COPYRIGHT NOTICE:

Bryan K. Epperson: Geographical Genetics

is published by Princeton University Press and copyrighted, © 2003, by Princeton University Press. All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher, except for reading and browsing via the World Wide Web. Users are not permitted to mount this file on any network servers.

For COURSE PACK and other PERMISSIONS, refer to entry on previous page. For more information, send e-mail to permissions@pupress.princeton.edu

CHAPTER 3

Ancient Events in Spatial–Temporal Processes

> The occurrence of new alleles each produced by a unique mutation or recombination [causes] . . . the “gametic pool” of Sewall Wright to be extended, as time progresses, to an indefinitely increasing number of new alleles; now called infinitallelism, which might have been the germ of the molecular clock had not . . . the (falsely Mendelian) consensus about fixed genes . . . paralyzed Post-Darwinism imagination.
>
> —Gustave Malécot (1998 personal communication)

Major events in the distant past may leave transient signatures in spatial and spatial–temporal patterns of genetic variation. Important ancient events include the time and place of genetic “innovations,” refugia, range expansions, colonizations or major immigration events, and fragmentation. Transient effects of ancient events contrast with stable patterns that can be produced by selection, genetic drift, and migration averaged over long periods. The study of transient effects of major events in the distant past calls for a somewhat different emphasis in the context of spatial–temporal processes. For example, in the theoretical works of Malécot (e.g., 1948) the focus was on deriving spatial distributions produced over long periods or at equilibrium, and they provide an entirely appropriate basis for empirical studies of stable geographical patterns of genetic variation. Equilibrium results do not show the transient effects of major events. Moreover, we should distinguish long-lasting yet transient effects of ancient events from more recent or shorter term transient effects. The latter may be analyzed using the migration matrix approach or short-term space–time correlations.

The signatures left by most nonrecurring ancient events are delible and continued gene flow will eventually erase them. Many empirical studies aim to detect the trace of an ancient event, in effect to parse off particularly important features of the past from the spatial–temporal context. Frequent goals are to infer the geographic origins of new genetic variants or by extension the geographic origins of species themselves. In this chapter examples are used to illustrate the primary issues of the general conditions required for valid separation of particular temporary spatial or spatial–temporal patterns from the space–time process in which they are embedded. Typically, such studies do the following: (1) use population differentiation per se of molecular variation at the present time; (2) conduct phylogenetic reconstruction to infer the gene genealogies, and the ancestral genotype (generally without including information on spatial proximity, structure, and migration in the probability models); and then (3) use present spatial patterns of types that are most like the inferred "ancestral" type, and hence infer the past location of ancestral type. Examination of such studies also illustrates some of the key distinctions of modern molecular data from gene frequency data. Phylogenetic studies can sometimes take advantage of a unique kind of temporal "depth" (Templeton 1998) to spatially distributed molecular data. Steps 1 and 2 may not encounter serious problems, particularly if the timescale on which mutations accumulate is much slower than that on which migrations occur, and when coalescences within populations are much more recent than those among populations (Nordborg 1997). The rationale for the third step has received the least attention, and the step often appears to be made subconsciously. While it may seem to be a safe assumption that a given gene sampled at the present time is most closely related to the past or ancient genes in the same population, the gene may also be very closely related to ancient genes in other populations, perhaps in some cases quite distant geographically. This is a recurring theme in the inference of major events in spatial–temporal processes.

The study of ancient events in population genetics generally requires extensive datasets. Most survey data are contemporary, but as ancient DNA samples become more available, they may become disproportionately important and informative. There are still few species for which sufficient, even purely spatial (i.e., not space–time) datasets exist. There can be no doubt that in the near future sample sizes orders of magnitudes larger will become available for some species, including humans. But because data requirements have been yet rarely met in surveys, much of this chapter is devoted to a specific example, the origin of anatomically modern humans. Other examples of ancient population genetics follow, for humans and other species.

OUT OF AFRICA

Studies of human genetics typically have especially large numbers of widely separated study populations, with large sample sizes and numbers of genetic markers, and this makes them well suited for inferring the location of an ancestral population from the distant past. We will examine issues around the so-called out-of-Africa hypothesis or theory of the origins of modern humans. I wish to make it very clear that this examination is not necessarily intended to challenge this theory. It very well may be true, and there are many considerations not addressed here that support the theory.

The paramount features of the out-of-Africa hypothesis surround the geographic location and isolation of the first *anatomically* modern humans. By some half-million years ago so-called archaic hominids had spread throughout much of the Old World. The simplest form of the out-of-Africa hypothesis is that the first anatomically modern humans evolved in a small population of probably less than 10,000 individuals, in complete reproductive isolation, somewhere in Africa, about 200,000 years ago. Once evolved, this population began to increase in size and spread geographically. One scenario has it that humans spread along the coastal zones of Africa, and later into Asia and Europe. As they spread further to various regions of the Old World, they must have come into contact with the archaics already present, but did not interbreed at all. All genes today are descended from those in the isolated original population in Africa. This theory has become widely accepted in the last decade or so, while enthusiasm for a contrasting theory, the “multiregional hypothesis” (e.g., Wolpoff 1989), waned. The multiregional hypothesis states that the gene pool of anatomically modern humans, us, contains substantial contributions from many prior regionalized and differentiated archaic populations. Much of the cited support for the out-of-Africa theory is genetic, in particular the pattern of genetic differentiation among geographically separated groups, especially ethnic populations that have not undergone recent global migrations. There is also some physical evidence for the theory (e.g., Klein 1995; Lahr 1996; Sokal et al. 1997b).

