Copyright © 1998 by the Genetics Society of America

Inference of Population History Using a Likelihood Approach

Gunter Weiss and Arndt von Haeseler

Institute of Zoology, University of Munich, D-80333 Munich, Germany

Manuscript received July 28, 1997
Accepted for publication March 23, 1998

ABSTRACT

We introduce an approach to revealing the likelihood of different population histories that utilizes an explicit model of sequence evolution for the DNA segment under study. Based on a phylogenetic tree reconstruction method we show that a Tamura-Nei model with heterogeneous mutation rates is a fair description of the evolutionary process of the hypervariable region *I* of the mitochondrial DNA from humans. Assuming this complex model still allows the estimation of population history parameters, we suggest a likelihood approach to conducting statistical inference within a class of expansion models. More precisely, the likelihood of the data is based on the mean pairwise differences between DNA sequences and the number of variable sites in a sample. The use of likelihood ratios enables comparison of different hypotheses about population history, such as constant population size during the past or an increase or decrease of population size starting at some point back in time. This method was applied to show that the population of the Basques has expanded, whereas that of the Biaka pygmies is most likely decreasing. The Nuu-Chah-Nulth data are consistent with a model of constant population.

IN the past decade much effort has been put into collecting and sequencing DNA from human populations from all over the world (Cavalli-Sforza *et al.* 1994; Kogelnik *et al.* 1997; Handt *et al.* 1998). While much of the interest in these problems comes from medical genetics, population geneticists also address questions about the origin of our species, its spread over the world, and relationships between contemporary populations. The noncoding hypervariable region *I* (HVR*I*) of mitochondrial DNA (mtDNA) is especially suited for this intraspecies analysis. This is due to the maternal inheritance of mtDNA, the absence of recombination, and the high mutation rate of this part of the genome. By now more than 4000 individual HVR*I* sequences have been published (Handt *et al.* 1998).

There are different approaches to analyzing this sort of data: Networks are a suggestive way to visualize the relationship of sequences in a sample (Bandelt *et al.* 1995). Tree reconstruction methods can give insights into the mode of evolution of the particular genomic region studied (Tamura and Nei 1993). If one intends to extract information about historical population events, coalescent theory provides a framework with which to incorporate parametric models of population history explicitly (Donnelly and Tavaré 1995). Coalescent models describe the evolution of a sample of DNA sequences in terms of stochastic processes. Thus, they allow application of statistical techniques for parameter estimation and model testing.

Here we study a class of expansion models that include three parameters: the initial population size, the point back in time when exponential growth (positive or negative) began, and the ratio of current to initial population size. Other authors (Rogers and Harpending 1992; Griffiths and Tavaré 1994b; Wakeley and Hey 1997) estimated the parameters of a so-called sudden expansion model assuming an infinitely many sites model (Watterson 1975).

We argue that a Tamura-Nei model with Γ-distributed rates is more appropriate for describing the evolution of HVR*I* sequences (Tamura and Nei 1993). Under this complex mutation model it is difficult to derive analytical formulae and infer population history parameters by a method of moments approach. Therefore, we make use of a likelihood ratio statistic that is based on two summary statistics, the mean pairwise sequence difference and the number of variable positions in a sample of DNA sequences. More precisely, for a whole range of population histories we compute the corresponding likelihoods given the data. A history is considered implausible if its likelihood is much smaller than that of the most probable population history. Since we are not aware of any analytical expression for the likelihood function under this complex model, we will resort to Monte Carlo simulations. We demonstrate the applicability of our approach by analyzing the demographic history of three human populations, namely the Basques, the Nuu-Chah-Nulth, and the Biaka pygmies.

THEORETICAL BACKGROUND

The coalescent (Kingman 1982a,b,c) is an efficient and elegant way to describe the ancestral relationship

Address for correspondence: Arndt von Haeseler, Institute of Zoology, University of Munich, Luisenstrasse 14, D-80333 Munich, Germany. E-mail: arndt@zi.biologie.uni-muenchen.de

of a sample of homologous DNA sequences in terms of probabilities. While some coalescent models are analytically tractable, a major feature of the coalescent is that it gives an exact framework to base the simulation on (Hudson 1991; Donnelly and Tavaré 1995). To generate a set of simulated DNA sequences we have to model two processes: First, the reproduction process (constant size or expansion model, panmictic or substructured population, etc.) determines shape and branch lengths of a sample genealogy; second, the model of the mutation process (infinitely or finitely many sites model, uniform or site-dependent mutation rate) specifies the way mutations are distributed along the sequences in the genealogy. The impact of the various assumptions of different reproduction models on genetic diversity measures, such as pairwise difference distributions, mean pairwise difference, or number of variable positions in a sample, was investigated by several authors (Tajima 1989b; Slatkin and Hudson 1991; Rogers and Harpending 1992; Marjoram and Donnelly 1993). Most of this theoretical work and therefore most of the data analysis has been done under the infinitely many sites model assumption (Watterson 1975). However, most HVR*I* sequence samples from different populations are not consistent with this model, which does not allow for any recurrent or parallel mutation. It is argued that this mutation process model serves as a useful approximation (Rogers 1992). Others claim that their method is "insensitive to the mutational process" (Rogers *et al.* 1996). On the other hand, reports that describe the substantial effects of different mutation processes on various genetic diversity measures abound (Lundstrom *et al.* 1992a,b; Bertorelle and Slatkin 1995; Aris-Brosou and Excoffier 1996; Tajima 1996). In agreement with the latter results we put forward that the specification of the model of the mutation process is crucial for the inference of the parameters of a population history. Consequently, we study the mutation process of a segment of the HVR*I* of human mtDNA.

