Intra-Deme Molecular Diversity in Spatially Expanding Populations

Nicolas Ray,† Mathias Currat,*† and Laurent Excoffier†*

*Genetics and Biometry Lab, Department of Anthropology and Ecology, University of Geneva, Geneva, Switzerland; and †Computational and Molecular Population Genetics Lab, Zoological Institute, University of Bern, Bern, Switzerland

We report here a simulation study examining the effect of a recent spatial expansion on the pattern of molecular diversity within a deme. We first simulate a range expansion in a virtual world consisting in a two-dimensional array of demes exchanging a given proportion of migrants (m) with their neighbors. The recorded demographic and migration histories are then used under a coalescent approach to generate the genetic diversity in a sample of genes. We find that the shape of the gene genealogies and the overall pattern of diversity within demes depend not only on the age of the expansion but also on the level of gene flow between neighboring demes, as measured by the product Nm, where N is the size of a deme. For small Nm values ($<$ approximately 20 migrants sent outwards per generation), a substantial proportion of coalescent events occur early in the genealogy, whereas with larger levels of gene flow, most coalescent events occur around the time of the onset of the spatial expansion. Gene genealogies are star shaped, and mismatch distributions are unimodal after a range expansion for large Nm values. In contrast, gene genealogies present a mixture of both very short and very long branch lengths, and mismatch distributions are multimodal for small Nm values. It follows that statistics used in tests of selective neutrality like Tajima's D statistic or Fu's F_S statistic will show very significant negative values after a spatial expansion only in demes with high Nm values. In the context of human evolution, this difference could explain very simply the fact that analyses of samples of mitochondrial DNA sequences reveal multimodal mismatch distributions in hunter-gatherers and unimodal distributions in post-Neolithic populations. Indeed, the current simulations show that a recent increase in deme size (resulting in a larger Nm value) is sufficient to prevent recent coalescent events and thus lead to unimodal mismatch distributions, even if deme sizes (and therefore Nm values) were previously much smaller. The fact that molecular diversity within deme is so dependent on recent levels of gene flow suggests that it should be possible to estimate Nm values from samples drawn from a single deme.

Introduction

The connection between the past history of a population and its neutral genetic diversity has become obvious with the advent of coalescent theory (Kingman 1982*a*, 1982*b*; Hudson 1990; Nordborg 2001). Although coalescent theory was initially developed in the context of a single population, it has been rapidly extended to include subdivided populations or populations connected by migration (the structured coalescent) (Notohara 1990; Marjoram and Donnelly 1994; Slatkin 1995; Rousset 1996, 1997; Nordborg 1997; Wilkinson-Herbots 1998; Wakeley 1999, 2000, 2001; Wakeley and Aliacar 2001). Past theoretical studies have focused on island or stepping-stone models within homogeneous environments. Focusing on a finite island model, Wakeley (1999, 2001) has shown that the coalescent process in a subdivided population could be divided into two distinct phases when the number of demes is large (much larger than the number of sampled genes). Going backward in time, the first phase, called the "scattering phase," is usually rapid and ends when all sampled genes are located in different demes. It is characterized by a series of initial coalescent events, with migration events scattering the gene lineages into different demes. The second phase, called the "collecting phase," is usually much longer and describes the coalescent process between the end of the scattering phase and the ultimate coalescent event. This phase is characterized by a large number of migration events and a few coalescent events that are only possible when a gene lineage has migrated into a deme already occupied by another gene lineage. Interestingly, the coalescent during the collecting phase is similar to that of an unsubdivided population but on a timescale proportional to the effective size of the whole population, itself depending on the number of demes, the migration rate between demes, and the deme size (Wakeley 1999). Additional realism has been recently incorporated by allowing for the occurrence of a potentially changing number of demes of unequal size connected by potentially changing rates of migration (Wakeley 2001; Wakeley and Aliacar 2001), showing that coalescent events would accumulate over time in small demes with low migrations rates (Wakeley 2001). Coalescent-based approaches have also been developed to estimate nonhomogeneous and asymmetric migration rates among demes of unequal sizes (Beerli and Felsenstein 1999, 2001), albeit under the assumption that the sampled demes actually exchange migrants.

The development of more realistic models that incorporate demographic history may allow for the explanation of complex patterns that may be apparent in population genetic data. A classical example of the influence of the demographic history of a population on its molecular diversity is a recent demographic expansion that leads to starlike phylogenies (Slatkin and Hudson 1991) and to unimodal distributions of the number of pairwise difference or mismatch distributions (Rogers and Harpending 1992). While this pattern could also be obtained by a complex mutation mechanism in the absence of large expansions, for instance, heterogeneity of mutation rates (Lundstrom, Taravé, and Ward 1992; Aris-Brosou and Excoffier 1996), the study of mitochondrial DNA in many human populations suggests that most human populations have experienced Pleistocene demographic expansions (Sherry et al. 1994; Rogers

Key words: mismatch distribution, spatial expansion, demographic expansion, human evolution, mitochondrial DNA, subdivided population.

E-mail: laurent.excoffier@zoo.unibe.ch.

Mol. Biol. Evol. 20(1):76–86. 2003
DOI: 10.1093/molbev/msg009
 ISSN: 0737-4038

1995; Rogers and Jorde 1995; Harpending et al. 1998; Excoffier and Schneider 1999; Schneider and Excoffier 1999). Similarly, microsatellite data from the Y chromosome were better explained with models based on past expansion than on stationarity (Pritchard et al. 1999). In contrast, analyses of Y chromosome single nucleotide polymorphism (SNP) did not provide any clear evidence for demographic expansions (Pereira et al. 2001). Studies with nuclear markers have also provided ambiguous results. Signals of expansion were found in some but not in all populations analyzed for microsatellite data (Reich and Goldstein 1998; Beaumont 1999; Goldstein et al. 1999). SNP studies showed no signs of expansion when single populations were considered (Nielsen 2000; Wakeley et al. 2001), whereas signals of expansions were found in a subdivided population model (Wakeley et al. 2001).

It is apparent that under existing demographic models, it is difficult to establish a clear and consistent explanation for the observed patterns of human molecular diversity. Discrepancies regarding signs of demographic expansions may be due to differences in demographic histories among regions (Reich and Goldstein 1998; Goldstein et al. 1999) and among ethnic groups (food producers vs. food gatherers) (Watson et al. 1996; Excoffier and Schneider 1999), differences between loci (Beaumont 1999), ascertainment bias in the choice of markers (Wakeley et al. 2001), or a lack of resolution of some markers (Pereira et al. 2001). However, these discrepancies could also result from making inferences based on erroneous models of population history (e.g., if the population is indeed subdivided) (Marjoram and Donnelly 1994).