The out-of-Africa theory became widely supported following phylogenetic studies of mitochondrial DNA, mtDNA. Mitochondrial DNA is strictly maternally inherited (e.g., Stoneking and Soodyall 1996), hence the genealogy of mtDNA is matrilineal. In a highly influential and widely publicized paper, Cann et al. (1987) studied samples of mtDNA from global populations and inferred a "mitochondrial Eve," the woman who carried the most recent common ancestor (MRCA) of all mitochondria today. In other words, all mitochondria lineages trace back to or "coalesce" in the mitochondrial Eve. Using polymorphic sites in the data and the molecular clock, Cann et al. (1987) estimated that "Eve" lived about 200,000 years ago. While this feature captured headlines and public imagination, the existence of a mitochondrial Eve is a necessary outcome of life, because *any set of genes must trace to a common ancestor at some time in the past*. Moreover, because mtDNA does not recombine, the entire mtDNA genome must all be descended from a single woman. The estimate that the time back to the most recent common ancestor, or TMRCA, was 200,000 years ago is interesting in part because it coincides with fossil evidence of the appearance of anatomically modern humans. However, this coincidence in itself also does not mean much, because there is no a priori reason to expect the TMRCA of a set of haplotypes or DNA sequences to necessarily coincide with the event of origination.

If there is selectively neutrality, the TMRCA should depend primarily on a function of the overall effective population size, and estimates of the TMRCA depend also on the mutation rates. For example, the TMRCA could in principle occur long after the origin of modern humans, in particular if human population size bottlenecks occurred after the origin. The TMRCA could also have been much earlier than the origination, if populations remained large prior to the formation and throughout the existence of modern humans. There could have been polymorphism within the theorized isolated single original population under the out-of-Africa hypothesis, depending on what its size was and for how long it was isolated. Based on the estimated value of TMRCA, the human effective population size, which is usually a function of the harmonic mean, since the time Eve existed has been estimated at around 10,000. A number of studies have argued that this is too small to fit the multiregional hypothesis, because, they maintain, the Old World population of archaics must have been much larger simply to have been sustained (e.g., Harpending et al. 1998).

Effective population size N_e is an important concept in population genetics, and it is worth noting some of its general properties. It simplifies the extension of theoretical models, originally constructed for an "ideal" population, to many other situations. Typically, the ideal population is constant in size, with monoecious random mating, and certain constraints on the variance of numbers of progeny produced per parent (e.g., see discussion about the process Equation 5.1 in chapter 5). Generally, an N_e is a function of the actual population size N and other, modifying factors, and it can simply be substituted for N in the model equations formulated for the ideal population. Equations are variously expressed in terms of probabilities of identity by descent, coalescences, or gene frequency covariances or variances, hence N_e is "effective" with respect to these measures. The most common are the "inbreeding effective number" and the "variance effective number" (e.g., see review by Crow and Denniston 1988). In some cases, but not others, they are equivalent. In general, population biological factors that substantially affect N_e include unequal numbers of females and males, large variance in numbers of offspring, and various forms of systemic inbreeding (e.g., Caballero and Hill 1992). However, when examining out-of-Africa the increase in population size is far more important. N_e is strongly disproportionately affected by any small sizes, as, for example, when N_e is a function of the harmonic mean (e.g., Crow and Denniston 1988).

While the TMRCA of mtDNA appears to be on fairly solid footing, it is much more difficult to determine the sequence and timing of how population size may have expanded and contracted during the last 200,000 years or so. The TMRCA for mtDNA is not so satisfying because it represents only the mitochondria. It does not mean that all other (e.g., autosomal) genes also coalesced at the same time, nor even that all genes came from the same population as the mitochondrial Eve. More recent analyses of larger datasets of mtDNA polymorphisms (e.g., Stoneking and Soodyall 1996) generally support the TMRCA reported by Cann et al. (1987). Together with the TMRCA, the genotype of the most recent common ancestor is also inferred from the gene genealogy, based solely on probabilities of mutations. It should be noted that the actual probabilities of common ancestry are also functions of geographic structure and migration rates, as well as mutation rates, and, although the former two factors may not make much difference, they were not considered in the calculations. However, this may not matter much if mutation effects occur on a much slower timescale than migration effects (e.g., Nordborg 1997; Fu 1997).

The geographic location of a most recent common ancestor is even more difficult to assess. Cann et al. (1987) found that among sampled populations, those in Africa were most similar to the inferred ancestral mtDNA sequence. Many studies have stated that (provided the data and gene genealogical conclusions are reliable) this means the ancestral sequence and the mitochondrial Eve existed in an African population. However, usually such statements have been made without further comment and are unsupported by any population genetic arguments or models. Recent theoretical developments have shown that the restricted presence of the inferred ancestral gene in specific present-day populations may or may not be closely related to the likelihoods of those populations having originated the ancestral gene, depending on the rates of mutation and migration (Epperson 1999a; 2002). The inferred ancestral gene (if selectively neutral) is in essence a randomly chosen representative of a population, hence relevant theoretical results can be expressed as space–time probabilities of identity by descent, a space–time extension of Malécot's definition of spatial probabilities of identity by descent. Under the conditions of the models (Epperson 1999a; 2002; and chapter 5), the relative values of these probabilities equal the probabilities of origination, because of the simple fact that one of the populations must have contained the ancestor of any given gene at present. Particularly important is how the probabilities of descent depend on the geographic distance between a potential ancestral (origination) geographically located population and the location of a present population. For example, consider Africa and Asia as alternative potential locations of origination, containing the ancestral mtDNA (mitochondrial Eve), and that now only African populations contain the ancestral type mtDNA. What are the *relative* likelihoods that Africa rather than Asia was the location? The two likelihoods depend on the migration rates and mutation rates (Epperson 1999a; 2002), and they should be nearly equal when the amount of migration is high and the rate of mutation is high. In such cases, the location of a gene today (e.g., ancestral type mtDNA in some African populations today) has almost nothing to do with where its ancestor was a long time ago. In other words, the spatial or geographic pattern of genetic variation today contains almost no information about where the origination was, whether one uses haplotype frequencies or gene genealogies based on phylogenetic reconstruction and degrees of differentiation (e.g., among DNA sequences).

