share the same mother. Therefore, coalescent events were rare. But prior to the expansion the population was small, and coalescent events were common. The result is that coalescent events are concentrated in a relatively brief interval prior to the expansion. This pattern is characteristic of expanded populations (see Rogers and Jorde [20] for another hypothetical example). It also appears in many gene genealogies estimated from human mtDNA [3].

[1] For example, if two lineages have been separate for $7/(2u)$ generations, the expected number of mutations separating the them is $2u \times 7/(2u) = 7$.

[2] This estimate was obtained by Rogers and Harpending [19] for the data of Cann, Stoneking, and Wilson [2], which I discuss further below.

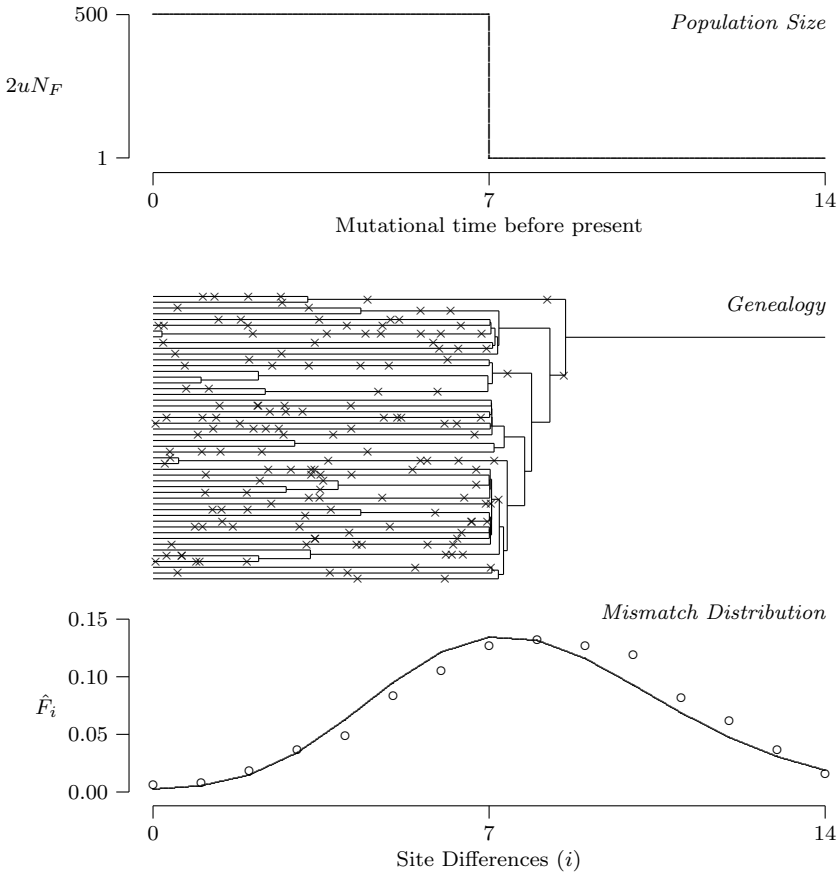

Figure 1: Mitochondrial genealogy and mismatch distribution of a hypothetical population

The top panel shows $2uN_F$ as a function of time before present, with time measured in units of $1/(2u)$ generations. Here, N_F is the effective female population size and u the aggregate mutation rate over the region of DNA under study. The population was small prior to time 7. The middle panel shows the genealogy of a sample of 50 individuals drawn from this population. The crosses represent mutations. The open circles in the bottom panel show these same data as a mismatch distribution. The solid line there shows the theoretical mismatch distribution for the parameters of the hypothetical population.

A wave such as that in Figure 1 might also have been produced by selection rather than population growth. Under this interpretation, Figure 1 tells a different story: A favorable mitochondrial allele appears by mutation at 7 units of mutational time before the present and then spreads rapidly to fixation. In the upper panel of the figure, N_F now refers not to the female population size but to the number of female descendants (or maternal ancestors) of the mutant female. There were few individuals in this lineage before the mutation (in fact there was only one: the mutant female has a single maternal ancestor in each generation) but many in the generations that followed the mutation. Although natural selection is responsible for the increase in the size of this lineage, the individuals within the lineage are selectively neutral with respect to each other, since each carries a copy of the same mitochondrial allele. Thus, it is appropriate to assume that variation within the lineage (and within our modern sample) is selectively neutral. This selective interpretation therefore leads to a genealogy and a mismatch distribution that are indistinguishable from that produced by population growth. Since the mismatch distribution is consistent with two interpretations (population growth and natural selection), the choice between these interpretations must be based on other data [20].

The crosses on the genealogy in Figure 1 represent mutations, which occur randomly along each branch. If this were a real population, we could count the mutational differences between pairs of individuals,[3] but we could not know either the true genealogy or the population history. These could only be estimated.

But any effort to infer the genealogy in Figure 1 from genetic data would be doomed to failure. In these data, nearly all of the 157 mutations occur after the expansion, in the part of the genealogy with few coalescent events. There are 35 coalescent events prior to the expansion but only 7 mutations. Consequently, no statistical method could succeed in telling us much about the topology of this genealogy—the data are essentially devoid of phylogenetic information. This example shows how a population expansion (or the selective sweep of a favorable allele) can lead to data with low phylogenetic resolution. With such data there is little point in trying to reconstruct the gene genealogy.

This is not to say that methods of phylogenetic inference are useless. Even when these methods cannot tell us the topology of the tree, they might still tell us that coalescent events were clustered in a narrow interval of time [3, 8]. This would imply a small effective population size during this interval. Thus, the data may tell us about population history even if they are devoid of phylogenetic information.

But when the sample of individuals is large, phylogenetic inference is a formidable business. There is no efficient way to search the immense set of possible genealogies for those which best describe the data, and computer runs take many hours. The method to be described below is a short-cut that avoids this problem. To introduce it, I turn once again to the hypothetical data in Figure 1.

Let us assume that each mutation produces a detectable nucleotide site difference (the so-called model of "infinite sites" [12] that is discussed in footnote 3). The number of site differences between each pair of individuals in

[3]Strictly speaking, this is not so. We can only count nucleotide (or restriction) site differences between pairs of individuals, and such a difference may reflect more than one mutation [13]. This issue is discussed further in the "Discussion" section.