While extensive studies have focused on the effect of population subdivision on the shape of gene genealogies (e.g., Notohara 1990; Marjoram and Donnelly 1994; Donnelly and Tavaré 1995; Nordborg 1997; Wakeley 2001), the effect of range or spatial expansions have thus far been neglected. In the case of modern humans, estimations of the age of the demographic expansions obtained from mtDNA sequence analyses point to the Pleistocene, and so these expansions could indeed represent a global increase in effective population size due to the spread of humans after a bottleneck. Although previous work has suggested that observed patterns of molecular diversity may have resulted from a simple demographic increase, the possibility also exists that these patterns are a signal of a range expansion after a speciation event (Excoffier and Schneider 2000). Although a range expansion certainly leads to an increase in the global effective size of a species, it is not known whether it leads to exactly the same molecular signal as a demographic expansion in a single unsubdivided population.

Despite advances in analytical techniques that allow for estimates of population parameters in more realistic settings, they may become intractable under complex evolutionary scenarios. It appears, therefore, that coalescent simulations are still useful and necessary to investigate the effect of such complex scenarios (such as nonconstant environments) on various aspects of the molecular diversity of populations. In this study, we use a simulation framework to study the combined effect of spatial and demographic expansions on patterns of within-deme molecular diversity in a simple two-dimensional landscape. After simulating a wave of advance (using a simple migration model with logistic regulation of deme size), a coalescent approach is used to simulate the genetic diversity of a sample conditional on the demographic history of the population. Different aspects of the molecular diversity are recorded, and factors with the potential to affect molecular diversity (place of origin, local deme size, size of gene flow between neighboring demes, and sampling location) are investigated and discussed.

Material and Methods

To efficiently simulate the molecular diversity expected in a sample drawn from a deme belonging to a large subdivided population having gone through a recent spatial expansion, we proceed in two steps. We first use a forward simulation scheme to generate the demography (density and gene flow) of a two-dimensional array of demes initially empty except for a single deme assumed to be at carrying capacity. We then use this resulting demographic information to simulate the molecular diversity of a set of DNA sequences drawn from a single deme using a coalescent backward approach.

Demographic Simulations

Simulations were performed in a subdivided population consisting of 2,500 demes arranged in a two-dimensional stepping-stone lattice of 50×50 demes. At the beginning of a simulation, a single deme of this population is occupied with a density equal to 100 (unless specified otherwise). This ancestral deme is the source of an isotropic spatial expansion. In our simulations, we have considered just two potential locations for this ancestral deme: one was located at the center of the lattice (at position <25; 25>), and the other located near the periphery (at position <5; 5>). After the onset of the spatial expansion process, the range of the population increases due to ongoing exchange of migrants between occupied demes and their neighbors. Emigrants are sent from a given deme having density N_t at time t to neighboring demes at rate m, so that $N_t m$ emigrants are sent outwards at each generation. If a gene is sent to an occupied deme, the movement results in gene flow. If not, the movement results in the colonization of a new deme. The emigration rate does not depend on the current density of the target deme, so that the same proportion of migrants are sent to empty or occupied demes. The number of emigrants $N_t m$ is then distributed equally among the neighboring demes. The density of each deme is limited by its carrying capacity K, and is regulated logistically as

$$N_{t+1} = N_t[1 + r([K - N_t]/K)],$$

where N_t is its density at time t, and r is the intrinsic rate of increase per generation (in the current study, r was constant at 0.1). At carrying capacity, Km migrants are thus exchanged between a deme and its neighbors. In the following, this number of migrants exchanged at equilib-

rium will be denoted by Nm to be consistent with published literature. For each generation, we implement a logistic regulation step followed by a round of migration. The demographic simulations are performed for 4,000 generations, and we store for each generation t the density of the j-th deme (N_{jt}) and the number of immigrants received from the k-th deme (I_{jkt}) in a database. This demographic database is then used to perform the genetic simulations using a coalescent approach described below.

Coalescent Simulations

Under neutrality, the genetic diversity of samples in a subdivided population is easy to simulate, as it depends only on the demographic and migration histories of the demes (e.g., Hudson 1990; Nordborg 2001). For this purpose, we have modified the coalescent simulation program SIMCOAL (Excoffier, Novembre, and Schneider 2000), allowing it to take into account the dynamic nature of deme densities and migration rates between adjacent demes. Starting at the present generation, we simulate the genealogy of genes sampled in a deme located, for convenience, at one of the two previously specified positions in the grid. Because we are interested in describing intra-deme diversity, we stress the fact that samples of genes are always drawn from a single deme. At each generation and going backward in time, genes can either move to a different deme or coalesce if they are not the single gene lineage in their deme. At generation t, the probability of emigration of a gene from deme j to deme k is computed according to the information recorded in the database created during the demographic simulation step and is equal to I_{jkt}/N_{jt}. After migration, the probability of a coalescence event in deme j depends both on the number of genes (i) present in deme j and on its density at time t as $i(i-1)/(2N_{jt})$. For each generation, we first implement a coalescence phase followed by a migration phase. As usually assumed in analytical treatments, a single coalescent event is allowed per deme per generation. In the case where the deme size is not much larger than the number of gene lineages (i) present in that deme, this strategy leads to slightly longer coalescence times (up to i generations) than if several coalescent events were allowed per generation. Because i is smaller than 30 in our current simulations, it is unlikely to affect the pattern of molecular diversity that is generated over thousands of generations. The coalescent process stops when there is a single gene lineage left in the array of demes. In the case when multiple gene lineages trace back to the ancestral deme at a time corresponding to the beginning of the forward simulation, the backward coalescent process proceeds further in this single deme of density equal to its initial density (100, unless specified otherwise). During the simulations, we record the locations and times of all coalescent events. For each simulated gene genealogy, we simulate mutations on the branches of the genealogy according to a Poisson process with rate μt, where μ is the mutation rate and t is the length (in generations) of a given branch. In the present case, we simulated an unbiased substitution process on a sequence of DNA of 300 bp, with $\mu = 0.001$ for the whole sequence, assuming a finite-site mutation model without heterogeneity of mutation rates. One thousand coalescent simulations were performed for each set of demographic parameters tested.

The distribution of a number of statistics were gathered from the simulated samples, including the number of segregating sites (S), the average number of pairwise differences (π), Tajima's D statistic (Tajima 1989), Fu's F_S statistic (Fu 1997), and the mismatch distribution. All analyses were performed using the software ARLEQUIN (Schneider, Roessli, and Excoffier 2000). Unless specified otherwise, summary statistics and mismatch distributions were obtained from the simulation of samples of 30 DNA sequences.