For example, in a system with one spatial dimension (appropriate if populations were mostly coastal), and with migration only between nearest neighbor populations with rate l_1 equal to 10 percent, and a per-sequence mutation rate k of 10^{-6}, an ancient (10,000 generations ago) gene from a population located 100 populations away is about 82 percent as likely to share identity as if from the same population itself (figure 3.1). With $l_1 = 0.10$ and $k = 10^{-4}$, a population located 100 populations away has probability 23 percent as large, in contrast to the purely spatial probabilities of identity by descent, which are only 4 percent as large. Moreover, when mutation rates are high the probabilities become very small, suggesting that they would be difficult to estimate from genetic data. The function of relative likelihoods of origination on distance is not quite so flat when migration rates are lower, for example, at 1 percent (figure 3.1), indicating that the amount of migration is critical within the range of 1–10 percent. We do not know what the actual migration rates were among ancient humans. If the system of populations was primarily two dimensional, then even greater flatness may be expected, as occurs for the space–time correlations of gene frequencies (Epperson 1993b).

The above models assume that migration is isotropic, i.e., the rates are the same in both directions. If migration were anisotropic, flowing more out of Africa than into it, it may be that migration causes distant populations to be *more* likely the ancestral source of present

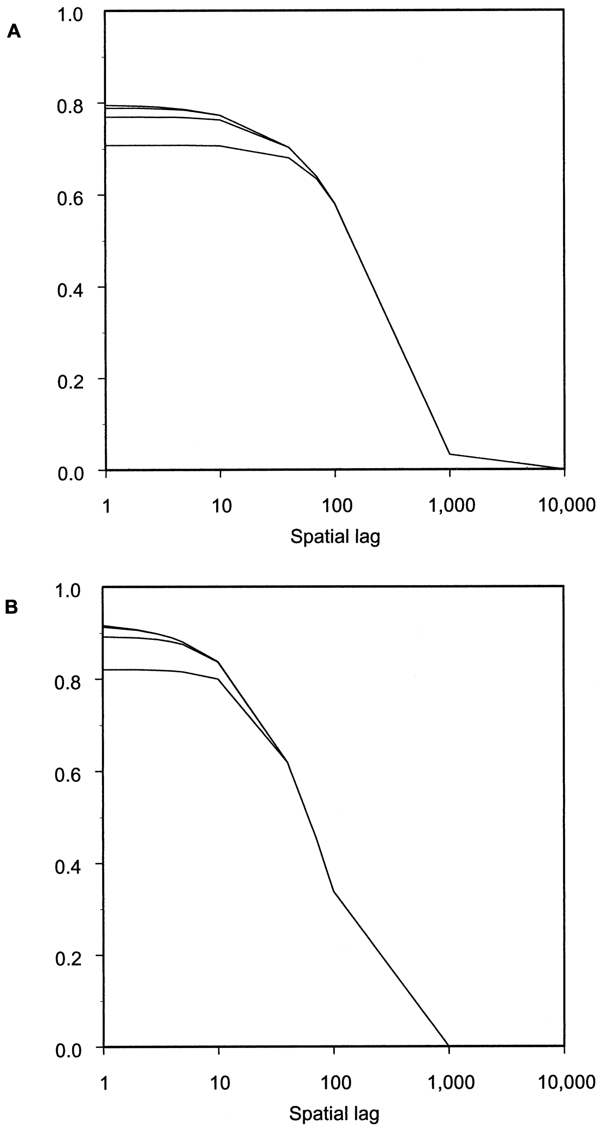

FIGURE 3.1A–D. Space–time probabilities of identity by descent in migration models with different parameter values. All cases are migration models with nonzero isotropic ($l_1 = l_{-1}$) migration only between adjacent populations in a system with one spatial dimension. For each graph, the X axis is spatial distance of separation on a log scale, and there are four curves representing, from top to bottom, time lags of zero (purely spatial

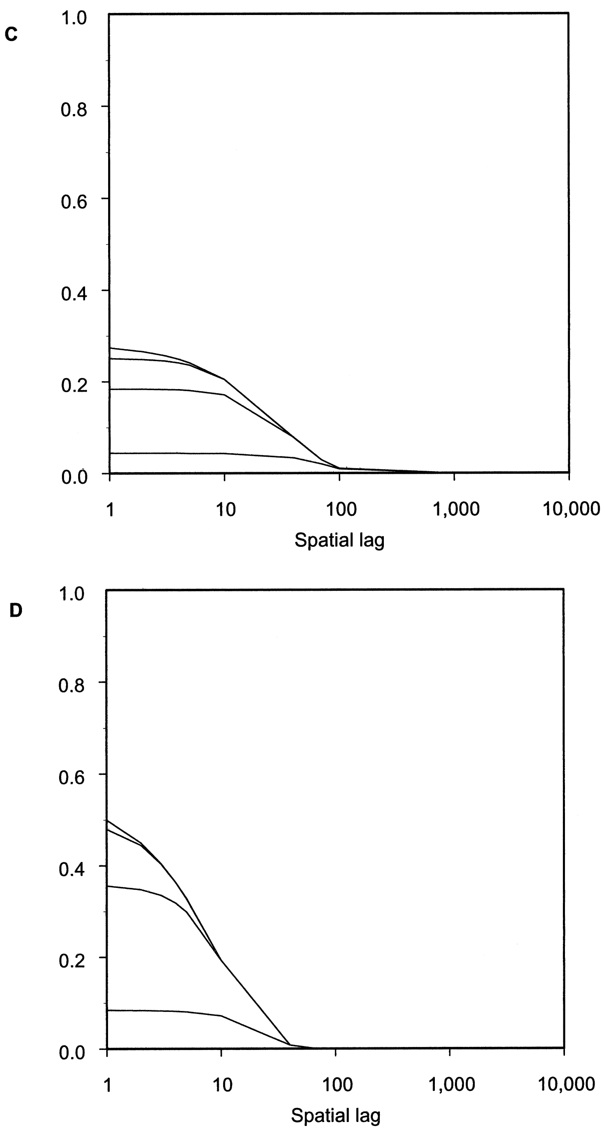

FIGURE 3.1 (*continued*) probabilities), 100, 1000, and 10,000 generations, respectively. The parameters for the four models are (A) $k = 10^{-6}$, $l_1 = 0.1$; (B) $k = 10^{-6}$, $l_1 = 0.01$; (C) $k = 10^{-4}$, $l_1 = 0.1$; and (D) $k = 10^{-4}$, $l_1 = 0.01$. Note that the "flattening" effects of large time lags on the spatial curves are greater in graphs C and D.