Figure 1 is then equal to the number of crosses along the path connecting them. For example, there are six site differences between the top-most pair of individuals in the genealogy. With 50 individuals in the sample, there are 1225 pairs of individuals, and we can count the differences between each pair. The open circles in the lower panel represent these 1225 differences as a scatter plot. Such plots are sometimes called "distributions of pairwise differences," and sometimes "mismatch distributions" [7, 6]. For simplicity, I will use the latter term. Notice that the mismatch distribution in Figure 1 peaks just to the right of the point $i = 7$. This is because, as the genealogy shows, many pairs of individuals are separated by a little more than 7 units of mutational time and are therefore expected to differ by a little more than 7 mutations (see footnote 1). Thus, the mismatch distribution peaks just prior to the expansion at a point corresponding to the part of the genealogy at which coalescent events are concentrated.

Had the expansion happened earlier, the peak would have been farther to the right. As time passes the peak will move from left to right, traversing one unit of the horizontal axis in $1/(2u)$ generations [19]. Thus, the distribution looks and acts like a wave moving very slowly from left to right. The horizontal position of the wave measures time since the expansion in mutational time units.

This example suggests that the mismatch distribution might provide information about the history of population size and/or natural selection. Unfortunately, there is no statistical theory to tell us how this information can best be extracted. We can of course define ad hoc statistics and explore their behavior through computer simulation, but it is difficult to know in advance which statistics are likely to prove useful. To use simulations effectively, we need some basis for intuition about the behavior of the mismatch distribution. To gain such intuition, it is useful to study the "theoretical mismatch distribution."

3 The theoretical mismatch distribution

There are no simple theoretical formulas for subdivided populations, so I shall rely on results for a population that mates at random. Even there, we have no explicit formulas for samples of arbitrary size and must make do with formulas for samples of only two individuals. Watterson [23] showed how to calculate the probability that two individuals would differ by i nucleotide sites in a population of constant size. His model is compared to simulated data in Figure 2. Watterson's theoretical mismatch distribution is drawn as a solid line in the lower panel. The contrast between it and the simulated mismatch distribution—shown by the open circles—could not be greater. Whereas the theoretical distribution declines smoothly from a maximum at $i = 0$, the simulated distribution is ragged, with multiple peaks and a maximum value at $i = 93$. To recover the theoretical curve, we would need to average a large number of simulated mismatch distributions with the same population history. With real data, this would require the impossible—averaging mismatch distributions from a series of parallel worlds [22].[4] Since we have only one world to study,

[4]One reviewer disagreed with this claim, so I will provide a proof: Let

$$\delta_{ij}(k) \equiv \begin{cases} 1 & \text{if the } i\text{th and } j\text{th DNA sequences in the sample differ by } k \text{ sites} \\ 0 & \text{otherwise} \end{cases}$$

If the sample was drawn at random, then for any distinct i and j the expectation $E[\delta_{ij}(k)]$ is by definition equal to F_k, the kth term of the theoretical mismatch distribution. The empirical

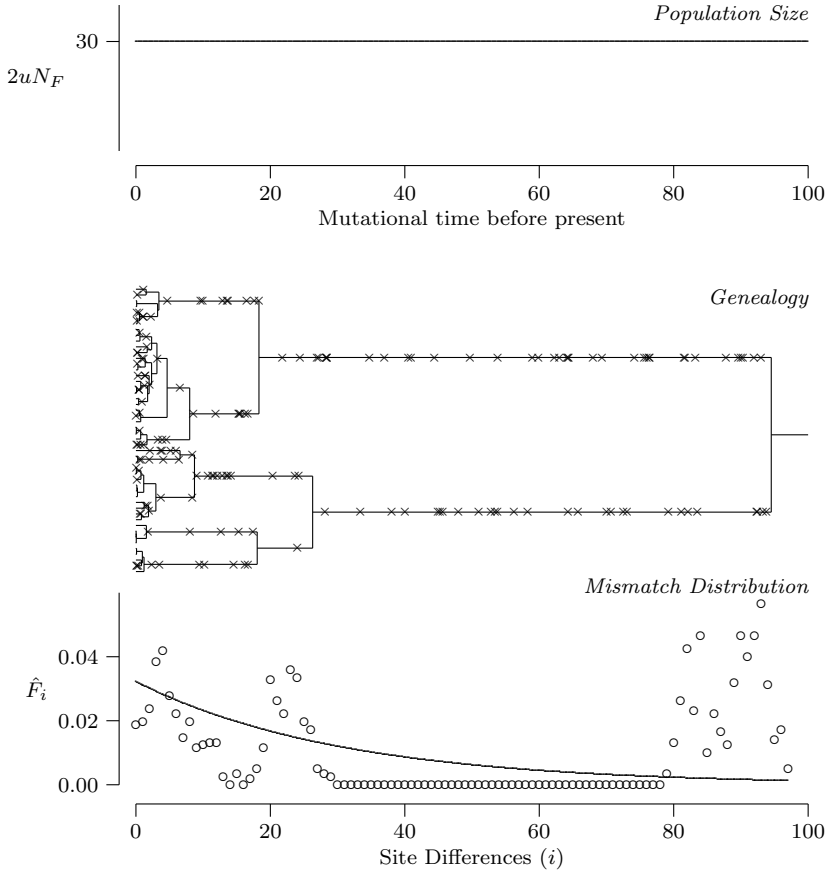

Figure 2: Mitochondrial genealogy and mismatch distribution of a hypothetical equilibrium population
This hypothetical population has always had a constant size, $N_F = 30/(2u)$. Definitions are as in figure 1.

there are grounds for skepticism about the utility of the theoretical mismatch distribution.

But let us press on, nonetheless, to the case in which population size has not been constant. Li [14, Equation 5] developed the relevant formula, which was used by Rogers and Harpending [19] to study population histories such as that in Figure 1. Their Equation (4) was used to draw the solid line representing the theoretical distribution in the lower panel of that figure. The agreement between theory and simulated data is much better in the non-equilibrium case (Figure 1) than in the equilibrium case (Figure 2). This is no fluke: theory and simulated data often agree in expanded populations, provided that the initial population was fairly small [19]. Thus, the theoretical formula may be useful after all as a basis for intuition about the empirical distributions of expanded populations.