Results

Spatial and Temporal Distribution of Coalescent Events

In figure 1, we show various aspects of the dynamics of the spatial expansion process for two different numbers of migrants (Nm) exchanged between neighboring demes. The first obvious result (fig. 1*A* and 1*B*) is that for low Nm values ($Nm = 10$), the speed of the colonization wave is slower (600 generations to colonize the 2,500 demes) than with high Nm values ($Nm = 500$) (400 generations to colonize the 2,500 demes). Note that this effect is not due to a difference in the proportion of migrants, since in both cases m was set to 0.1. This is rather due to the fact that for low Nm values, the deme takes longer to fill than for higher Nm values, and therefore migration commences later. The migration pattern also influences the timing and the location of coalescent events. A majority of coalescent events are recent, having occurred during the scattering phase (*sensu* Wakeley 1999) and been geographically located close to the sampling location for $Nm = 10$ (fig. 1*C* and *E*), while for $Nm = 500$, they are mainly older and located close to the origin of the expansion (fig. 1*D* and *F*). Note that the ultimate coalescent event is older than 4,000 generations in 96.1% of the cases (and thus occurs in the ancestral deme) when $Nm = 10$, compared with 100% of the simulated cases when $Nm = 500$.

Patterns of Molecular Diversity

We have studied different aspects of DNA sequence polymorphism within a single deme drawn from a population that has experienced a range expansion. The results of the analysis of simulated samples are reported in table 1 for different levels of migration among neighboring demes. A drastic difference is found between demes exchanging 20 migrants or less per generation and demes exchanging higher numbers of migrants. Whereas the average number of pairwise differences only slightly increases with larger Nm values (going from $\pi = 6.3$ for $Nm = 5$ to $\pi = 8$ for Nm values ≥ 200), the number of segregating sites increases much more drastically (going from $S = 30$ for $Nm = 5$ to $S = 96.9$ for $Nm = 1{,}000$). This difference can be attributed to the timing of the coalescent events, which is indeed different for small or large Nm values. Because a majority of coalescent events occur in the scattering phase for small Nm values and much later (around the onset of the expansion) for large

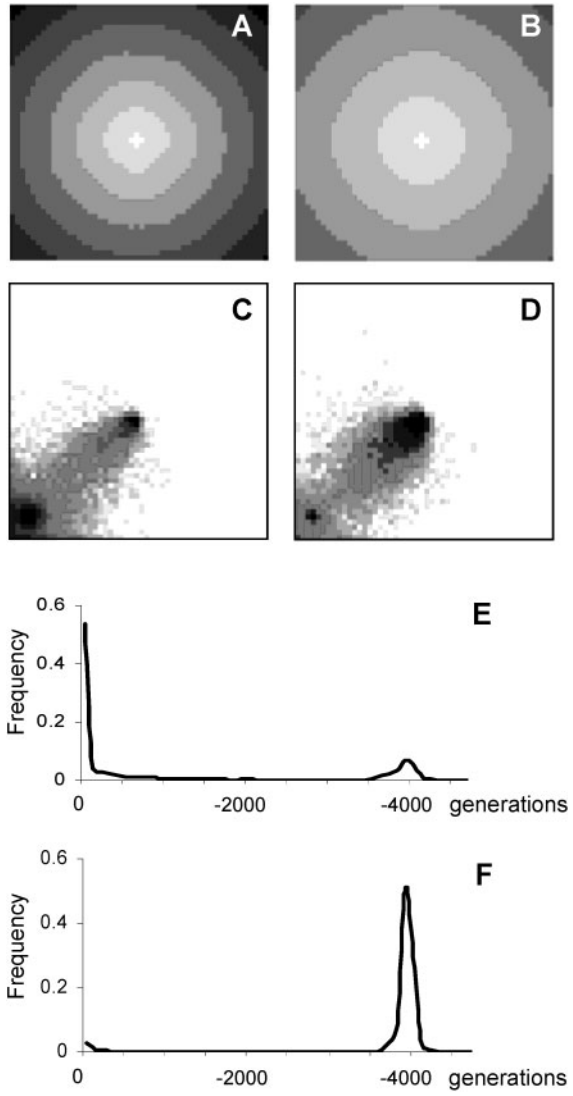

FIG. 1.—Summary of the dynamics of the spatial expansion process and its associated coalescent in a simulated subdivided population of 2,500 demes arranged as a two-dimensional stepping-stone (50 × 50 demes). *A and B*, Dynamics of the spatial expansion showing the progressive colonization of demes, with the spatial expansion starting from the central deme. Each shade of gray denotes the limit of the area of occupied demes after a further 100-generation step. The central white "cross" is the state of the expansion after just one generation. *C and D*, Empirical spatial distribution of the coalescence events obtained from 1,000 simulations of the genealogies of 30 genes sampled in a single deme in the lower left periphery (deme at position <5; 5>) of the population. Gray intensity is related to the total number of coalescent events having occurred in a given cell. Cells where no coalescent events occurred are shown in white. *E and F*, Empirical time distribution of the coalescence events obtained from 1,000 simulations. Time before present is represented on the X axis. In *A*, *C*, and E, the number of migrants exchanged between neighboring demes is lower ($Nm = 10$) than in *B*, *D*, and *F* ($Nm = 500$). In all cases, m is set to 0.1.

Nm values (see fig. 1*E* and *F*), the total length of the gene genealogy is much larger for samples of genes drawn from a deme with high *Nm* than those drawn from demes with small *Nm*. The difference in the timing of coalescence events and the overall shape and length of genealogies of genes drawn from demes with low or high *Nm* values can be seen in figure 2, where we show three random genealogies for three values of *Nm* (5, 25, and 200). As could be inferred from figure 1, there are many recent coalescent events in demes sending only a few migrants to neighboring demes, whereas recent coalescent events are rare in demes sending many migrants, resulting in very long terminal branches in the genealogies. Note that very similar gene genealogies with long terminal branches are observed in an unsubdivided population after a demographic expansion (Slatkin and Hudson 1991; Rogers and Harpending 1992).

The sampling location and the geographical location of the expansion have no effect on the pattern of molecular diversity in our homogeneous environment for large *Nm* values, whereas we observe a slight reduction in genetic diversity for demes that are sampled in the periphery of the simulated population for low *Nm* values (independent of the origin of the expansion) (table 2). Tajima's *D* statistic seems sensitive to the sampling location for low *Nm* values, as demes sampled in the center show a significant negative *D* value in 22% to 26 % of the cases at the 5% level, whereas demes sampled in the periphery only show significant *D* values in 7% to 8% of the simulations.