The mutation process of HVR*I* sequences: Since DNA sequence evolution generally does not adhere to the infinitely many sites model, it is essential to investigate the process of sequence evolution in greater detail. We suggest decoupling the estimation of mutation parameters from the analysis of a specific population sample. This approach requires a reasonably large number of sequences from a widespread distribution of different populations. To model the mutation process we assume that evolution at a given site follows a time-continuous, time-homogeneous Markov chain. The most general model we allow for is the so-called Tamura-Nei model (1993). This model is most conveniently summarized in the rate matrix R (Scheme 1), where the omitted diagonal elements are set such that the elements in each row add up to zero. The entry $r_{i,j}$ (i,j = A,C,G,T, $i \neq j$)

$$R = \begin{pmatrix} — & \pi_C & 2\kappa\frac{2}{\xi+1}\pi_G & \pi_T \\ \pi_A & — & \pi_G & 2\kappa\frac{2\xi}{\xi+1}\pi_T \\ 2\kappa\frac{2}{\xi+1}\pi_A & \pi_C & — & \pi_T \\ \pi_A & 2\kappa\frac{2\xi}{\xi+1}\pi_C & \pi_G & — \end{pmatrix},$$

Scheme 1

describes the substitution rate of nucleotide i to nucleotide j. The model comprises parameters for the base frequencies π_A, π_C, π_G, π_T, a transition/transversion parameter κ, and a pyrimidine/purine transition parameter ξ. Note that other models are special cases of the Tamura-Nei model. If $\xi = 1$, the Hasegawa Kishino Yano (HKY) model (Hasegawa *et al.* 1985) is defined. If we, moreover, assume homogeneous base frequencies, the HKY model reduces to Kimura's two-parameter model (Kimura 1980). Finally, letting $\kappa = 0.5$, the Jukes-Cantor model is obtained (Jukes and Cantor 1969). To take rate heterogeneity into account, we assume that each site evolves according to the same rate matrix R, but that the actual rate of a given site follows a γ-distribution

$$f(r) = \frac{\alpha^\alpha}{\Gamma(\alpha)} e^{-r\alpha} r^{\alpha-1},$$

with mean 1 and variance $1/\alpha$ (Uzzel and Corbin 1971; Wakeley 1993).

For a specific model the corresponding mutation parameters are estimated by employing a phylogenetic approach. That is, a random sample of lineages is drawn from the collection of sequences. Subsequently, a maximum likelihood tree using the *PUZZLE* program is reconstructed and the mutation parameters are estimated from the tree (Strimmer and von Haeseler 1996). To ensure that the estimates are not affected by the sample and by the population dynamics, the entire procedure is repeated several times for different random samples of lineages. This phylogenetic approach produces parameter estimates of the mutation process that are virtually independent of the demographic history.

However, as outlined above, we still have the choice between a variety of complex models. To find out the most appropriate model of evolution for HVR*I* sequences, we employed Goldman's suggestion (1993) of a likelihood ratio test (*e.g.*, Cox 1961, 1962). That is, for a sample of lineages a maximum likelihood *PUZZLE* tree is estimated assuming a HKY model of sequence evolution, say. Let L_0 be the corresponding maximum likelihood value. Then, assuming a Tamura-Nei model the maximum likelihood tree with likelihood L_1 is estimated. Unfortunately, the usual χ^2 approximation for the distribution of $\Delta = -2\ln(L_0/L_1)$ under H_0 cannot be used in the phylogenetic context (Goldman

1993). Therefore, Δ is compared to the empirical distribution, which is obtained as follows: Sequences were generated under hypothesis H_0 (*e.g.*, the estimated ML tree under HKY), using the program *Seq-Gen* (Rambaut and Grassly 1997). For each simulated data set L_0/, L_1 and Δ are computed as described for the true data. H_0 are computed as described for the true data. H_0 is rejected if the observed Δ-value of the data falls in the upper 5% tail of the empirical distribution. Usually, 100 sets of sequences are generated. If H_0 is rejected, it is assumed that the parameters of the evolutionary model from H_1 are more appropriate for describing sequence evolution.

Starting with the most simple model (*e.g.*, Jukes-Cantor 1969) one can gradually increase the complexity of the models until H_0 is no longer rejected or until we cannot increase the complexity of the model (in our case the Tamura-Nei model with γ-distributed rates is the most complex model).