This argument is not rigorous, but it doesn't need to be. I am trying to justify using the theory as a basis for intuition, not as a basis for inference. Below, the theory will suggest which statistics should be calculated, and how they might be related to parameters describing population history. But these suggestions are only tentative, and are therefore checked by computer simulation. Thus, statistical inference will be justified by computer simulation, not by appeal to the theoretical mismatch distribution.

4 What should be estimated?

We cannot hope for a complete description of the population's history. That would require one parameter—the population's size—for each time period. If the population were subdivided, we would need additional parameters for migration rates in each time period. Yet it is never possible to estimate more than a few parameters at once. We must content ourselves with some simplified representation of population history.

Fortunately, the theoretical mismatch distribution suggests that a simple model may be useful. The population history in Figure 1 has just three parameters: N_0 (the female population size before expansion), N_1 (the post-expansion size), and t (the time in generations since the expansion). Unfortunately, these parameters are all confounded with the mutation rate so that the mismatch distribution depends only on

$$\theta_0 \equiv 2uN_0 \quad (1)$$

$$\theta_1 \equiv 2uN_1 \quad (2)$$

$$\tau \equiv 2ut \quad (3)$$

where as before u is the aggregate mutation rate over the region of DNA under study. Because N_0, N_1, and t are confounded with u, it is not possible to

mismatch distribution is $\hat{F}_k \equiv \binom{n}{2}^{-1} \sum_{i<j} \delta_{ij}(k)$, where n is the number of DNA sequences in the sample. Its expectation is therefore

$$E[\hat{F}_k] = \binom{n}{2}^{-1} \sum_{i<j} E[\delta_{ij}(k)] = F_k$$

Thus, the theoretical mismatch distribution is the expectation of the empirical distribution. Were it possible to average independent realizations of $\hat{F}_k$, the law of large numbers would guarantee that this average would converge to F_k as the number of cases became large.

estimate any of these parameters directly from genetic data. Consequently, I address myself to the problem of estimating θ_0, θ_1, and τ.

This three-parameter model is not complex enough to describe the history of any real population, but it may nonetheless provide a fair description of the data. Analysis of the theoretical mismatch distribution [19] shows that the three-parameter model is robust in several ways: (1) When a population's size is small, convergence to the 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 not unreasonable to assume that the pre-expansion population was at equilibrium. (2) Instantaneous growth has an effect on the mismatch distribution that is indistinguishable from exponential growth over thousands of years. (3) After the population has grown large, subsequent episodes of growth and minor bottlenecks have little effect on the mismatch distribution. Because of these properties, the three-parameter model should prove useful even in populations whose histories are more complex than that in Figure 1.

This conclusion is based on the theoretical mismatch distribution, and should therefore be regarded with caution. It may not hold in circumstances where empirical and theoretical distributions tend to differ. Consider therefore the hypothetical population whose history, genealogy, and mismatch distribution are shown in Figure 3. This population is similar to that in Figure 1 in that it too experienced a burst of growth at mutational time 7, was small for a long while before, and was generally large thereafter. But there the similarity ends. The population in Figure 3 has seen several growth spurts and minor bottlenecks. Even the spurt at time 7 is different, being exponential rather than instantaneous. Yet none of this has any important effect. The genealogies of the two populations both show the same pattern—coalescent events are concentrated in a brief interval just prior to the population expansion. The lower panel of Figure 3 includes both the theoretical distribution of the complex population history (shown as a solid line) and of the simple population history (the dotted line). There is hardly any difference between them. This illustrates the robustness of the theoretical mismatch distribution [19]. The empirical distributions shown in Figures 1 and 3 are also very similar; the difference between them is no larger than that typically seen in different distributions simulated with the same population history. Were we to analyze the data in Figure 3 using the simpler population history of Figure 1, we would not be led astray. We would conclude correctly that a substantial episode of growth had occurred at around $\tau = 7$. Thus, the simple three-parameter model of population history can be useful even with populations whose histories are far more complex.

It would be easy to over-interpret these examples. While they suggest that the empirical mismatch distribution is insensitive to many details of population history, they don't amount to a proof. Furthermore, they deal only with the history of population size and therefore tell us nothing about the effect of other assumptions that may be violated. It is not even true that the difference between the population histories of Figures 1 and 3 has no effect on the mismatch distributions, for the second history very occasionally produces distributions with a second peak very far to the right. I found 2 such distributions in 100 trials with the complex history but none with the simpler history. All the other simulated distributions looked similar to those in Figures 1 and 3. Since the two histories produce similar distributions 98 times in 100, it is reasonable to

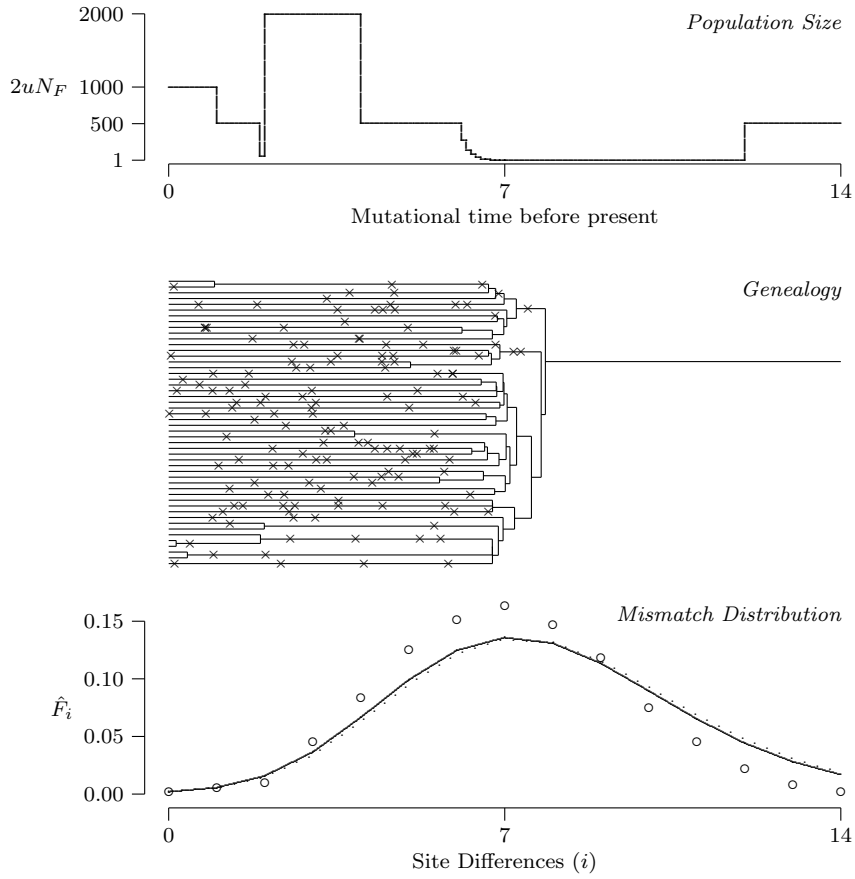

Figure 3: The minimal effect of a complex population history
The dotted line shows for comparison the theoretical mismatch distribution from Figure 1. Everything else is as defined in Figure 1.

estimate the parameters of the simple history in either population. The rare outliers of the complex history will however affect the statistical properties of these estimates. For example, the mean depth of genealogies from the complex history is about twice that of genealogies from the simpler history. This will require further discussion below in the section on confidence intervals.