Mismatch Distributions

The empirical distributions of the number of differences between pairs of genes (identified here as mismatch distributions for sake of brevity) are shown in figure 3 for a subset of the cases described in table 1. In agreement with figure 1*E* and *F*, the average mismatch distributions observed in demes with small *Nm* values show two modes, whereas those observed in demes with large *Nm* values ($Nm > 50$) show a single mode. The first mode in demes with low *Nm* values corresponds to the zero-difference class, which is due to pairs of genes with a recent ancestor, whereas the second mode corresponds to pairs of genes having a common ancestor around the time of the onset of the spatial expansion. The 90% empirical confidence intervals for the mismatch distributions presented in figure 3 also show that the variance of the mismatch distributions is much larger for low than for large *Nm* values. In figure 4, we report four random simulated mismatch distributions for demes with either low (10) or large (500) *Nm* values. We see that while the average mismatch distribution for low *Nm* values is bimodal, single realizations of the coalescent in such cases can lead to multimodal and very ragged distributions. In contrast, the mismatch distributions in demes with large *Nm* values are most often unimodal and closer to their expectation, in agreement with the reduced variance shown in figure 3.

In figure 5, we report the mismatch distributions obtained for very different combinations of carrying capacities (K) and m values leading to the same *Nm* value at equilibrium (when $N = K$). It is clear from this figure that the average shape of the mismatch distributions (and therefore the underlying coalescent process) depends mainly on the value of the product $N \times m$ and almost not on the absolute values of deme size or migration rate. We note, however, that for a given low *Nm* value, there is a slight decrease in the zero-frequency class with larger N

Table 1
Summary Statistics Describing the Pattern of Polymorphism Found in a Sample of 30 DNA Sequences

Nm	π[a]	Var(π)	S[b]	Var(S)	D[c]	$P(D) < 0.05$[d]	F_S[e]	$P(F_S) < 0.05$[f]
5 . . .	6.3	13.4	30.0	57.0	−0.55	0.03	1.56	0.00
10 . . .	7.0	11.7	41.7	75.7	−1.20	0.26	−0.42	0.01
20 . . .	7.5	9.8	56.9	85.9	−1.77	0.87	−3.63	0.29
50 . . .	7.8	8.3	76.4	97.6	−2.25	1.00	−11.38	0.99
100 . . .	7.9	7.8	85.2	88.0	−2.40	1.00	−15.93	1.00
200 . . .	8.0	7.6	90.8	72.8	−2.48	1.00	−19.61	1.00
250 . . .	8.0	7.5	93.1	74.6	−2.52	1.00	−21.81	1.00
500 . . .	8.0	7.5	96.0	69.2	−2.55	1.00	−23.13	1.00
1000 . . .	8.0	7.3	96.9	71.4	−2.57	1.00	−23.98	1.00

NOTE.—The sequences are 300 bp drawn from a single deme after a spatial expansion that occurred $\tau = 2Tu = 8$ units of times ago. In this case $T = 4{,}000$ generations, $u = 0.001$, and the sampled deme was located in the center of the array of demes shown in figure 1*A* and *B*, at the same location as the origin of the expansion.

[a] Mean number of differences between all pairs of sequence in the sample.

[b] Number of segregating sites.

[c] Tajima's D statistic (Tajima 1989).

[d] Probability that Tajima's D statistic is found significant at the 5% level estimated from 1,000 simulations.

[e] Fu's F_S statistic (Fu 1997).

[f] Probability that Fu's F_S statistic is found significant at the 5% level estimated from 1,000 simulations.

values (fig. 5, left column with $K = 500$ and $K = 1{,}000$, as compared with $K = 100$). Note that no such effect is observed for large *Nm* values, as shown in the right column of figure 5. This phenomenon may be due to the fact that with low N values (implying a large m value), several gene lineages may initially comigrate in the same deme and subsequently coalesce, whereas with smaller m values, gene lineages will migrate once at a time. The comigration of genes in the same deme thus slightly increases the probability of recent coalescent events, leading to the slightly larger probabilities of no differences between genes sampled in demes of small size and exchanging a large fraction of genes with their few neighbors.

The age of the expansion seems to affect the pattern of diversity in a more drastic way for low than for large *Nm* values (table 3). For $Nm = 500$, Tajima's D and Fu's F_S statistics are very efficient in detecting departure from population equilibrium, irrespective of the age of the expansion. In contrast, for $Nm = 10$, Tajima's D statistic is much less powerful, showing departure from equilibrium between only one fourth and one third of the cases. For the same amount of gene flow, the behavior of the test based on Fu's F_S statistic is markedly different. The hypothesis of selective neutrality and population equilibrium will be more often rejected for relatively recent expansions ($\tau < 3$) than for older expansions ($\tau > 5$), which is somewhat counterintuitive. However, we may propose the following explanation. Since Fu's F_S statistic is the logit of the probability to observe k or more alleles conditional on π, the observed average number of pairwise differences, the behavior of this test can be explained by understanding the behavior of k and π under spatial expansions. As visible on the first row of figure 2, a range expansion with limited gene flow among demes produces an intra-deme gene genealogy with both many recent and many old coalescent events. The age of the old coalescent events depend essentially of the age of the expansion, whereas the age of the recent coalescent events depends on the size of the deme. For a sufficiently large mutation rate, the age of the expansion will not affect much k, but it will have a large effect on π. Thus, the probability of observing a given number k or more alleles will increase with older expansions, leading to less negative F_S values, as shown in table 3. The effect of the age of the expansion on the mismatch distribution is clearer, and much like in the case

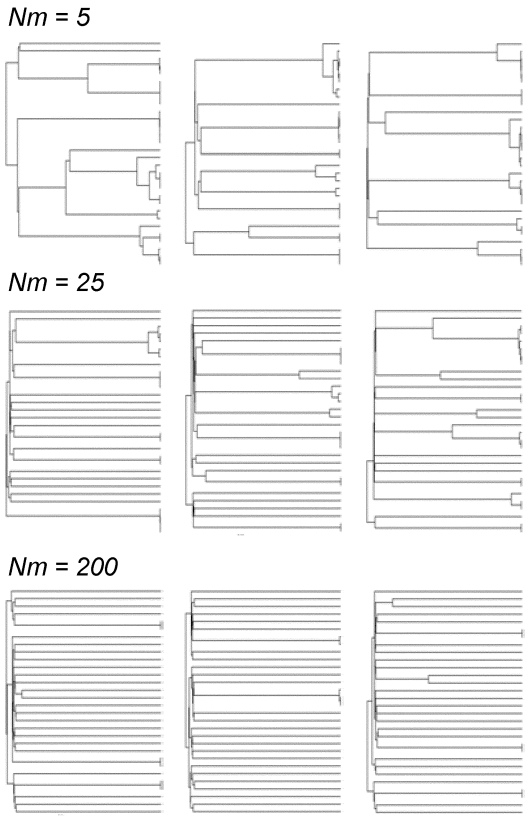

FIG. 2.—Gene genealogies after a spatial expansion. Three random genealogies of 30 genes are shown for *Nm* values of 5, 25, and 200 migrants exchanged between neighboring demes. The spatial expansion occurred $\tau = 8$ units of time ago, as indicated in the footnote of table 1.