The entire procedure was applied to estimate the best evolutionary model for the 360-bp region [corresponding to positions 16024 through 16383 (Anderson *et al.* 1981)] of HVR*I* from human mtDNA, from which 2298 different lineages are collected (Handt *et al.* 1998). In the HVR*I* region a pronounced transition/transversion ratio as well as rate heterogeneity is observed (Kocher and Wilson 1991; Hasegawa *et al.* 1993; Wakeley 1993). Thus, the infinitely many sites model is violated.

To pick the most appropriate model of sequence evolution, 10×25 random lineages were samples from the collection (Handt *et al.* 1998), and for each sample the test procedure was performed. For all 10 repeats the Tamura-Nei model with γ-distributed rates was the best model as indicated by the likelihood of the tree. All simpler models were rejected on the 5% level, except for the HKY model with γ-distributed rates, which was rejected only in 1 out of the 10 cases. Since the Tamura-Nei model with γ-distributed rates always provided a higher likelihood, this model is assumed in the following. Finally, the model parameters (*cf.* Scheme 1) were estimated from 50 sets of 50 lineages that were randomly chosen from the HVR*I* sequences collection (Handt *et al.* 1998). For each set, parameters were estimated from the maximum likelihood *PUZZLE* tree. Table 1 shows the results. Our estimates are in good agreement with those from similar studies, even though direct comparisons are complicated by the different regions studied. For example, we obtained one α of 0.26 and other estimates ranging from 0.11 to 0.47 (Tamura and Nei 1993; Wakeley 1993). Similarly, we inferred a value of 15.5 for the transition/transversion ratio κ, which corresponds with other observations (Horai and Hayasaka 1990; Tamura and Nei 1993). For the following analysis of population genetics data we used the Tamura-Nei model with γ-distributed rates and the parameter values from column three of Table 1.

TABLE 1

Estimated parameters of the Tamura-Nei model with Γ-distributed rates for HVR*I* sequences of human mtDNA

Parameter	Identifier	Mean	Standard deviation
Base frequency of A	π_A	32.7%	—
Base frequency of C	π_C	33.8%	—
Base frequency of G	π_G	11.4%	—
Base frequency of T	π_T	22.1%	—
Transition/transversion	κ	15.5	5.0
Pyrimidine/purine trans.	ξ	1.85	0.48
Rate heterogeneity	α	0.26	0.05

A class of population history models: The basic model of population history is the Wright-Fisher model, assuming a panmictic population of constant size with nonoverlapping generations (Fisher 1930; Wright 1931). Several other models, including population structure (with or without migration) or models with variable population sizes, have been proposed (Hudson 1991; Slatkin and Hudson 1991; Rogers and Harpending 1992).

We will investigate a class of expansion model where a Wright-Fisher population at equilibrium started to grow or decrease exponentially at a certain time τ in the past to the current population size. This model is defined by three parameters: $\theta = 2N_0\mu$ is the population parameter at equilibrium in the past. Here, N_0 denotes the initial effective population size and μ is the mutation rate per sequence and generation. τ is the time when the population size started to change, where τ is measured in units of $1/\mu$. Finally, ρ defines the ratio of current to initial population size. We get the basic model of constant size as a special case of the expansion model by setting ρ equal to one. Then the time parameter τ remains unspecified. Figure 1 sketches the three scenarios.

Inference of population history parameters: In the following, we assume that the mutation process is known. Thus, the evolution of a sample of sequences is fully characterized by the three parameters θ, τ, and ρ. To estimate these three parameters a likelihood method is reasonable. Instead of conditioning the likelihood of the parameter set on the full sequence data as the elegant approach of Griffiths and Tavaré (1994a,b) does, we summarize the data by the mean pairwise difference K and the number of variable positions S of the sample. While we do not know yet to what extent this restriction affects the estimates of (θ,τ,ρ), there are several reasons to proceed in that manner: First, we are more interested in detecting major demographic signals in the data than in giving precise joint estimates (together with the inherent large variances of these estimates). Second, by conditioning on both K and S, we make implicit use of the relationship between the mean pairwise difference and the number of variable posi-

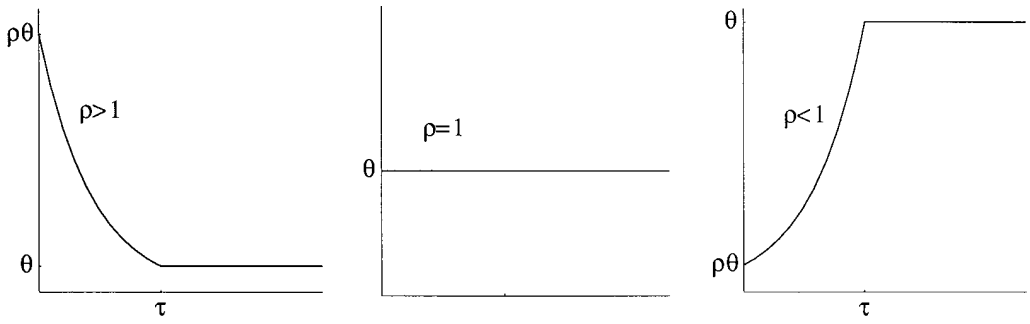

Figure 1.—Three types of population histories included in the class of models considered. The left part shows an exponential increase in population size that started at time τ in the past; the right part shows an exponential decrease. In the middle a model of constant population size is indicated. The *x*-axis represents time from present backward and the *y*-axis scaled population size.