5 Estimators

Statisticians have developed various methods for finding statistics to estimate particular parameters. The simplest of these, called the *method of moments*, proceeds by equating theoretical moments (the mean, the variance, and so on) with observed moments and solving the resulting equations. There is no good reason for confidence that this method will yield well-behaved estimators. The problem is that mismatch distributions do not offer a set of independent, identically distributed observations. Each pair of individuals in the data is correlated to a greater or lesser degree with many others. Thus, computer simulations will be needed not only to determine the statistical properties of the estimators proposed below but also to verify that they behave as estimators at all.

Given three parameters, θ_0, θ_1, and τ, a straightforward application of the method of moments would use three equations, obtained from the theoretical formulas for the mean, the variance, and the skewness. This approach works poorly here, because it is often impossible to solve these three equations.

However, the theoretical mismatch distribution tells us that there is usually very little information about θ_1 anyway. A large value of θ_1 has essentially the same effect as an infinite value. Therefore, we may hope to estimate θ_0 and τ from a model in which $\theta_1 \to \infty$. In this case, there are only two equations, and these have a simple solution [18]:

$$\hat{\theta}_0 = \sqrt{v - m} \tag{4}$$

$$\hat{\tau} = m - \hat{\theta}_0 \tag{5}$$

where m is the mean and v the variance of the empirical mismatch distribution. I propose to use these statistics as estimators.

This proposal, however, is only based on intuition—the intuition provided by the theoretical mismatch distribution. To justify this intuition it is necessary to show that $\hat{\theta}_0$ and $\hat{\tau}$ do in fact behave as estimators. For each set of parameter values, the sampling distributions of the estimators should ideally be narrow, and centered around the true parameter values.

I have verified this behavior elsewhere [18] and will here include only a single figure. Figure 4 was obtained by simulating 1000 data sets at each of several values of τ and using each of these to calculate $\hat{\tau}$. At each value of τ, the 1000 estimates were used to estimate the quantiles of $\hat{\tau}$, and these quantiles are shown in the figure. At each value of τ, the median (shown as a solid line) is close to the bold dot that marks the true value of τ, and the distribution is relatively narrow. Thus, not only does $\hat{\tau}$ behave as an estimator, it is an estimator with admirable statistical properties. A similar analysis [18] shows that $\hat{\theta}_0$ is also well behaved when $\theta_0 \geq 1$ but is incapable of discriminating values in the range $0 < \theta_0 \leq 1$. This is not a serious limitation. It means only that when $\hat{\theta}_0 \approx 1$, the confidence interval will reach all the way to zero.

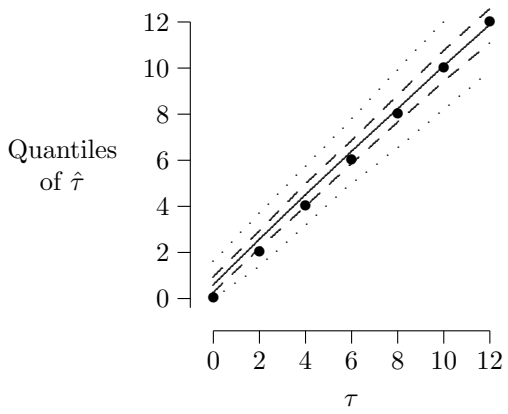

Figure 4: Quantiles of $\hat{\tau}$

At least 1000 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 147 subjects.

6 Simulating data sets with population structure

Thus far, I have merely summarized previous work, which assumed a randomly mating population. What happens when these methods are applied in a subdivided population? To find out, I have implemented the geographically structured coalescent algorithm described by Hudson [9]. My implementation 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;

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 last epoch, which I denote as 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.[5]

As the algorithm moves backward into the past, two types of event 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.

[5]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.

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 backwards 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, $S \equiv \sum_j s_j$ (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[6]

$$h = [SM_i + (R - S)/2]/\gamma_i \tag{6}$$

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

The algorithm first sets $S = n$, $R = K_L(n/K_L)^2$, 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. I 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 $S = 1$. Mutations are then added along each branch.

Step 1 Let T_i denote the amount of time that we have already traveled (backwards) 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 off 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

[6]Let m denote the migration rate per generation, g the group size, and $M \equiv mg$. The hazard per generation is

$$h^* \equiv \sum_j [s_j m + s_j(s_j - 1)/(2g)] = (1/g)[SM + (R - S)/2]$$

The cumulative hazard in t generations is

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

where $\tau \equiv 2ut$ and $\gamma \equiv 2ug$. Equation 6 follows from the observation that, by definition, the hazard h in mutational time obeys $h\tau \equiv h^* t$.

individuals to the new groups. Individuals will enter the new groups only through migration.[7]

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 6 implies that the event is a migration with probability

$$P_M = \frac{SM_i}{SM_i + (R - S)/2}$$

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 S by 1, and reset R and h.

Mutation I use the infinite sites model of mutation, which implies that the number of mutations along each branch is a Poisson random variable with parameter ut, where u is the mutation rate and t the length of the branch in generations [12]. In mutational time, branch lengths equal $\tau \equiv 2ut$ and the Poisson distribution has parameter $\tau/2$.

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.

7 Confidence regions

A 95% confidence region is a set of parameter values constructed by any procedure that guarantees the following property: *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 [11, p. 110].

I shall apply this method to hypotheses about population history. A hypothesis is rejected at significance level α if it implies that data sets "at least as

[7] 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.