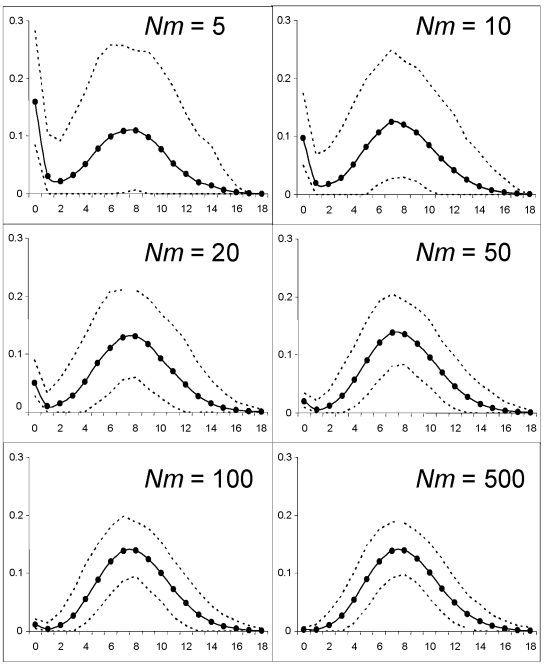

FIG. 3.—Average mismatch distributions after a spatial expansion for different *Nm* values. The Y axis stands for the average probability that two DNA sequences differ at a given number of sites represented on the X axis. The solid lines are average mismatch distributions obtained from 1,000 simulations of the coalescent of 30 genes drawn in a single deme after a spatial expansion having occurred $\tau = 8$ units of time ago. Dotted lines delimit an empirical 90% confidence interval for the mismatch distribution.

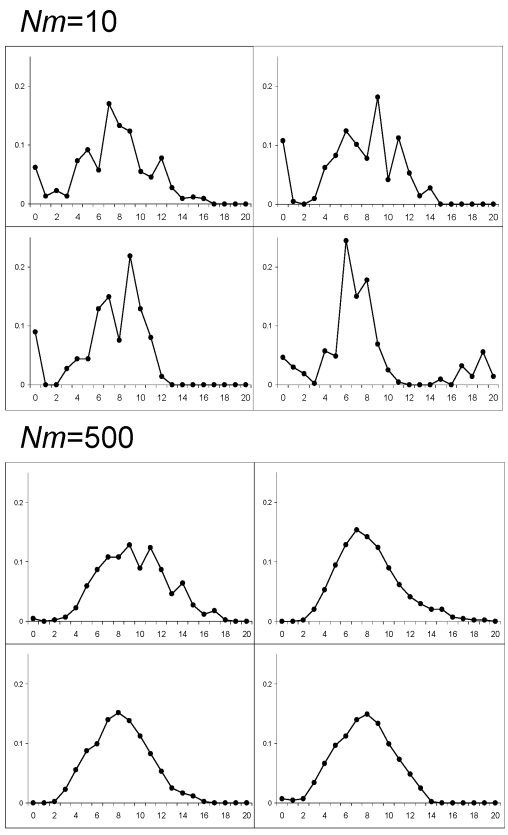

FIG. 4.—Mismatch distributions obtained for single realizations of the coalescent after a spatial expansion for demes exchanging either a low (*Nm*=10) or a large (*Nm* = 500) numbers of genes each generation with neighboring demes.

of demographic expansion in unsubdivided populations (Rogers and Harpending 1992), the mismatch distribution mode is shifted to the right with older expansion times (data not shown).

Discussion

Implication for Human Mitochondrial DNA Diversity

Previous interpretations concerning the pattern of diversity in mitochondrial mtDNA have relied on the assumption that populations were unsubdivided (Slatkin and Hudson 1991; Rogers and Harpending 1992; Rogers 1995; Weiss, Henking, and von Haeseler 1997; Excoffier and Schneider 1999), with some exceptions (e.g., Marjoram and Donnelly 1994). Under this paradigm, unimodal mismatch distributions have been interpreted as being due to past demographic expansions (Slatkin and Hudson 1991; Rogers and Harpending 1992). However, although it is true that most human populations show approximately unimodal mismatch distributions compatible with Pleistocene population expansions (fig. 6*A*), almost all present or recent hunter-gatherer groups show very ragged distributions and in particular a high proportion of pairs of sequences that are similar, thus showing no differences (fig. 6*B*). This contrast has been interpreted as the consequence of a recent (post-Neolithic) contraction

Table 2
Influence of the Sampling Location and the Expansion Origin on Patterns of Molecular Diversity

Expansion Origin	Sampling Location	*Nm*	π	Var(π)	*S*	Var(*S*)	*D*	$P(D) < 0.05$	F_{Sz}	$P(F_S) < 0.05$
Periphery . .	Periphery	10	6.6	11.8	34.7	60.5	−0.86	0.08	−0.04	0.00
Periphery . .	Center	10	6.9	11.0	40.3	71.7	−1.16	0.22	−0.52	0.01
Center	Periphery	10	6.6	11.8	34.3	54.6	−0.82	0.07	−0.09	0.01
Center	Center	10	7.0	11.7	41.7	75.7	−1.20	0.26	−0.42	0.01
Periphery . .	Periphery	500	8.0	7.4	94.4	65.0	−2.53	1.00	−22.89	1.00
Periphery . .	Center	500	8.0	7.4	95.7	65.9	−2.55	1.00	−23.21	1.00
Center	Periphery	500	7.9	7.3	93.2	77.6	−2.53	1.00	−23.10	1.00
Center	Center	500	8.0	7.5	96.0	69.2	−2.55	1.00	−23.13	1.00

NOTE.—Center refers to the central deme in our simulated array of 50 × 50 demes. It is thus located at position <25; 25>. Periphery refers to a deme located in the periphery of the simulated array, at position <5; 5>. The remaining headers are identical to those described in table 1.

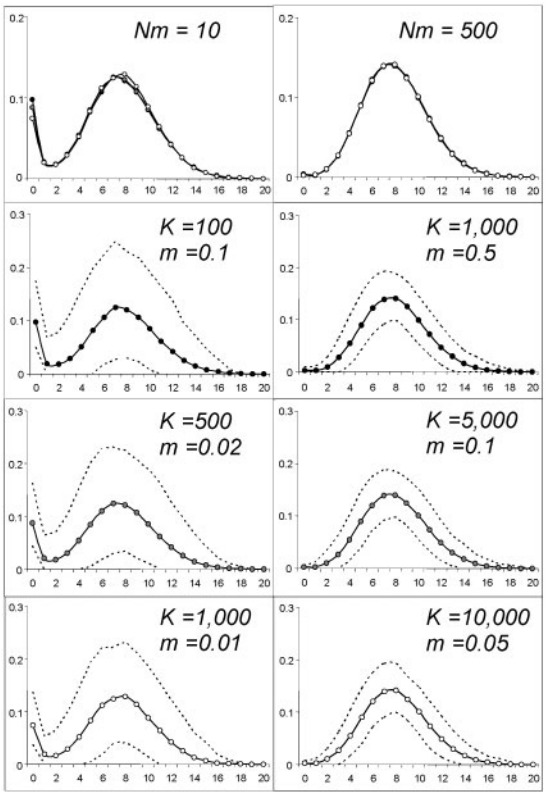

FIG. 5.—Mismatch distributions obtained for three different combinations of carrying capacities (K) and proportion of migrants exchanged with neighboring demes (m), leading to similar Nm values at carrying capacity ($N = K$). Left panels: $Nm = 10$; right panels: $Nm = 500$. The averaged mismatch distributions corresponding to the three different cases are superimposed on the top panels and are shown separately with their 90% confidence intervals on the three lower panels.