variants than the same (or very nearby) populations (Epperson 1993b). In other words, the mtDNA data, particularly the similarity of present day African populations to the mtDNA Eve, could indicate that Africa is *less* likely than other populations (perhaps even Asia) to be the ancestral location of Eve! The anisotropic model has not been examined for space–time probabilities of identity by descent, but there may be reasons to expect them to be similar to the space–time correlations, which can show this effect of anisotropy (Epperson 1993b). However, both models also assume that populations have always existed at a fixed size *N*. Thus their application to the out-of-Africa hypothesis, which implies that populations exist only as they become founded, is clouded. I want to make very clear that I am not arguing here that humans originated in Asia. There are as yet no completely satisfactory theoretical results, but the results do illustrate how difficult inference of origination can be. Such considerations also point to the potential importance of "space–time" genetic data, in particular from ancient DNA samples, and the space–time probabilities of identity by descent or space–time coalescence may provide the probability theory needed to properly utilize such data.

Other evidence for the mtDNA Eve and out-of-Africa theory has been assigned explicitly in terms of details of the geographic distribution of haplotypes together with the degree of differences (reflecting numbers of mutations) among haplotypes and the inferred ages of haplotypes. Haplotypes that have the largest number of inferred mutational changes from the inferred ancestral type are assumed to be more recent. This temporal dimension could in principle contain critical information in an explicitly nonequilibrium context (e.g., Templeton 1998). The geographic distribution may contrast under isolation by distance versus range expansion. If Old World human populations had been stable and exchanging migrations, in a restricted manner, throughout existence, then old haplotypes should be widespread and younger ones less so. Cann et al. (1987) argue that a range expansion, such as is implicit with out-of-Africa, would have caused some ancient haplotypes to remain in the origination area and not spread globally. Having been lost during colonization events and by mutations, they are restricted to the origination areas today. Some of the more recently mutated haplotypes can become widespread if they were not lost during colonizations associated with expansion.

This is, in fact, the pattern observed in the mtDNA data, where Africa is the origination area (Cann et al. 1987). However, there is no basis for this idea from theoretical models, either analytic or computer simulated (Templeton et al. 1995). Theoretical studies of range expansions are in their infancy. Generalities may not be tractable because the pattern may depend on many details about population bottlenecks in the ancestral area, losses of haplotypes in colonization events, intervening migration among established populations, and other details about the sequence of events as populations spread geographically.

The best-developed statistical method tailored to geographic distributions of phylogenies is the nested cladistic analysis of Templeton (1993). The method first constructs a phylogenetic tree for the haplotypes. Such "gene genealogies" may be based on coalescence theory, or other methods such as parsimony. As was noted earlier, this ignores the geographical structure of genetic variation, although this is probably acceptable under many conditions. Next, the unrooted haplotype tree is partitioned into a hierarchical structure, according to the nested cladistic criteria (e.g., see figure 3.2). The most recently mutated haplotypes occur at the "tips" of the tree, the next most recent ones occur at the first nodes in from the tips, and the nodes positioned more toward the interior of the tree represent still older halotypes (Templeton et al. 1995). Among various possible procedures, nested cladistic analysis then uses the tree to first form groups of tip haplotypes and the nearest interior node, i.e., haplotypes that differ by one polymorphic site (one mutation), because each such group has a greater degree of shared ancestry. Then these "one-step clades" are pruned off the tree, and the procedure is repeated for tip haplotypes of the pruned tree. The pruning and forming of one-step clades is repeated until all haplotypes have been classified into one-step clades. The one-step clades are usually still connected by a tree, and the next step is to use them in the same way to classify all one-step clades into two-step clades. This is repeated until the entire tree has been hierarchically classified. Sometimes special rules are needed if there are ambiguities (reticulation) in the tree or symmetries (Templeton et al. 1995). This is a rigorous method, but it may involve some loss of information and it relies on the topology of the tree being true.

The frequencies of the nested clades (sets of related haplotypes) are tested for significant differences among geographic localities. Nested clades can be used to test for the presence of specific geographic patterns. The same could be done for the frequencies of haplotypes themselves, but the use of nested clades may be helpful because it combines data of like type, e.g., haplotypes that share recent ancestry. Templeton et al. (1995) developed a nested contingency table analysis, where geographic location is one of the categorical variables, and the others are clade types nested in hierarchical levels. Exact tests can be constructed, and statistical significance implies that the clade types are not evenly distributed across the geographic locations of populations. However, this inference in itself does not describe the features of the geographic pattern that might be used to distinguish which types of processes have occurred. One way to proceed further is to calculate the geographic distances over which different clades are distributed. Various measures of distance could be used, such as the Euclidean or Great Circle distance measures. For each clade X, the average distance from where each representative of it is found to the geographic center of its range can be calculated, $D_c(X)$ (Templeton et al. 1995). Hence $D_c(X)$ is a measure of how widespread clade X is. This value may be contrasted with the average distance of clade X types from the geographical center of the next higher level clade group of which X is a member, $D_n(X)$. The contrasts may distinguish different processes, including standard genetic isolation by distance versus range expansion. The former could be understood from examination of the space–time probabilities of identity by descent, because it includes information on the relative likelihoods of descendants of an ancestral haplotype being found at varying distances from its origin, in the standard isolation by distance model.