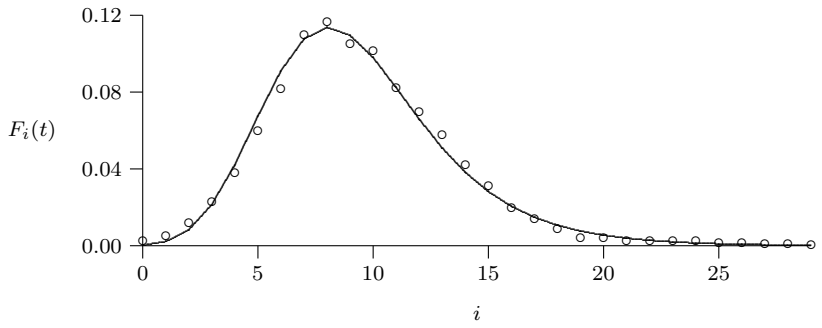

Figure 5: The model of sudden expansion fit to the data of Cann, Stoneking and Wilson

The open circles show the mismatch distribution (the relative frequencies of pairs of individuals whose mtDNA samples differ by i sites) of Cann, Stoneking, and Wilson [2, fig. 1]. The solid line shows the fit of the non-equilibrium distribution [19] obtained using the two-parameter method of moments estimator [18]. This figure summarizes comparisons among $n = 147$ subjects.

extreme" as the real data occur with a frequency less than α. Given a precise definition of "at least as extreme," it is easy to estimate the frequency of such events. I generate a large number of simulated data sets using the coalescent algorithm described above, and estimate α by the fraction of the simulated data sets that are at least as extreme as the observed data.

It remains to decide when a simulated data set will be deemed at least as extreme as the observed data. This decision might be made in any number of ways, all of which would yield valid confidence regions. Yet these definitions would not all be equally useful. I have tried to design a test that would yield small confidence regions in expanded populations. The test is described in full elsewhere [18] and is summarized only briefly here: A simulated data set is deemed to be at least as extreme as that observed if (a) the Mahalanobis distance [10, pp. 423–424] between the vector $(\hat{\theta}_0, \hat{\tau})$ and the simulation mean is at least as large for the simulated data as for the real data, and (b) the mean squared error (MSE) between empirical and theoretical mismatch distributions is at least as large for the simulated data as for the observed data. Occasionally, the covariance matrix used in calculating the Mahalanobis distance turns out to be singluar and this test fails to provide an answer.

This method yields information not only about θ_0 and τ, but also about θ_1, because for given values of θ_0 and τ, the MSE tends to decrease with θ_1. Simulations show that this method yields confidence regions that usually enclose the true parameters values and are reasonably small in expanded populations under random mating [18].

In structured populations, these confidence intervals could presumably be made smaller by using a test that distinguishes within-group from between-group variation [6]. Nonetheless, I will work instead with distributions calculated from the population as a whole, using the test developed for my earlier, random-mating model [18]. This will make it possible to assess the bias that is introduced when that model is applied to data from a structured population. Thus, I take no account of population structure when estimating parameters.

The statistical properties of these estimates, however, will be investigated under various assumptions about population structure. To make comparisons simple, I also analyze the same data set that was used in that earlier publcation. These data were published by Cann, Stoneking, and Wilson [2] and are shown in Figure 5.[8]

8 Population structure

In this section I consider three models of population history. The first—that of world-wide random mating—is not realistic. I include it because it is simple, because it has been used before [18], and because I want to find out whether it yields useful answers even when reality is more complex. The second model assumes that the human population is been subdivided as far as we can see back into the past, while the third assumes that subdivision appeared more recently. These models represent the multiregional and the replacement models of modern human origins, respectively.

8.1 Random mating

My earlier confidence intervals [18] made the simplest possible assumption about population structure: that of an undivided randomly mating population. With the present model, this case corresponds to a population history of the following form:

Epoch	θ_i	M_i	τ_i	K_i
1	θ_1	0	τ	1
0	θ_0	0	∞	1

In each epoch, there is only a single subdivision ($K_i = 1$) and there can of course be no migration between subdivisions ($M_i = 0$). In an earlier publication [18] I used this population history together with the data in Figure 5 to infer the confidence region shown in Figure 6. The open circles there represent hypotheses that were rejected at the 0.05 significance level. The filled circles represent hypotheses that could not be rejected. Thus, the filled circles delimit a 95% confidence region for the parameters defined in Equations 1–3. The confidence region implies that $\theta_0 < 10$, $\theta_1/\theta_0 > 100$, and that $4 < \tau < 9$. Yet these results rely on an assumption that has surely been violated: They assume random mating, whereas the human population is geographically structured.

The history of this structure is a matter of debate. There are two competing views, which I discuss below.

8.2 The multiregional hypothesis

The multiregional hypothesis of modern human origins [24, 4] holds that our species evolved within a widespread population that has inhabited much of Europe, Africa, and Asia for the past million years. Favorable mutations arising

[8] Do not read too much into the close fit between the observed and the theoretical distributions. This fit provides no strong support either for the theory or for the statistical methods. See my earlier paper [18] and the discussion above dealing with Figure 4.