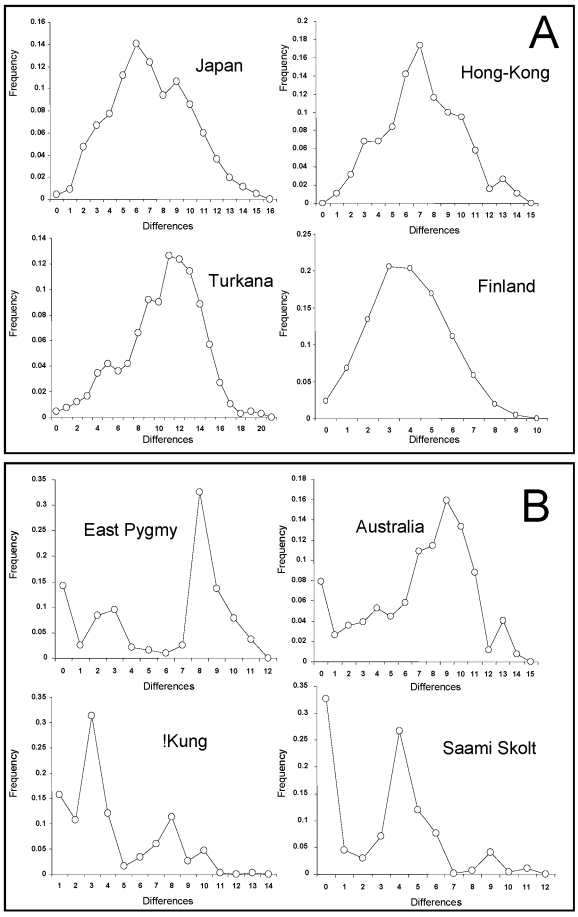

FIG. 6.—Observed mismatch distributions in human populations analyzed for mtDNA hypervariable region 1 (HVR1). Data drawn from samples referenced in Excoffier and Schneider (1999). *A*, Post-Neolithic populations. *B*, Present or former hunter-gatherer populations.

of the size of hunter-gatherer populations, resulting from the fragmentation of their habitat leading to contraction of their effective size (Excoffier and Schneider 1999).

Our present results would however lead to a simpler and very different interpretation of the differences in the shape of mismatch distribution between post-Neolithic and hunter-gatherer populations. By assuming that the present distribution of human populations results from some spatial range expansion, this contrast would simply result from the much larger deme size of Neolithic populations (resulting in much larger Nm values) than hunter-gatherer populations. While our simulations have assumed constant deme sizes from the onset of the range expansion to the present time, it is easy to simulate a range expansion with

Table 3
Different Statistics Summarizing the Pattern of Molecular Diversity After Range Expansion

Nm	Range Expansion[a]	π	Var(π)	S	Var(S)	D	$P(D) < 0.05$	F_S	$P(F_S) < 0.05$
10	1	1.1	0.9	7.1	8.0	−1.17	0.31	−2.74	0.57
	3	2.8	3.0	17.9	23.7	−1.28	0.32	−3.38	0.45
	5	4.5	5.8	27.6	40.3	−1.25	0.29	−2.21	0.17
	7	6.1	9.5	36.7	62.9	−1.19	0.26	−0.90	0.02
	8	7.0	11.7	41.7	75.7	−1.20	0.26	−0.42	0.01
500	1	1.2	1.1	14.5	13.6	−2.26	0.99	−11.50	1.00
	3	3.1	2.9	40.4	37.2	−2.55	1.00	−23.06	1.00
	5	5.1	4.8	64.4	48.0	−2.58	1.00	−24.55	1.00
	7	7.1	6.8	85.9	75.0	−2.56	1.00	−24.06	1.00
	8	8.0	7.5	96.0	69.2	−2.55	1.00	−23.13	1.00

NOTE.—The different times of expansions are $T = 500$, 1,500, 2,500, 3,500, and 4,000 generations ago, and the different migration intensities between neighboring demes are $Nm = 10$ and 500.

[a] Date of the onset of the range expansion τ, in units of mutation rate u, as $\tau = 2Tu$, where T is the time of the expansion in number of generations, and $u = 0.001$.

Table 4
Pattern of Molecular Diversity After a Spatial Expansion

#	K_0	K_1	Size Exp.	π	Var(π)	S	Var(S)	D	$P(D) < 0.05$	F_S	$P(F_S) < 0.05$
A	1000	1000	—	9.6	11.7	90.1	95.6	−2.20	1.00	−13.9	1.00
B	100	1000	500	7.8	8.1	73.4	81.2	−2.19	1.00	−14.2	1.00
C	100	1000	100	7.4	8.8	60.4	85.0	−1.91	0.95	−8.1	0.85
D	100	1000	50	7.5	9.6	56.0	84.8	−1.75	0.84	−5.3	0.54
E	100	1000	10	7.1	11.4	43.3	74.1	−1.26	0.29	−0.8	0.01
F.	100	100	—	7.0	11.7	41.7	75.7	−1.20	0.26	−0.4	0.01

NOTE.—The expansion started 4,000 generations ago and was followed by a more recent global demographic expansion at different times in the past (T = 10, 50, 100, and 500 generations ago). A fraction m = 0.10 of migrants are constantly exchanged between neighboring demes. K_0 = Carrying capacity of the demes before the demographic expansion. K_1 = Carrying capacity of the demes after the demographic expansion. Size Exp. is the time in generations (before present) at which the demographic expansion occurs.

small deme size and a recent increase in deme size resulting in higher levels of gene flow with surrounding demes. The results of such simulations are shown in table 4 for the pattern of molecular diversity and in figure 7 for the mismatch distributions.