A translation of the predictions of Cann et al. (1987) for haplotypes in expansions into a form for nested clades in expansions was also developed by Templeton and colleagues. Templeton et al. (1995) suggest that a fairly recent range expansion should show large values of $D_c(X)$ and $D_n(X)$ for some but not all tip clades (i.e., wide range expansion for haplotypes that are "young," formed during the range expansion). It should also show small values for some interior nodes, because, according to Cann et al. (1987), some old or ancestral types persist only in a limited area at or near the origin. There may be other details that differentiate contiguous (short-distance) range expansion versus long-distance colonizations, in the expansion process. These patterns contrast those expected for both standard isolation by distance and population fragmentation.

Templeton (1998) demonstrated several advantages of a rigorous statistical method like nested cladistic analysis over graphical methods (e.g., Avise 1994). For example, such methods can be used to study the effects of sampling (Templeton 1998). However, more powerful methods, based on haplotype distributions in explicit space–time models, might be developed in the future. For example, violations in the assumptions of the three steps outlined at the beginning of this chapter could be substantial, and further study is needed to understand the conditions under which they are. Moreover, there may be loss of power. For example, Smouse (1998) has weighed some pros and cons of using phylogenetic trees in biogeographical studies.

Templeton (1993, 1997) reexamined the mtDNA datasets for the out-of-Africa problem using nested cladistic analysis. The results indicated that restricted gene flow, as in standard isolation by distance among long-standing populations existing throughout the Old World, was the prevailing factor. No intercontinental range expansions were evident among Old World populations. In contrast, there was evidence of a much more recent expansion within Europe as well as one associated with colonization of the New World (Templeton 1997).

Many studies argued that the fact that there is greater mtDNA variation in Africa (Vigilant et al. 1991) indicates origination in Africa. The general thinking is that genetic variation was lost through repeated founder effects as new populations were founded and humans spread throughout the rest of the Old World. However, loss of variation in populations from founder effects is temporary if there is gene flow among populations afterward. Theoretical results on the "flattening" of space–time probabilities of identity by descent and space–time correlations (Epperson 1993b), as time goes on, provide an indirect framework for how long differences in diversity among populations should persist. Moreover, the scenario of repeated founder effects is just one of many nonequilibrium processes that can explain the higher level of diversity within African populations. For example, a quite opposite scenario is that the diversity reflects the temporary effects of a "recent" admixture event, such as a major migration back into Africa, as has been suggested based on nested cladistic analysis of Y-chromosome data (Hammer et al. 1998). Finally, it is worth noting that a few samples of DNA from Neanderthal remains (Krings et al. 1997) may provide some support for an expansion throughout Europe during that time period. Space–time probabilities and space–time coalescence probabilities could provide the probability theory needed to more precisely interpret these samples, which are likely to become increasingly important.

Other methods of using modern molecular data for detecting population expansions are based on the frequency distribution of the numbers of pairs of haplotypes that have any given number *i* of mutational differences. Theoretical models of single, unstructured, populations show that if the population has undergone a fairly recent expansion, the frequency distribution of pairwise differences will have a shape (Rogers and Harpending 1992) that differs from the one expected for a population that has had stationary size (Watterson 1975). Rogers and Harpending (1992) showed how the distribution appears immediately after an expansion, and how it changes over the succeeding generations. Let $x_i(\tau)$ be the frequency of *i* pairwise differences, at time τ, measured in units of $\tau = 2\mu t$ (t = number of generations; μ = rate of mutation), expired since an instantaneous population expansion event. This distribution differs markedly from the equilibrium distribution (for any time *t*), where $x_i(t)$ is geometrically decreasing with *i*. Actual distributions for any given gene, DNA sequence, or haplotype may differ substantially because of stochasticity (Slatkin and Hudson 1991). The distribution has a sizable mode at intermediate values of *i* following a sudden increase in population size. The properties of the mode resemble those of a wave, because it tends to move toward the right (increasing values of *i*) as time proceeds further (Rogers and Harpending 1992). Contractions in population size can also cause modes in the distributions (Rogers and Harpending 1992), although the distribution tends to be more “ragged,” with many minor peaks, or less smoothly varying with *i*, compared to that following expansion (Harpending 1994). The value at which the mode occurs depends on the values of θ, pre- and postexpansion (or contraction), where θ is $2N\mu$, and N is either the population size (for haplotype data such as mitochondrial haplotypes), or the number of autosomal genes (for diploid genotypes).

Rogers and Harpending (1992) showed that the distribution of pairwise difference among the worldwide mtDNA samples of Cann et al. (1987), with 147 sampled individuals, exhibits the proper shape. Indeed, it could be closely fitted by sudden expansion, with initial (preexpansion) $\theta = 2.44$, secondary (post-sudden expansion) $\theta = 410.69$, and with a value of $\tau = 7.18$. Using available information on mutation rates, this corresponds to an expansion from about 800–1600 females to 137,000–274,000 about 60,000–120,000 years ago (Rogers and Harpending 1992). These results are reassuring, in part because some sort of expansion to the present size of some six billion should be reflected in the data. Moreover, when mtDNA data on regional populations (Di Rienzo and Wilson 1991), are analyzed separately, the distributions also exhibit modes, including datasets from Sardinia, Japan, the Middle East, and American Indians, all of which have undergone recent expansions (Rogers and Harpending 1992). The overall shape of the worldwide distribution has been argued to support the out-of-Africa theory, again largely because it suggests that the population size during the era in which modern humans evolved was too small to have been maintained across the Old World. Hence widespread archaics could not have contributed to our gene pool (Rogers and Harpending 1992), using the same logic mentioned earlier in this chapter. Note that this argument alone does not specify the geographic location of origination; it is based on a single population model, and hence is not spatially or geographically explicit.