tions, which proved useful in Tajima's *D*-statistic (1989a). Assuming the infinitely many sites model Tajima (1989b) showed that changes in population size have different effects on *K* and *S*. Third, the reduction of the data simplifies and speeds up our method of evaluating the likelihood via simulations.

The set of parameters that maximizes the likelihood $\mathrm{lik}(\theta,\tau,\rho|k,s)$ defines the most probable population history within the class of models described in the previous paragraph. To rate the plausibility of a parameter set (θ_0,τ_0,ρ_0) we use the likelihood ratio $\mathrm{lik}(\theta_0,\tau_0,\rho_0|k,s)/L_A$, where L_A is the maximum likelihood value within the considered class of population histories.

The remaining question is how to compute the likelihoods for a given data set. To the best of our knowledge analytical formulae for the likelihood functions based on a complex model of sequence evolution are not available, even in the case of a population of constant size. Therefore, we will use computer simulations to determine the likelihood value of the parameter sets.

Simulations: Coalescent theory provides an efficient way to simulate the evolution of a sample of sequences under various population histories. The coalescent is a stochastic process that counts the number of distinct ancestors in a genealogy of a sample of size *n* as one moves back in time. If the current population size *N* is large and time is measured in *N* generations, the coalescent provides a good approximation to the ancestral process of the sample (Kingman 1982b,c; Tavaré 1984; Donnelly and Tavaré 1995). The coalescent starts with initial state *n*. Each coalescent event of two ancestors in the sample reduces the number of distinct ancestors by exactly one; *i.e.*, the state value decreases by one. After $n-1$ coalescent events the process ends in state one and the most recent common ancestor of the sample is found. The continuous time approximation of the coalescent ensures that the jump size of the process is exactly one; *i.e.*, only one coalescent even can occur at one time. The times S_j, $j = n, \ldots, 2$, when the process changes state (coalescent times) are random quantities that depend on the population history. If the sampled sequences stem from a Wright-Fisher population of constant size, the times $T_j = S_j - S_{j+1}$ ($S_{n+1} = 0$) follow an exponential distribution with mean $2/j(j-1)$, $j = n, \ldots, 2$.

If the population size varies deterministically, the coalescent approximation still holds, but the coalescent times must be rescaled appropriately (Griffiths and Tavaré 1994b; Donnelly and Tavaré 1995): Let $Nv(x)$ denote the population size at time x in the past. The relative size function $v(x)$ for the class of models under consideration (Figure 1) is given by

$$v(x) = \begin{cases} 1, & \text{if } \rho = 1 \\ \rho^{-x/\tau}, & \text{if } \rho \neq 1 \text{ and } x < \tau \\ \rho^{-1}, & \text{if } \rho \neq 1 \text{ and } x \geq \tau. \end{cases} \tag{1}$$

Furthermore, we define the population size intensity function Λ:

$$\Lambda(t) \equiv \int_0^t \frac{1}{v(x)}dx \tag{2}$$

$$= \begin{cases} t, & \text{if } \rho = 1 \\ \dfrac{\tau}{\ln \rho}(\rho^{t/\tau} - 1), & \text{if } \rho \neq 1 \text{ and } t < \tau \\ \dfrac{\tau}{\ln \rho}(\rho - 1) + \rho(t - \tau), & \text{if } \rho \neq 1 \text{ and } t \geq \tau. \end{cases}$$

When population size varies, the sequence S_n, S_{n-1}, $\ldots$, S_2 is still Markovian. Although the times T_j are not independent anymore, it is straightforward to simulate the coalescent times (Griffiths and Tavaré 1994b):

For $j = n, \ldots, 2$ the times S_j are determined by recursively solving the equation

$$\Lambda(S_j) - \Lambda(S_{j+1}) = E_j, \tag{3}$$

where S_{n+1} is set equal to zero and E_j is an exponential random variable with mean $2/j(j-1)$. When the $n-1$ coalescent times are determined, the genealogy of the

sample is produced by randomly merging two ancestral lineages at each coalescent time.

Sequences according to the genealogy are then produced as follows: We generate an ancestral sequence at the root of the genealogy according to the stationary base composition in Table 1. This sequence then evolves along the genealogy under the Tamura-Nei model with rate heterogeneity. Thus, one simulated data set is produced.