We find that demographic expansions having occurred more than 100 generations ago and resulting in a 10-fold increase in Nm values (from $Nm = 10$ to $Nm = 100$) would lead to unimodal distributions (fig. 7*B* and *C*), as if their size had always been 10-fold higher (fig. 7*A*). In contrast, more recent demographic expansions would lead to a greater number of recent coalescent events and multimodal distributions (fig. 7*D* and *E*), as if deme size had always been low (fig. 7*F*). Patterns of molecular diversity show a very similar trend (table 4), with demographic expansions having occurred more than 50 generations ago resulting in a clear rejection of neutrality and population equilibrium with Tajima's D or Fu's F_S statistic. These simulations clearly show that relatively recent demographic expansions leading to overall larger Nm values lead to patterns of molecular diversity equivalent to those expected in demes having always exchanged large numbers of individuals with their neighbors. It thus seems that the Nm value prevailing during the scattering phase (*sensu* Wakeley 1999) of the gene lineages is the factor that will primarily determine the overall pattern of diversity observed within demes. Large Nm values during the scattering phase are sufficient to prevent recent coalescent events. The ancestral lineages of almost all sampled genes will thus be found in different demes at the end of the scattering phase. After that point, if the number of demes is much larger than the number of remaining lineages, the size of the demes (and their associated Nm value) will have almost no effect on the pattern of coalescence until the onset of the spatial expansion. This property should make the model quite robust to the likely complex histories of natural populations going through long-term size fluctuations.

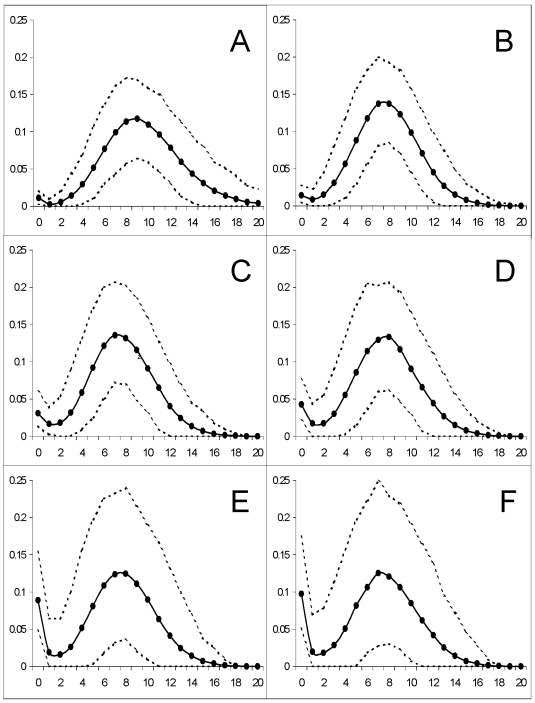

FIG. 7.—Observed mismatch distributions after a spatial and an optional demographic expansion. Cases *A* through *F* correspond to demographic histories defined in table 4. *A*, Demes with constant K = 1,000. *B*, Demographic growth from $K_0 = 100$ to $K_1 = 1{,}000$ occurring 500 generation before present. *C*, Same as *B* but demographic growth occurring 100 generations before present. *D*, Same as *B* but demographic growth occurring 50 generations before present. *E*, Same as *B* but demographic growth occurring 10 generations before present. *F*, Demes with constant $K = 100$.

The present simulation results thus explain the difference between the mismatch distributions of hunter-gatherer and post-Neolithic populations by the simple fact that food gatherers have generally lower densities than food producers (if one assumes that both groups have approximately similar emigration rates). However, additional factors may have led to different patterns of molecular diversity in these communities. It remains true that present hunter-gatherer communities currently live in environments that are unfavorable and more fragmented than before (Lewin 1988), which could have reduced considerably their effective population size and thus led to multimodal mismatch distribution (Excoffier and Schneider 1999). Such a process would certainly reinforce the difference in recent deme size between the two types of communities and contribute to the extreme raggedness of hunter-gatherer mismatch distributions. But we feel that a realistic model of population differentiation should necessarily take into account the subdivision of human populations. Therefore, a scenario with global demographic growth and subsequent bottlenecks to explain observed differences between patterns of diversity in food-producing and food-gathering populations appears much less par-

simonious and less likely than simply taking into account the finite spatial structure of the demes and the low census size of hunter-gatherers.

Distinction Between Spatial and Demographic Expansions

We find that although spatial expansions also involve a demographic increase at the level of the population as a whole, they do not necessarily lead to a molecular signature similar to that of sudden demographic expansions in unsubdivided populations. This is the case only if the amount of gene flow is large between neighboring demes. For relatively low levels of gene flow ($Nm < 20$), recent coalescent events and therefore multimodal mismatch distributions can be expected in a quite large fraction of simulations (table 1), even if the global size of the population has been increased by several orders of magnitude after the expansion. The dependence between the amount of gene flow between demes and the average level of genetic diversity (π) within deme observed after a spatial expansion is different from that expected in a subdivided population at equilibrium. Several studies have indeed shown that the average coalescence time between a pair of genes should only depend on the total size of the population, if demes are all either directly or indirectly interconnected (Slatkin 1987; Strobeck 1987; Hey 1991) and if the number of demes is constant (Nagylaki 1998). Examination of table 1 suggests that demes with low levels of gene flow should show lower average levels of diversity (both lower π and lower S values) than demes with high gene flow after a spatial or range expansion. Also note that what we call "low levels of gene flow" are still cases where Nm is much greater than 1, which is generally the value above which spatially arranged demes are assumed to evolve as a single unit (e.g., Maruyama 1971). This result underlines the need to further study spatial models of populations out of equilibrium.

Another prediction that may reveal differences between models of demographic and spatial expansions is the relationship between the geographical location of the sample and its genetic diversity. Results shown in table 2 suggest that demes sampled in the periphery of the present population range may show slightly reduced levels of molecular diversity for low Nm values, regardless of the origin of the expansion. This may be due to the fact that gene lineages are less free to diffuse to different demes in the scattering phase when they are close to the border of the expansion range. They would thus spend more time within the same deme and have therefore more time to coalesce. The spatial diffusion constraints during the scattering phase would lead to an excess of recent coalescent events as compared with genes sampled in more central demes. This suggest that the pattern of molecular diversity within samples should be affected by the presence of geographical barriers preventing a free diffusion of genes to neighboring demes for species having low dispersal abilities. Note that this effect would be quite different from the reduced diversity expected in marginal populations and resulting from a demic diffusion process from a given source (Rendine, Piazza, and Cavalli-Sforza 1986; Sokal, Oden, and Wilson 1991; Barbujani, Sokal, and Oden 1995), where one would expect a loss of genetic diversity due to a succession of small founder effects. However, a clearer distinction between demographic and spatial expansions should emerge from the study of samples of genes taken from different demes, which should be the object of a different study.