In a geographic setting, the definition of "population" may be confounded. This is a good example of the difficulties in trying to parse out a purely temporal effect (sudden expansion of the total population size of the modern human species) from the more complex spatial–temporal process and context in which it occurred. Such analyses must distinguish overall effective population size (averaged over long periods), from sudden expansions and contractions, as well as the role of migrations and range expansions. Moreover, analyses must consider the complications that may arise from migration among populations and even the definition of population itself. For example, Graven et al. (1995) analyzed a large data set collected from Senegalese Mandenka, and maintained that the distribution of pairwise differences did not have the wave typical of sudden expansion. However, Eller and Harpending (1996) conducted extensive computer simulations of the dataset and found that the data did not reject the possibility of demographic expansion nor did it reject stationarity. They also found that at least the postexpansion population sizes were rather high, and noted that the definition of population from which the sample was taken should be considered to include more than just the Mandenka. They point out that the Mandenka presumably have been exchanging migrants and genes with other regional populations, and hence that West Africa populations as a whole should be considered as the reference population for the samples. Moreover, even if one population, such the Mandenka, has not experienced an expansion, this does not mean that the same is true for other African populations (Eller and Harpending 1996).

Genetic data of a second type, Y-chromosome haplotypes, provides added information on the origin of modern humans, in terms of demographics of males (Hammer and Zegura 1996; Hammer et al. 1997; Underhill et al. 1997). The nonrecombining part of the Y chromosome is the patrilineal counterpart to mtDNA. Harpending et al. (1998) analyzed sequence differences among approximately 20,000 sites from Y chromosomes of 718 sampled individuals (Underhill et al. 1997). They concluded that the distribution of frequencies of polymorphism was consistent with population expansion, as well as inconsistent with that expected from a selective sweep.

Hammer et al. (1998) conducted a nested cladistic analysis of a large data set of Y chromosomes, from 1544 individuals representing 35 populations in total. By obtaining nine polymorphic (diallelic) sites, ten different haplotypes were found. In the first step of their cladistic analysis, the authors used parsimony to reconstruct an unambiguous tree rooted with outgroups of four species of great apes. For the ten Y-chromosome haplotypes, there were five one-step clades, and three two-step clades (figure 3.2). The null hypothesis was strongly rejected for the entire cladogram, as well as for the majority of one-step and two-step clades (Hammer et al. 1998), indicating that there is a spatial pattern of shared ancestry. The most common pattern for Y-chromosome clades was isolation by distance, but there were also three evident range expansions (Hammer et al. 1998). Most importantly, there appeared to have been a range expansion based on some of the oldest clades, because they had small D_c values, and this supported out-of-Africa. Nonetheless, as has been noted earlier, this re-

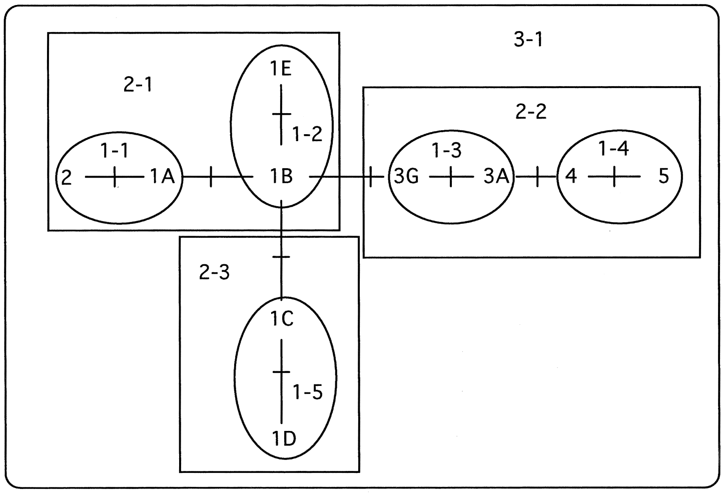

FIGURE 3.2. Nested cladogram for *Y*-chromosome haplotype data. The structure of the cladogram is indicated by lines connecting ten haplotypes, 1A, 1B, 1C, 1D, 1E, 2, 3A, 3G, 4, and 5. There are five one-step clades contained in ovals and denoted 1–1 through 1–5, three two-step clades contained in rectangles and denoted 2–1 through 2–3, and a single three-step clade that contains all ten haplotypes. Original figure provided by M. Hammer (1998).

lies on the unproven presumption that some old haplotypes are likely to persist in the original population in which they arose. Another range expansion was global, as it should be. Perhaps most interestingly, the third expansion was out of Asia and *into* Africa, but one in which Y chromosomes were not completely replaced. This would have caused significant admixture in the African populations, and could help to explain the greater levels of genetic diversity in Africa today. Interestingly, the difference in patterns for Y chromosomes and mtDNA suggests differences in the demographics of males and females.

The limitations of the mtDNA and Y-chromosome data also indicate that conclusive evidence is more likely to come from the nuclear genome. Most of the reported mtDNA variation is limited to a single "hypervariable" segment of about 1 kilobase in size (Brown 1985). The mtDNA pattern and the Y-chromosome pattern are each just one outcome of a process that presumably has been subjected to a large degree of stochasticity. Moreover, both the mitochondrial genome and Y chromosome are, in principle, highly subject to natural selection, and it may be difficult to separate out the effects of any "selective sweeps," for example, large changes in mtDNA haplotype frequencies, perhaps even fixation and losses, due to natural selection (e.g., Knight et al. 1996; Jorde et al. 1997). On the other hand, the population genetics of nuclear genes are in some cases more difficult to model, especially if there is recombination. Statistical methods are also sometimes more complicated. Whereas for haplotypes such as mtDNA a single type is observed per sampled individual, for autosomal loci two genes are sampled from an individual, and they may not be independent (e.g., if there is inbreeding).

Data on isozyme loci and nuclear restriction fragment length polymorphism (RFLP) do not particularly support out-of-Africa. These markers are not more diverse in African populations (e.g., Bowcock et al. 1991; Nei and Roychoudhury 1993), in contrast to mtDNA. However, it has been argued (Jorde et al. 1997) that this may reflect an ascertainment bias, because many of the markers were developed as part of a search for polymorphic markers among European populations.