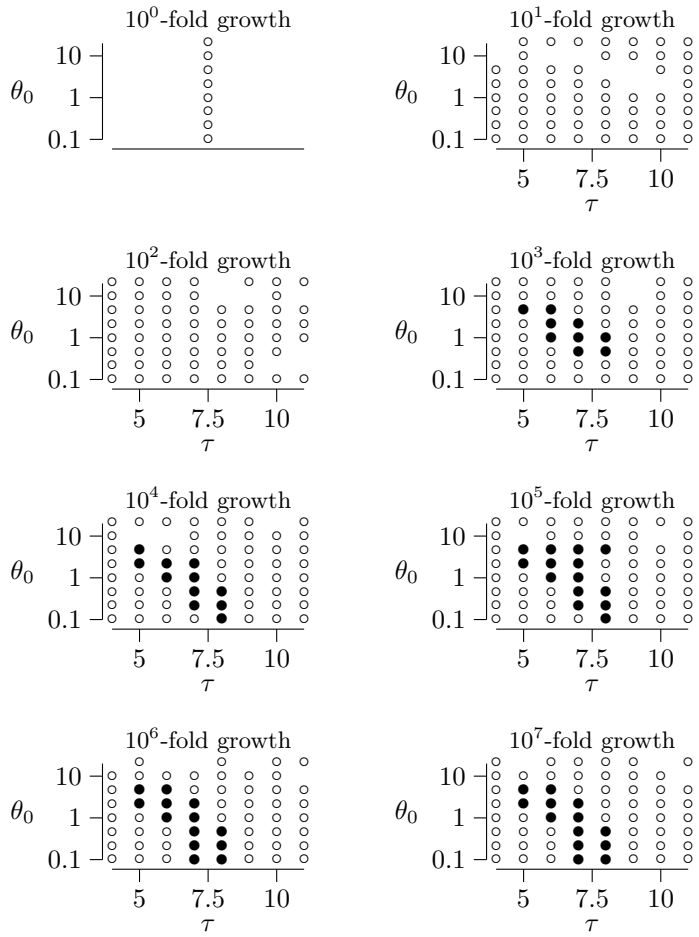

Figure 6: A 95% confidence region for the CSW data under the assumption of random mating

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 $\theta_1/\theta_0 = 10^x$. Missing circles indicate parameter values for which no test was possible because the covariance matrix of $(\hat{\theta}_0, \hat{\tau})$ was singular. Data are from Cann, Stoneking, and Wilson [2]. Reproduced from Rogers [18].

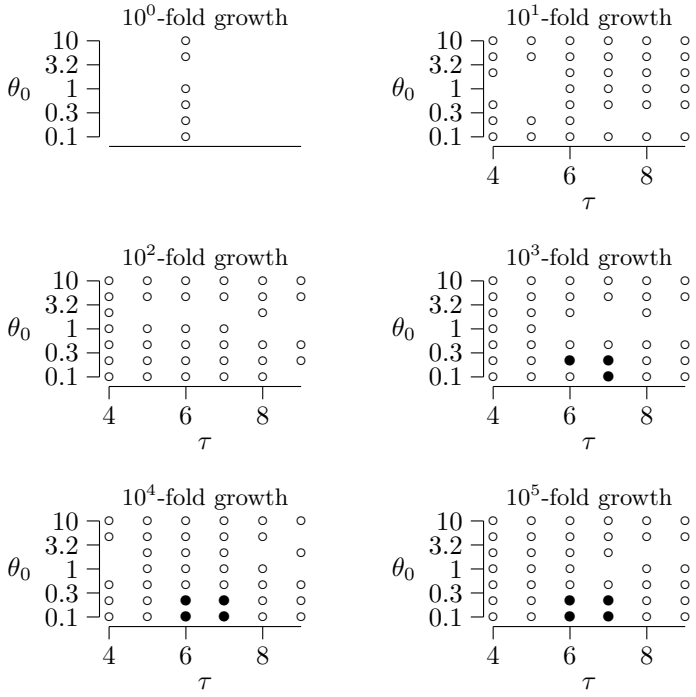

Figure 7: Multiregional hypothesis, $M = 0.1$

in one location spread to others by gene flow rather than by replacement of entire populations.

This hypothesis implies that the geographic structure of our species goes back at least a million years, originating well before the common mitochondrial ancestor. For our purposes, this is equivalent to assuming that the structure has existed forever. Thus, I use a population history of form

Epoch	θ_i	M_i	τ_i	K_i
1	θ_1	M	τ	3
0	θ_0	M	∞	3

with $K_0 = K_1 = 3$ to represent the three major races, and migration is measured by the parameter M. In words, this history assumes that the population has always been divided into three groups, which have always exchanged M migrants per generation. The history allows for a change in population size from θ_0 to θ_1 at τ units of mutational time before the present.

This history was used to generate the confidence regions shown in Figures 7–9. The three confidence regions differ in their assumptions regarding the level M of migration. Figure 7 assumes that migration is weak ($M = 0.1$), Figure 8 that it is moderate ($M = 1$), and Figure 9 that it is strong ($M = 10$). In all of three figures, the confidence region is smaller than that under random mating and includes no parameters values that are not also included within the random-mating confidence region. There is one important difference: The multiregional hypothesis requires that $\theta_0 < 2.15$, whereas the model of random mating requires only that $\theta_0 < 10$.