Recent Range Expansions as a Way to Examine Patterns of Dispersal from Single Samples

Recent range expansions and speciations are thought to have been quite common in the Quaternary, following or due to ice ages, respectively (for a review, see e.g., Taberlet et al. 1998; Hewitt 2000). It is therefore likely that the traces of recent spatial expansions could be found in many species other than humans, in fact in all populations that would have gone through very small sizes during former ice ages spent in refuge areas, from where they would have then reexpanded. Interestingly, the fact that some populations would have expanded from a refuge area would not only tell us something about their global dispersal abilities but could also bring important information on their recent rate of dispersal outside their demes. Since the shape of the mismatch distribution, and particularly the frequency of recent coalescent events, depends on recent migration rates, it should be possible to estimate emigration rates by sampling individuals from the same deme and examining their pattern of molecular diversity. Applied to sex-linked markers, this could allow one to study potential sex-biased dispersal and/or different effective size between sexes. An estimation procedure for Nm values inferred from a single sample drawn from a recently expanding population is currently under investigation, and it will be the subject of a forthcoming paper. Available methods for estimating levels of gene flow usually rely on the availability of a series of samples. Gene flow is then inferred between demes from which the samples are supposed to be drawn (see e.g., Beerli and Felsenstein 2001). This implies that sampled demes actually exchange migrants and that one is able to define the geographical limit of the deme. The validity of these two assumptions is generally quite difficult to assess and would not be required from the analysis of single samples. We are therefore confident that the analysis of patterns of molecular diversity from single deme samples would allow one to get important insights on the life history of the numerous populations having gone through recent range expansions.

Acknowledgments

Thanks to Pierre Berthier and Stefan Schneider for their computing and programming assistance. We are grateful to Grant Hamilton for his careful reading of the manuscript and to Montgomery Slatkin and Henry Harpending for their comments on an earlier version of the manuscript. We are also indebted to John Wakeley for his many suggestions helping to improve various aspects of the manuscript. This work was supported by a Swiss NSF grant No 31–054059–98 to L.E.

Literature Cited

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.

Barbujani, G., R. R. Sokal, and N. L. Oden. 1995. Indo-European origins: a computer-simulation test of five hypotheses. Am. J. Physical Anthropol. **96**:109–132.

Beaumont, M. A. 1999. Detecting population expansion and decline using microsatellites. Genetics **153**:2013–2029.

Beerli, P., and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152**:763–773.

———. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA **98**:4563–4568.

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

Excoffier, L., J. Novembre, and S. Schneider. 2000. SIMCOAL: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J. Hered. **91**:506–510.

Excoffier, L., and L. Schneider. 2000. The demography of human populations inferred from patterns of mitochondrial DNA diversity. Pp. 101–108 *in* C. Renfrew and K. Boyle, eds. Archaeogenetics: DNA and the population prehistory of Europe. McDonald Institute for Archeological Research, Cambridge.

Excoffier, L., and S. Schneider. 1999. Why hunter-gatherer populations do not show sign of Pleistocene demographic expansions. Proc. Natl. Acad. Sci. USA **96**: 10597–10602.

Fu, Y.-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. Genetics **147**:915–925.

Goldstein, D. B., G. W. Roemer, D. A. Smith, D. E. Reich, A. Bergman, and R. K. Wayne. 1999. The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics **151**: 797–801.

Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers, and S. T. Sherry. 1998. Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95**:1961–1967.

Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. Nature **405**:907–13.

Hey, J. 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. Theor. Popul. Biol. **39**:30–48.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. J. Futuyma and J. D. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, New York.

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

———. 1982*b*. On the genealogy of large populations. J. Appl. Proba. **19A**:27–43.

Lewin, R. 1988. New views emerge on hunters and gatherers. Science **240**:1146–1148.

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

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

Maruyama, T. 1971. Analysis of population structure. II. Two-dimensional stepping stone models of finite lengths and other geographically structured populations. Ann. Hum. Genet. **35**:179–196.

Nagylaki, T. 1998. The expected number of heterozygous sites in a subdivided population. Genetics **149**:1599–1604.

Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154**:931–942.

Nordborg, M. 1997. Structured coalescent processes on different time scales. Genetics **146**:1501–1514.

———. 2001. Coalescent theory. Pp. 179–212 *in* D. Balding, M. Bishop, and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, New York.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population. J. Math. Biol. **29**: 59–75.

Pereira, L., I. Dupanloup, Z. H. Rosser, M. A. Jobling, and G. Barbujani. 2001. Y-chromosome mismatch distributions in Europe. Mol. Biol. Evol. **18**:1259–1271.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16**:1791–1798.

Reich, D. E., and D. B. Goldstein. 1998. Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95**:8119–8123.

Rendine, S., A. Piazza, and L. L. Cavalli-Sforza. 1986. Simulation and separation by principal components of multiple demic expansions in Europe. Am. Nat. **128**:681–706.

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

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

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

Rousset, F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142**:1357–1362.

———. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics **145**:1219–1228.

Schneider, S., and L. Excoffier. 1999. Estimation of demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152**:1079–1089.

Schneider, S., D. Roessli, and L. Excoffier. 2000. ARLEQUIN: a software for population genetics data analysis. Version 2.000. University of Geneva, Geneva, Switzerland.

Sherry, S. T., A. R. Rogers, H. Harpending, H. Soodyall, T. Jenkins, and M. Stoneking. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. Hum. Biol. **66**:761–775.

Slatkin, M. 1987. The average number of sites separating DNA sequences drawn from a subdivided population. Theor. Popul. Biol. **32**:42–49.

———. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

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

Sokal, R. R., N. L. Oden, and C. Wilson. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature **351**:143–145.

Strobeck, K. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117**:149–153.

Taberlet, P., L. Fumagalli, A. G. Wust-Saucy, and J. F. Cosson. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Mol. Ecol. **7**:453–464.

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

Wakeley, J. 1999. Nonequilibrium migration in human history. Genetics **153**:1863–1871.

———. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution **54**:1092–1101.

———. 2001. The coalescent in an island model of population subdivision with variation among demes. Theor. Popul. Biol. **59**:133–144.

Wakeley, J., and N. Aliacar. 2001. Gene genealogies in a metapopulation. Genetics **159**:893–905.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero, and K. Ardlie. 2001. The discovery of single-nucleotide polymorphisms-and inferences about human demographic history. Am. J. Hum. Genet. **69**:1332–1347.

Watson, E., K. Bauer, R. Aman, G. Weiss, A. von Haeseler, and S. Paabo. 1996. mtDNA sequence diversity in Africa. Am. J. Hum. Genet. **59**:437–444.

Weiss, G., A. Henking, and A. von Haeseler. 1997. Distribution of pairwise differences in growing populations. Pp. 81–95 *in* P. Donnely and S. Tavaré, eds. Progress in population genetics and human evolution. Springer Verlag, New York.

Wilkinson-Herbots, H. M. 1998. Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. **37**:535–585.

Naruya Saitou, Associate Editor

Accepted September 13, 2002