One of the first extensive nuclear DNA sequence datasets collected was for the HLA major histocompatibility complex. In comparison to most of the other sequence datasets, HLA has very high levels of polymorphism, and the inferred phylogenetic tree is deep. It appears that many alleles are very old, and that the coalescences of some alleles occurred on the order of millions of years ago (Ayala 1995; Erlich et al. 1996), and hence many alleles were segregating in the founding population(s) of modern humans. Correspondingly, the inferred long-term effective population size is large, on the order of 100,000 (Takahata 1993). Sherry et al. (1997) argue that because the time frame includes a long period before the formation of modern humans, when population sizes may have been large relative to the founding population of modern humans in the out-of-Africa scenario, this is not inconsistent with out-of-Africa. However, Ayala (1995) argued that the data suggest that the population size was around 100,000 during the period in which modern humans arose. This was countered by an argument that a smaller effective population size (10,000), together with a certain level of diversifying selection, could also explain the data (Erlich et al. 1996). Nonetheless, the possibly confounding role of diversifying selection makes conclusions from the HLA data difficult.

Extensive nuclear data are available in the form of polymorphisms for *Alu* retroposable elements. More than 500,000 *Alu* exist in families of up to 500–2000 mostly noncoding (Deininger and Batzer 1993) and presumably selectively neutral elements. Some elements have inserted recently enough that not all humans carry an element at a given insertion site, and other inserted elements are fixed, although they vary in sequence because of mutations (Batzer et al. 1996, 1997). Knight et al. (1996) surveyed 29–60 worldwide individuals for three insertions that had varying levels of nucleotide diversity. One insertion had a diversity value that, by comparison to nucleotide divergence of humans from apes, suggested an average age of sequence divergence in the range of 30,000 to 55,000 years ago. Knight et al. (1996) found that the data fit more closely a one-population model, which reflects the out-of-Africa hypothesis, than they fit a model of two isolated populations that diverged 1.5 million years ago. However, there are many possible forms of the multiregional model, and migration could affect the conclusions.

Sherry et al. (1997) examined 13 *Alu* elements from 122 individuals. The ratio of dimorphic (not present in all individuals) elements to elements fixed in humans but absent from apes, can be used to estimate effective population size, under the assumption that an element is never lost once it becomes inserted in a new chromosomal location. However, it is not clear that this assumption is consistent with loss through genetic drift. Sherry et al. (1997) estimated that the population size is about 18,000, effective over the past one to two million years, and they showed that the data are consistent with a population bottleneck and subsequent expansion. However, other scenarios are possible. There does not seem to be any reason to assume that the insertions occurred simultaneously with the population isolation (and contraction in size) associated with the speciation in Africa (Stanley 1997). Again, the strongest argument is simply that 18,000 is too small to include the archaics that had spread throughout the

Old World (Harpending et al. 1998). Finally, in these studies there was no particular evidence that the founder population was in Africa.

A recent study of diversity of microsatellites does suggest that African populations have slightly greater nuclear genomic diversity (Jorde et al. 1997). Among 60 unlinked microsatellite loci from 255 individuals sampled from all over the Old World, average heterozygosity was slightly greater in Africans ($H = 0.76$) than in Asians (0.70) or Europeans (0.73), but the difference was not statistically significant. When allele size was taken into account, reflecting multiple mutations in a stepwise fashion, Africans had statistically significant greater variances. Moreover, the contrasts were greater for those microsatellite loci that had overall smaller variances in allele sizes. This suggests that loci with higher variance also had higher mutation rates, and thus the pattern created by expansion has been erased to a greater degree for the more variable loci. While high mutation rate tends to cause greater spatial patterning, in terms of the slope of the isolation by distance function, it also causes all spatial correlations to be small, making detection of genetic patterns, especially from the distant past, more difficult (Epperson 2002). Other explanations, such as recent admixture in Africa, are also possible.

The above discussions do not mean that the out-of-Africa hypothesis is incorrect. Perhaps there are just too many coincidences in the datasets, each of which points in some way toward out-of-Africa. This data-rich example illustrates the statistical and stochastic issues in inferring events from the distant past in a spatial–temporal context. It is also important to note that this subject, like many subjects in human genetics, is of inherent interest to many people. There can be little doubt that much more data, much of it sequence data from the nuclear genome, will be generated, and to some degree this may offset the stochasticity inherent in such processes. Much larger datasets may detect even the faintest of signals, which we may conclude become very faint indeed over such lengths of time. This also means that we need precise models of the processes, under out-of-Africa and competing theories, together with optimal points of comparison. Finally, the importance of the role of ancient DNA samples is yet to be understood, particularly, how many of such samples are required to give representations of past populations at a point in time and space.

PATTERNS CAUSED BY MORE RECENT EXPANSIONS OF HUMAN POPULATIONS

Spatial correlation and trend surface analysis indicate that there are both isolation by distance and clines for some genetic markers in Europe (Sokal and Wartenberg 1981; Bocquet-Appel and Sokal 1989), and, as was discussed in chapter 2, particularly the latter may have been caused by a population expansion from the southeast. Sokal (1988) analyzed the spatial patterns for 107 genetic variables, grouped into 27 systems, at over 3000 locations in Europe. He used a multivariate extension of the Mantel test statistic for the association of one matrix of genetic distances with several other matrices of distances (Smouse et al. 1986). In Europe, the matrix of genetic distances among samples, GEN, is highly correlated with geographic distance, GEO, as expected for any kind of genetic isolation by distance. However, GEN was also highly correlated with linguistic differences, LAN, and, most importantly, there remains a positive partial correlation of GEN and LAN, conditioned on GEO. This shows that language, which tends to have a branching structure often with known polarity, is a major factor in traces of genetic heritage, and that "language barriers" among populations can reduce gene flow between them (Sokal 1988).