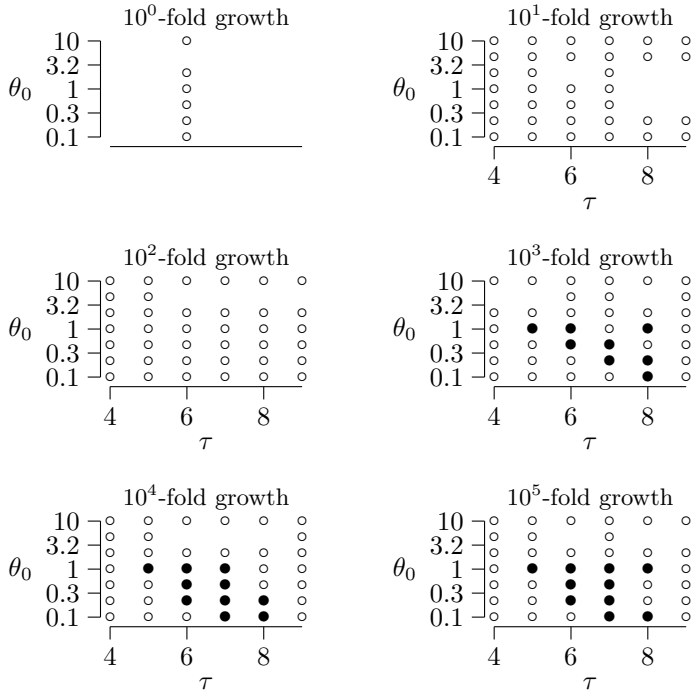

Figure 8: Multiregional hypothesis, $M = 1$

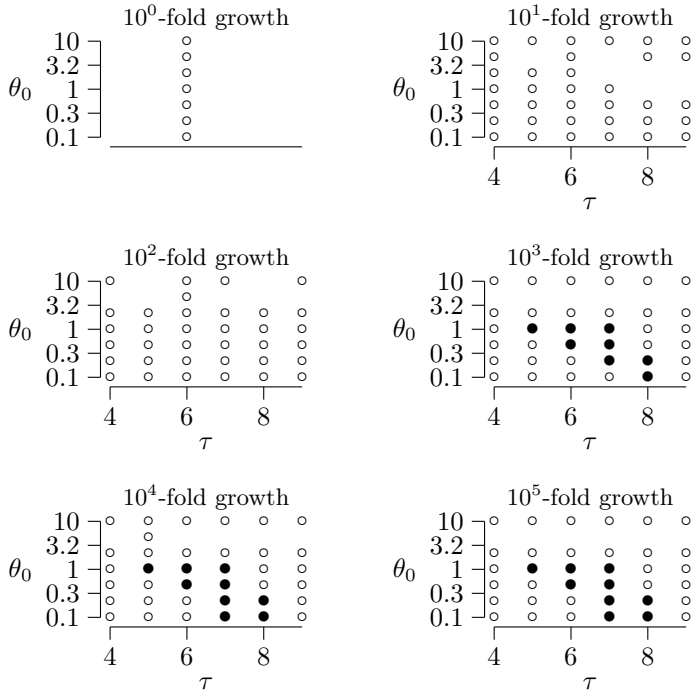

Figure 9: Multiregional hypothesis, $M = 10$

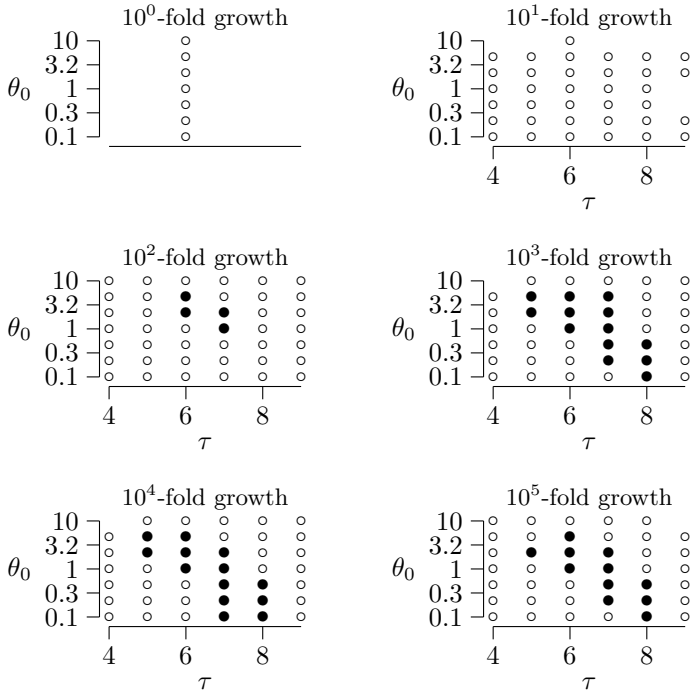

Figure 10: Replacement model, $M = 0.1$

Using published estimates of u, my previous paper found that the random-mating result requires that $N_0 < 7000$ breeding females [18].[9] Since this number seemed too small to populate the continents of Europe, Africa, and Asia, I viewed this estimate as evidence against the multiregional hypothesis. Yet now it is clear that my earlier estimate was too generous. The reduced upper bound on θ_0 implies that $N_0 < 1500$ breeding females. Thus, if the wave in the mitochondrial data reflects population growth rather than selection, then the analysis *with* population structure rejects the multiregional hypothesis even more strongly than the one *without*.

It seems clear, moreover, that this conclusion would not be altered much by other assumptions about M. The upper bound on θ_0 increases with M (compare Figures 7–9), so smaller values of M would not lead to any favorable assessment of the multiregional hypothesis. On the other hand, larger values of M might do so. However, as M grows large, the confidence region will approach the random-mating result, which still allows only 7000 breeding females. Thus, it is not possible to salvage the multiregional hypothesis by a judicious choice of M.

8.3 The replacement hypothesis

The replacement hypothesis holds that modern humans evolved in one region and spread from there throughout the world some 50,000–100,000 years ago,

[9] I used the smaller of the two published estimates of u ($\hat{u} = 7.5 \times 10^{-4}$) in order to make the upper bound on N_0 as large as possible.

replacing earlier peoples as they went. This view implies that the geographic structure of the human population is less ancient, having developed after the initial expansion of modern humans. Thus, the population history becomes

Epoch	θ_i	M_i	τ_i	K_i
1	θ_1	M	τ	3
0	θ_0	M	∞	1

The difference here is that $K_0 = 1$, rather than 3. In words, this says that prior to the expansion that occurred τ time units ago, there was only 1 ancestral population rather than 3.

The confidence interval generated under this hypothesis is shown in Figure 10 and is similar to that in Figure 6. The main difference seems to be that the hypothesis of random mating rejects 100-fold growth, while the replacement hypothesis does not. This makes sense, since subdivision increases a population's effective size [16]. Thus, a 100-fold increase that is combined with subdivision is equivalent to a larger increase without subdivision. Apart from this amendment, the replacement hypothesis is well approximated by a model of random mating and allows the initial population to be nearly five-fold larger than does the multiregional hypothesis. This paradoxical result says that if our ancestors were all in one place then their population may have been of moderate size ($N_0 < 7000$), yet if their population was subdivided then it must have been extremely small ($N_0 < 1500$).

9 Discussion

Two studies [15, 6] have shown that when the initial population is relatively large, and especially if it is geographically structured, mismatch distributions are often rough and ragged like that in Figure 2. Smooth waves such as that in Figure 1 occur only when the initial population is extremely small. These results have been interpreted as evidence that statistical inference from mismatch distributions is a perilous business [15], but I would argue otherwise. Indeed, the raggedness of these distributions makes my confidence regions smaller. According to the confidence regions, the upper bound on θ_0 is 2.15 when the initial population is strongly structured but 10 when it mates at random. This is because intermediate values such as $\theta_0 = 5$ produce ragged mismatch distributions if the initial population is structured but not if it mates at random. Since ragged distributions look nothing like the observed distribution (Figure 5), the statistical method rejects the parameter values that generate them. Far from being a problem, raggedness made the present estimates more accurate [5].

On the other hand, raggedness is not always a blessing. When the observed mismatch distribution is ragged, the present statistical methods yield large confidence regions [18]. But this is as it should be. Large confidence regions may be disappointing, but they are unlikely to lead us astray. They demand no more caution than is normal in statistical analysis.