The approximate likelihood value $\text{lik}(\theta,\tau,\rho|k,s)$ for a real data set with mean pairwise difference k and s variable positions is based on $j = 1, \ldots, B$ simulations according to the specified parameters (θ, τ, ρ). For a simulated data set j we computed the mean pairwise difference k_j and the number of variable positions s_j. Because the mean pairwise difference is virtually a continuous variable, we introduce the indicator variable

$$I_\delta(j) = \begin{cases} 1, & \text{if } |k_j - k| \le \delta \text{ and } s_j = s \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where δ is a small positive number that defines an interval with k as the center. Thus, the indicator variable equals one for simulations that are reasonably close to our summary statistics observed in the data. The choice of δ is a compromise between the efficiency of the simulations and the precision of the approximation. In our investigation $\delta = 0.2$ proved to be useful when $B = 25{,}000$. The likelihood value of a parameter set is then approximated by

$$\text{lik}(\theta,\tau,\rho|k,s) \approx \frac{1}{B}\sum_{j=1}^{B} I_\delta(j). \quad (5)$$

We determine the likelihood values on a grid of parameter combinations to approximate the maximum likelihood value under the most general hypothesis and the specific subset of parameters defined by the null hypothesis.

APPLICATION TO SEQUENCE DATA

We demonstrate the features of our method by analyzing published data sets from three populations: The Basques are linguistic isolates living on both sides of the border between France and Spain (Bertranpetit *et al.* 1995); the Nuu-Chah-Nulth (NCN) are a native-American tribe peopling the western coast of Vancouver Island (Canada) and the Olympic Peninsula of Washington state (USA) (Ward *et al.* 1991); and the Biaka pygmies live as hunter-gatherers in the Central African Republic (Vigilant *et al.* 1991). Sample sizes, n, mean pairwise sequence differences, k, and numbers of variables positions, s, of these three data sets are given in Table 2.

TABLE 2

Sample sizes, mean pairwise sequence differences, and numbers of variable positions of the analyzed data sets

Population	n[a]	k[b]	s[c]
Basques	45	3.24	32
Nuu-Chah-Nulth	63	5.32	26
Biaka pygmies	17	8.12	21

[a] Sample sizes.
[b] Sequence differences.
[c] Numbers of variable positions.

Assuming the model of evolution as specified in Table 1, we determined for each population the likelihood values for various parameter combinations according to Equation 5. The number of simulations B was equal to 25,000 and $\delta = 0.2$. Thereby, ρ was set equal to 10^z, z being an integer. In this notation $z = 0$ reflects the case of constant population size over time. If $z > 0$ the population size is increasing, and if $z < 0$, it is decreasing. The likelihood values determined for τ were a multiple of one-fourth and θ a multiple of one-half.

The parameter set that yields the highest likelihood (L_A^{pop}) is regarded as the most probable population history. Since this point estimate does not reflect the uncertainty inherent in the stochastic models, we rate each parameter set (θ_0,τ_0,ρ_0) using the likelihood ratio $\text{lik}(\theta_0,\tau_0,\rho_0|k,s) / L_A^{\text{pop}}$. Figures 2–4 show graphical representation of these results. Each panel refers to the specified value of ρ. The abscissa and ordinate represent the parameters τ and θ, respectively. (Note that if $\rho = 1$, τ is not defined and the panel is only one dimensional.) The different gray levels reflect the likelihood ratio of the corresponding parameter sets. Dark colors represent high values of the likelihood ratio (see the legend in Figure 2). If the usual χ^2 theory applies, then plain white color would indicate parameter combinations that are rejected on the 5% level. Conversely, colored boxes belong to a 95% confidence set.

Figure 2 represents the analysis of the Basques. Parameters sets that favor a recent expansion (small τ) from a small population (small θ) receive high support from the data. Models of population decrease or constant population size get virtually no support. The corresponding panels are plain white and therefore omitted from Figure 2. Thus, we conclude that this data set is consistent with a model exponential growth. The analysis further suggests that the Basques stem from a rather small founding population (as measured by mitochondrial variability) that started to increase in size one to four mutational time units ago.

The picture of the Nuu-Chah-Nulth data set (Figure 3) shows a different result. Neither models of population decrease nor increase were rejected. The most probable parameter set is that of constant size and $\theta = 7.5$. Figure 3 indicates also high support for a very recent expansion ($\tau \le 0.5$). However, this result suggests that