Sokal et al. (1992) examined a similar dataset further, to test some specific theories about the nature of the expansion, in particular whether it was associated with three waves of Kurgan migrations or with the immigration of Indo-Europeans together with agriculture. Using Mantel procedures, they constructed two additional distance matrices that specified the relevant features of the two hypotheses. Both hypotheses failed to cause the multiple regressions to fit significantly better, and hence neither help to explain the spatial patterns seen today. Nonetheless, this analysis is a good example of how various hypotheses as well as external information about specific space–time processes may be included and tested in proper statistical models, without constructing phylogenetic trees (Smouse 1998). The results suggest that expansion or demic diffusion of Indo-Europeans (Sokal et al. 1991) is not necessarily closely tied to the spread of agriculture. Sokal et al. (2000) showed that cancer incidences are also associated with genetics and ethnohistory in Europe.

Recently, spatial autocorrelation analysis called AIDAs (autocorrelation indices for DNA analysis [Bertorelle and Barbujani 1995]) was conducted on a large sample (>2600) of mtDNA sequences in Europe, the Near East, and the Caucasus (Simoni et al. 2000). It was concluded that Paleolithic expansion and a Neolithic diffusion of agricultural communities may be responsible for the north–south clines in genetic differentiation of mtDNA diversity, although the pattern for nuclear genes reflects a somewhat more complex process. This appears to fit with Templeton's (1997) nested cladistic analysis of Old World mtDNA, in which the only range expansion detected was a relatively recent (compared to the scale of out-of-Africa) expansion within Europe. In contrast, the Y-chromosome data for Europe was dominated by isolation by distance (Hammer et al. 1998), suggesting that demographic diffusion of women was more important than that of men in the expansion in Europe, although one Y-chromosome haplotype does seem to reflect demic diffusion (Ammerman and Cavalli-Sforza 1984) and expansion in Europe between 5000 and 10,000 years ago (Hammer et al. 1997).

Similarly, other fairly recent range expansions have been detected, using nested cladistic analyses, in the settlement of remote islands in the Pacific (Sykes et al. 1995), and of Siberia and North America (Torroni et al. 1993a,b). In all of these cases there was strong independent evidence of expansions, and thus these genetic studies provided a good test of the ability to detect range expansions with genetic data. The success, in considerable contrast to the out-of-Africa hypothesis, may be related to the fact that the posited events are much more recent. They should be more detectable, according to general considerations in the theory based on space–time probabilities of identity by descent and coalescences (Epperson 1999a; 2002). As time proceeds the signature left in spatial patterns declines, especially if migration rates are relatively high (Epperson 1999a).

PATTERNS IN OTHER SPECIES

The ability to detect the temporary effects of ancient events in a spatial–temporal context is well documented for many species other than human. For example, Templeton (1998) reviewed studies on mtDNA in 11 animal species or subspecies. In every case there was prior knowledge that the species had fairly recently expanded their ranges, and in all cases but one expansion was indicated in the present geographic pattern of genetic variation, as examined with nested cladistic analyses (Templeton 1998). Total sample sizes ranged from 34 to 613 individuals and with 4–58 distinct haplotypes. In three cases, two subspecies of darter, *Etheostoma blennoides blennoides* and *E. b. pholidotum* (Wilson 1997), and the gopher, *Geomys bursarius*, the species are North American with current ranges that include large areas that were under glaciers during the Pleistocene. They expanded ranges northward since the Wisconsinian glaciation, which reached glacial maximum about 18,000 years ago. Similarly, two North American subspecies of salamander, *Ambystoma tigrinum tigrinum* and *A. t. mavortium* (Templeton et al. 1995), have expanded into areas that while not glaciated were nonetheless uninhabitable during the time of glaciation. Range expansions were detectable in all five cases, fragmentation was also detected for the darters, and a regional colonization event was detected for the gopher. Similarly, the lichen grasshopper, *Trimerotropis sáxatalis*, expanded into the Ozarks from the southwest and this expansion was followed by a fragmentation event (Gerber 1994). Range expansion and fragmentation were also detected in the fish *Galaxias truttaceus* in colonizing lakes created by melting Pleistocene glaciers in Tasmania (Templeton 1998).

Expansions that are very recent were also detected (Templeton 1998). Both contiguous (gradual) expansion and (long-distance) colonization of *Drosophila melanogaster* were detected in the recent global expansion from Africa. Colonization of the macaque monkey, by Portuguese sailors in the 1500s, was detected on the Island of Mauritius, and the Phillippines and/or Indonesia were identified as the possible sources. Similarly, range expansion of *Canis latrans* in North America since around 1900 was detected. Among the examples, only the expansion of *Drosophila buzzatii* in Europe, from South America, was not detectable. Apparently, an extreme bottleneck occurred in the initial Iberian colonization, where only one mtDNA haplotype was found (Rossi et al. 1996), and this haplotype was also the most frequent one in South American flies, as well as being interior in the phylogenetic tree. Moreover, mutation has not created any new, widespread haplotypes in Europe, perhaps because the haplotypes were defined by restriction site variations, which have low mutation rates. Hence geographically widespread young haplotypes (tip clades) are not observed (Templeton 1998).

Most examples of successful identification, whether through the use of phylogenetic trees or spatial analysis without trees (see Templeton 1998; Smouse 1998; Goldstein and Harvey 1999), of events of population expansion and fragmentation are for those events that are relatively recent. Such events are essentially by definition temporary, nonequilibrium processes. We may expect that the signatures in geographic patterning are diminished for events from very long ago. On the other hand, if events are too recent, and there is insufficient time for mutation to create new variants, not all of ancient features will be manifest in geographical genetic patterns, and different models and expectations must be purported. Finally, some aspects of the inference of origins of polymorphism, ancestral sources of variation, or the location of origin of a species itself, are very clearly manifest and testable. These are sometimes contradicted in present patterns of genetic variation. For example, in red pine, *Pinus resinosa*, the geographic region that has by far the greatest levels of chloroplast DNA diversity (Echt et al. 1998; Walter and Epperson 2001) simply cannot be the center of origin (of postglacial populations), because this area was buried by the Wisconsin glacier. The most likely explanation is that this area is recently admixed between two or more lineages, derived ultimately from separate refugial populations. Nonetheless, this violates an assumption often made, that centers of diversity are also centers of origination, and it points out that admixture can also be an important ancient event.