The results presented here suggest that geographic structure affects the mismatch distribution primarily by way of effective population size. Effective size is larger if a population is structured than if it mates at random [16]. Consequently, a population with a structured initial population behaves like one with a large initial population—its mismatch distribution tends to be ragged. By

the same token, a population that expands and at the same time subdivides behaves like a population that has undergone a much larger expansion. This accounts for the fact that a 100-fold expansion is rejected under the model of random mating but not under the replacement hypothesis. A smaller expansion is allowed in the latter case because geographic structure makes the effective expansion larger than the expansion in numbers.

The expansion that is inferred here can be interpreted in two ways. It may have been either an expansion in population size or the expansion in frequency of an advantageous mitochondrial allele. Harpending et al. [6] and Rogers and Jorde [20] present arguments in favor of the former interpretation, but additional data will be needed to settle the issue.

There is also cause for concern about my use of the infinite sites model of mutation. R. Lundstrom (unpublished data) has shown that with a finite number of sites and mutation rates that vary from site to site, waves can be generated in the theoretical distributions even of equilibrium populations. Fortunately, my own calculations indicate that this effect is unlikely to be important in the theoretical distributions considered here [17]. But even if this effect is negligible in theoretical distributions, its effect on the statistical distribution of my estimates may be important [1].

Finally, it has been argued that mismatch distributions should be interpreted with special caution because we have only one world to study [15, 1]. Because of this limitation, the empirical mismatch distribution amounts to a single observation from a distribution that (depending on parameter values) may be highly variable. This problem is real, for it makes parameter estimates less precise—confidence regions would surely be smaller with data from parallel worlds. Yet it is not a fatal problem, since the confidence regions are small enough to be useful even without parallel worlds.

10 Summary

This paper began with a review showing how the theoretical mismatch distribution has been useful in previous research as a basis for intuition. It then introduced a method for inferring confidence regions with data from subdivided populations, emphasizing that these methods are based on computer simulation, not on the theoretical mismatch distribution. The statistical methods show that if the pattern in the mitochondrial data reflects population growth rather than selection, then (1) the multiregional hypothesis of modern human origins is rejected more strongly than before, (2) the replacement hypothesis of modern human origins is not rejected but allows the expansion of population size to be smaller than did my earlier random-mating model, (3) the model of random mating yields a confidence region that encompasses that of the multiregional hypothesis and differs only slightly from that of the replacement hypothesis. (4) Population structure does not make confidence intervals larger, at least for the data considered here. Consequently, these results provide no support for the view that population structure reduces the value of mismatch distributions for statistical inference.

Acknowledgements

I thank Peter Donnelly, Henry Harpending, Lynn Jorde, and two anonymous reviewers for comments. This work was supported in part by a grant from NSF (DBS–9310105).

References

[1] Giorgio Bertorelle and Montgomery Slatkin. The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. manuscript, 1994.

[2] Rebecca L. Cann, Mark Stoneking, and Allan C. Wilson. Mitochondrial DNA and human evolution. *Nature*, 325(1):31–36, January 1987.

[3] Anna Di Rienzo and Alan C. Wilson. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proceedings of the National Academy of Sciences, USA*, 88:1597–1601, 1991.

[4] David W. Frayer, Milford H. Wolpoff, Alan G. Thorne, Fred H. Smith, and Geoffrey G. Pope. Theories of modern human origins: The paleontological test. *American Anthropologist*, 95(1):14–50, 1993.

[5] Henry Harpending. Signature of ancient population growth in a low resolution mitochondrial DNA mismatch distribution. *Human Biology*, 66(4):591–600, 1994.

[6] Henry C. Harpending, Stephen T. Sherry, Alan R. Rogers, and Mark Stoneking. The genetic structure of ancient human populations. *Current Anthropology*, 34:483–496, 1993.

[7] Daniel L Hartl and Andrew G. Clark. *Principles of Population Genetics*. Sinauer, Sunderland, MA, 2nd edition, 1989.

[8] Masami Hasegawa, Anna Di Rienzo, and Alan C. Wilson. Toward a more accurate time scale for the human mitochondrial DNA tree. *Journal of Molecular Evolution*, 37:347–354, 1993.

[9] Richard R. Hudson. Gene genealogies and the coalescent process. In Douglas Futuyma and Janis Antonovics, editors, *Oxford Surveys in Evolutionary Biology*, volume 7, pages 1–44. Oxford University Press, Oxford, 1990.

[10] Albert Jacquard. *The Genetic Structure of Populations*. Springer-Verlag, New York, 1974.

[11] M. Kendall and A. Stuart. *The Advanced Theory of Statistics. II. Inference and Relationship*. MacMillan, New York, fourth edition, 1979.

[12] Motoo Kimura. Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology*, 2:174–208, 1971.

[13] Thomas Kocher and Allan Wilson. Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. In S. Osawa and T. Honjo, editors, *Evolution of Life: Fossils, Molecules, and Culture*, pages 391–413. Springer-Verlag, New York, 1991.

[14] Wen-Hsiung Li. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics*, 85:331–337, 1977.

[15] Paul Marjoram and Peter Donnelly. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics*, 136:673–683, February 1994.

[16] Masatoshi Nei and Nayouki Takahata. Effective population size, genetic diversity, and coalescence time in subdivided populations. *Journal of Molecular Evolution*, 37:240–244, 1993.

[17] Alan R. Rogers. Error introduced by the infinite sites model. *Molecular Biology and Evolution*, 9:1181–1184, 1992.

[18] Alan R. Rogers. Genetic evidence for a Pleistocene population explosion. *Evolution*, 1995. In press.

[19] Alan R. Rogers and Henry C. Harpending. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, 9:552–569, 1992.

[20] Alan R. Rogers and Lynn B. Jorde. Genetic evidence on modern human origins. *Human Biology*, 67(1):1–36, February 1995. In press.

[21] Stephen Sherry, Alan R. Rogers, Henry C. Harpending, Himla Soodyall, Trefor Jenkins, and Mark Stoneking. Mismatch distributions of mtDNA reveal recent human population expansions. *Human Biology*, 66(5):761–775, Oct 1994.

[22] Montgomery Slatkin and Richard R. Hudson. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129:555–562, 1991.

[23] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7:256–276, 1975.

[24] Milford H. Wolpoff. Multiregional evolution: The fossil alternative to Eden. In Paul Mellars and Chris Stringer, editors, *The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans*, pages 62–108. Princeton University Press, Princeton, New Jersey, 1989.