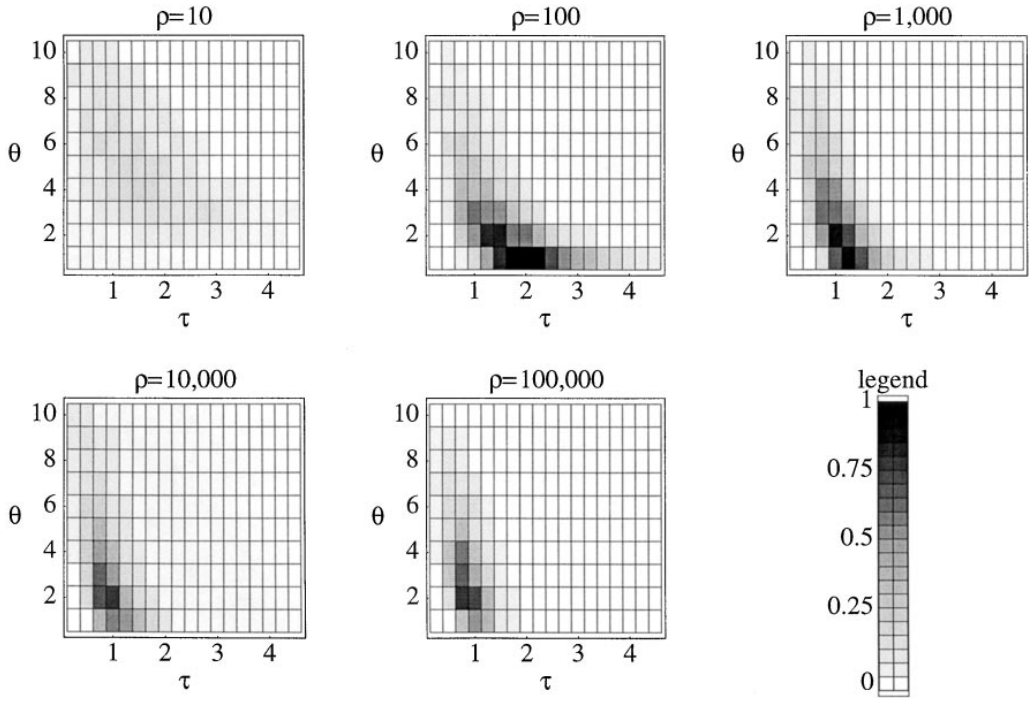

Figure 2.—Result of the likelihood analysis of the Basque data (Bertran-Petit *et al.* 1995). Dark color represent high values of the likelihood ratio (see legend). The parameter combination $\rho = 100$, $\tau = 2.25$, and $\theta = 1$ yielded the highest likelihood ratio. If the usual χ^2 theory applies, colored boxes belong to a 95% confidence set.

further analysis of this data set is reasonable under the constant size assumption.

The analysis of the Biaka pygmies reveals an interesting result (Figure 4). Models that assume a moderate decrease in population size ($\rho = 0.1$) obtain a high support. Also, the long-term constant size model is conceivable. The decrease could be the result of an expansion of other populations into the habitat of the Biaka pygmies, pushing them back to more restricted areas.

DISCUSSION

Effects of variable population size on genetic diversity measures are confounded with the effects of rate variation in the sequences (Tajima 1989b; Slatkin and Hudson 1991; Lundstrom *et al.* 1992a; Rogers and Harpending 1992; Marjoram and Donnelly 1993; Bertorelle and Slatkin 1995; Aris-Brosou and Excoffier 1996; Tajima 1996). Here, we try to resolve the interaction by studying first the appropriate model of sequence evolution and subsequently employing this model in a coalescent framework.

The large amount of available HVR*I* sequence data (Handt *et al.* 1998) enabled us to infer this model of sequence evolution for this particular region of the human genome. This was done independently from population dynamics by using a likelihood-based tree reconstruction method (Strimmer and von Haeseler 1996). Fixing the model of sequence evolution and thereby separating the dynamics of the mutation process from the dynamics of population history makes inference of population history parameters possible. Note that we do not include the uncertainty inherent in estimating the mutation process parameters. This may lead to underestimation of the variability in the estimates of population history parameters.

We studied a class of population histories that allows for positive or negative growth of a population starting from a population at equilibrium in the past. Even though these models are more general than the frequently used constant-size assumption, it is obvious that they are simplistic and do not describe the evolution of a population in its full complexity. Nevertheless, the application section shows that this approach detects signals of major changes in population size. Moreover, we rated demographic scenarios by their likelihood ratio. This yields a set of plausible parameter combinations consistent with the data. On the basis of these parameters, further questions, such as the estimation of the time to the most recent common ancestor, can be addressed (Tajima 1983; Griffiths and Tavaré 1994c; Fu 1996b; Tavaré *et al.* 1997).

The proposed likelihood approach makes implicit use of the relationship of the mean pairwise difference and the number of variable positions. Therefore, it is far more applicable than "mismatch analysis" (Rogers 1995; Rogers *et al.* 1996), which is restricted to data sets that show smooth and unimodal pairwise difference distributions. While the restriction of the data to two summary statistics has its advantage in speeding up the evaluation of likelihood values via simulations and concurrently retaining information about population history, the drawback of the data summary is a loss in the precision of parameter estimates. We note that inclusion of additional information, like patterns of segregating sites (Fu 1996a; Wakeley and Hey 1997), should result

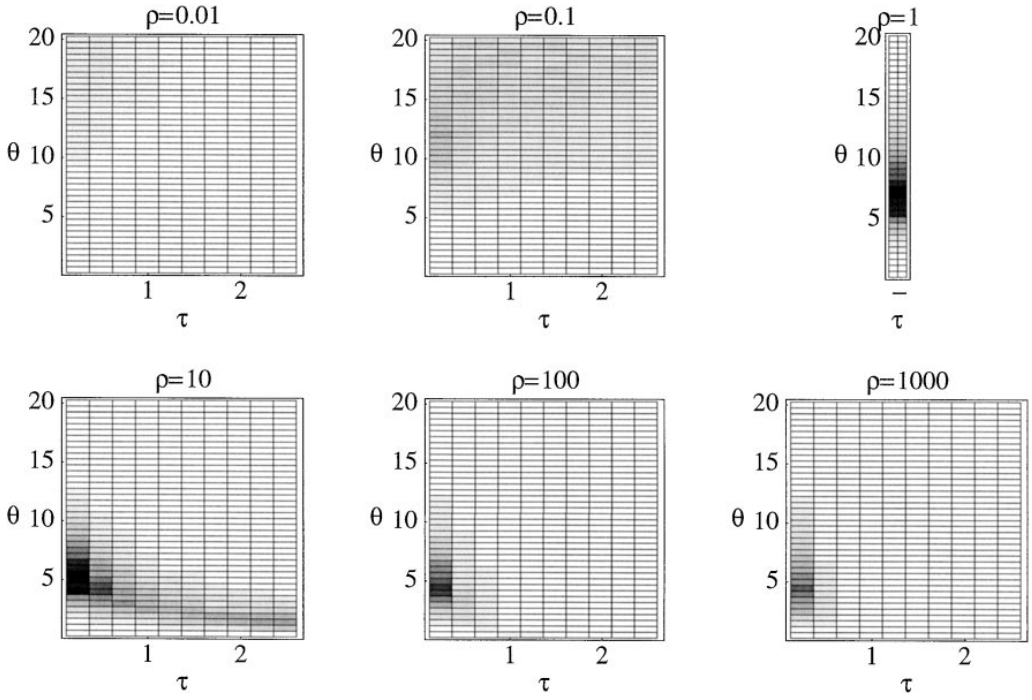

Figure 3.—Result of the likelihood analysis of the Nuu-Chah-Nulth data (Ward *et al.* 1991). The parameter combination $\rho = 1$, $\theta = 7.5$ yielded the highest likelihood ratio. For further description see legend of Figure 2.

in sharper estimates. Conditioning the likelihood on the entire information in the data as proposed by Griffiths and Tavaré (1994a,b) would reduce the variability of the estimates even further. However, the Markov Chain, Monte Carlo-based calculation of the likelihood is computationally intensive when applied to our complex model of sequence evolution.

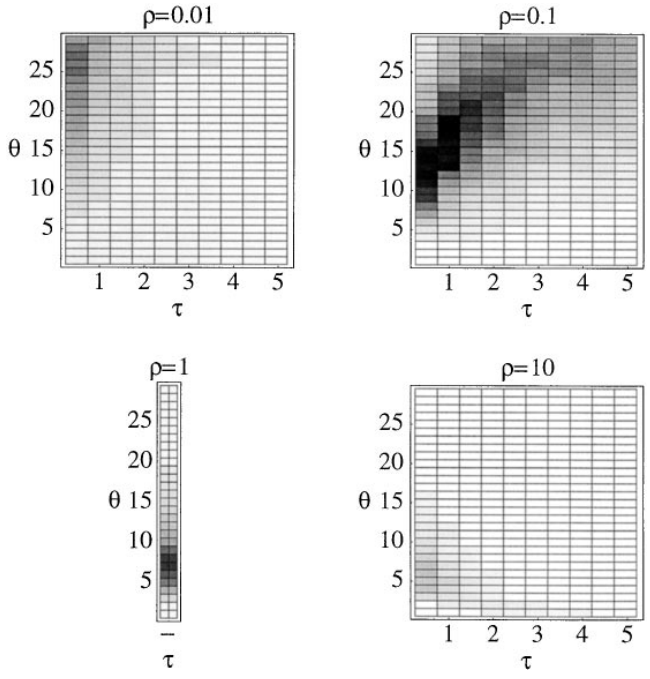

Figure 4.—Result of the likelihood analysis of the Biaka pygmie data (Vigilant *et al.* 1991). The parameter combination $\rho = 0.1$, $\tau = 1$, and $\theta = 18$ yielded the highest likelihood ratio. For further description see the legend of Figure 2.

In principle, the likelihood ratio approach is applicable to other classes of population history. The extension to demographic scenarios that are different from the exponential growth model is straightforward. This could be done by appropriately altering the rescaling of the coalescent times (Equations 1 and 2). Also, simple subpopulation models with migration (Wakeley and Hey 1997) can be explored by conditioning on suitable summary statistics.

With the prospective large amount of available DNA sequence data, the refinement of our understanding of the evolution of different parts of our genome, and the development of applicative methods of analysis, we will be able to improve our knowledge of the history of our species in the near future.

The authors are grateful to Korbinian Strimmer for helpful comments on the use of *PUZZLE* and Sonja Meyer for providing the HVR*I* data collection and discussing the estimation of mutation parameters. We thank Ellen Baake and Svante Pääbo for fruitful discussions and improving the manuscript. Critical comments from two anonymous referees and the editor are also gratefully acknowledged. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DGF) to A.v.H. The simulation program described here is available upon request (arndt@zi.biologie.uni-muenchen.de, gweiss@zi.biologie.uni-muenchen.de).

LITERATURE CITED

Anderson, S., A. T. Bankier, B. G. Barell, M. H. L. de Bruijn, A. R. Coulson *et al.*, 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.

Bandelt, H. J., P. Forster, B. C. Sykes and M. B. Richards, 1995 Mitochondrial portraits of human populations using median networks. Genetics **141:** 743–753.

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.

Bertranpetit, J., J. Sala, F. Calafell, P. A. Underhill, P. Moral, *et al.*, 1995 Human mitochondrial DNA variation and the origin of Basques. Am. J. Hum. Genet. **59:** 63–81.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1994 *The History and Geography of Human Genes.* Princeton University Press, Princeton.

Cox, D. R., 1961 Tests of separate families of hypotheses. Proc. 4th Berkeley Symp. (Univ. California Press) **1:** 105–123.

Cox, D. R., 1962 Further results on tests of separate families of hypotheses. J. R. Stat. Soc. **B 24:** 406–424.

Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon, Oxford.

Fu, Y.-X., 1996a New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

Fu, Y.-X., 1996b Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. Genetics **144:** 829–838.

Goldman, N., 1993 Statistical tests of models of DNA substitution. J. Mol. Evol. **36:** 182–198.

Griffiths, R. C., and S. Tavaré, 1994a Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Griffiths, R. C., and S. Tavaré, 1994b Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B **344:** 403–410.

Griffiths, R. C., and S. Tavaré, 1994c Ancestral inference in population genetics. Stat. Sci. **9:** 307–319.

Handt, O., S. Meyer and A. von Haeseler, 1998 Compilation of human mtDNA control region sequences. Nucleic Acids Res. **26:** 126–130.

Hasegawa, M., H. Kishino and K. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Hasegawa, M., A. Di Rienzo, T. D. Kocher and A. C. Wilson, 1993 Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. **37:** 347–354.

Horai, S., and K. Hayasaka, 1990 Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am. J. Hum. Genet. **46:** 828–842.

Hudson, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyama and J. Antonovics. Oxford University Press, Oxford.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kimura, M., 1980 A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Applns. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kingman, J. F. C., 1982c Exchangeability and the evolution of large populations, pp. 97–112 in *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino. North-Holland Publishing Company, Amsterdam.

Kocher, T. D., and A. C. Wilson, 1991 Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and protein coding regions, pp. 391–413 in *Evolution of Life: Fossils, Molecules and Culture*, edited by S. Osawa and T. Honio. Springer Verlag, Tokyo.

Kogelnik, A. M., M. T. Lott, M. D. Brown, S. B. Navathe and D. C. Wallace, 1997 MITOMAP: an update on the status of the human mitochondrial genome database. Nucleic Acids Res. **25:** 196–199.

Lundstrom, R., S. Tavaré and R. H. Ward, 1992a Modelling the evolution of the human mitochondrial genome. Math. Biosci. **112:** 319–335.

Lundstrom, R., S. Tavaré and R. H. Ward, 1992b Estimating substitution rates from molecular data using the coalescent. Proc. Natl. Acad. Sci. **89:** 5961–5965.

Marjoram, P., and P. Donnelly, 1993 Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136:** 673–683.

Rambaut, A., and N. C. Grassly, 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13:** 235–238.

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

Rogers, A. R., 1995 Genetic evidence for a pleistocene population explosion. Evolution **49:** 608–615.

Rogers, A. R., A. E. Fraley, M. J. Bamshad, W. S. Watkins and L. B. Jorde, 1996 Mitochondrial mismatch analysis is insensitive to the mutational process. Mol. Biol. Evol. **13:** 895–902.

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

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

Strimmer, K., and A. von Haeseler, 1996 Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13:** 964–969.

Tajima, F., 1983 Evolutionary relationships of DNA sequences in finite populations. Genetics **105:** 437–460.

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

Tajima, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

Tajima, F., 1996 The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics **143:** 1457–1465.

Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Pop. Biol. **26:** 119–164.

Tavaré, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

Uzzel, T., and K. W. Corbin, 1971 Fitting discrete probability distributions to evolutionary events. Science **172:** 1089–1096.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson, 1991 African populations and the evolution of mitochondrial DNA. Science **253:** 1503–1507.

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

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

Ward, R. H., B. L. Frazier, K. Dew-Jager and S. Pääbo, 1991 Extensive mitochondrial diversity within a single Amerindian tribe. Nat. Acad. Sci. **88:** 8720–8724.

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

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

Communicating editor: S. Tavaré

View publication stats