Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history

ALAN R. TEMPLETON

Department of Biology, Washington University, St Louis, Missouri 63130-4899, USA

Abstract

Since the 1920s, population geneticists have had measures that describe how genetic variation is distributed spatially within a species’ geographical range. Modern genetic survey techniques frequently yield information on the evolutionary relationships among the alleles or haplotypes as well as information on allele frequencies and their spatial distributions. This evolutionary information is often expressed in the form of an estimated haplotype or allele tree. Traditional statistics of population structure, such as \( F \) statistics, do not make use of evolutionary genealogical information, so it is necessary to develop new statistical estimators and tests that explicitly incorporate information from the haplotype tree. One such technique is to use the haplotype tree to define a nested series of branches (clades), thereby allowing an evolutionary nested analysis of the spatial distribution of genetic variation. Such a nested analysis can be performed regarding the geographical sampling locations either as categorical or continuous variables (i.e. some measure of spatial distance). It is shown that such nested phylogeographical analyses have more power to detect geographical associations than traditional, nonhistorical analyses and, as a consequence, allow a broader range of gene-flow parameters to be estimated in a precise fashion. More importantly, such nested analyses can discriminate between phylogeographical associations due to recurrent but restricted gene flow vs. historical events operating at the population level (e.g. past fragmentation, colonization, or range expansion events). Restricted gene flow and historical events can be intertwined, and the cladistic analyses can reconstruct their temporal juxtapositions, thereby yielding great insight into both the evolutionary history and population structure of the species. Examples are given that illustrate these properties, concentrating on the detection of range expansion events.

Keywords: coalescence, colonization, fragmentation, gene flow, haplotype tree, population structure, range expansion

Introduction

Starting with Wahlund (1928), population geneticists have realized that genetic survey data can reveal information about population subdivision. Wright (1931, 1943) introduced \( F \) statistics as a way of utilizing allele frequency data gathered in different geographical locations to quantify population subdivision and estimate the amount of gene flow. However, modern genetic surveys using restriction site or DNA sequence data also provide information on the evolutionary relationships of the genetic variation being scored, which is often portrayed as an allele or haplotype tree. Consequently, information is now available about the alleles’ existence through evolutionary time as well as geographical space. Can this new temporal information be used to shed more light upon the spatial distribution of current allelic variation? The purpose of this article is to answer this question by worked examples that make use of the evolutionary time dimension provided by haplotype trees.

Traditional \( F \) statistics do not use temporal information on allelic variation, but several new statistical procedures can make use of haplotype trees (Hudson et al. 1992; Slatkin 1989, 1993; Slatkin & Maddison 1989, 1990; Templeton 1993; Templeton et al. 1995; Templeton & Georgiadis 1996). This article provides worked examples demonstrating that these new approaches have enhanced
power over \( F \) statistics, allow greater precision of gene-flow estimation, and can separate population structure (recurrent forces such as gene flow) from historical events (fragmentation and range expansion events). Because the detection of range expansion events has proven to be particularly controversial (Templeton 1993, 1994, 1996a; Ayala 1995), the validity of the criteria used to detect range expansions given in Templeton et al. (1995) is examined by applying their methodology to several data sets for which strong a priori evidence exists of range expansions. Finally, it is shown that the method of Templeton et al. (1995) does not merely identify and geographically localize the various factors influencing the spatial distribution of genetic variation but it also estimates the dynamic structure and temporal juxtaposition of these evolutionary factors.

**Detecting and estimating restricted gene flow**

Wright (1931) showed that there is a nonlinear relationship between the amount of gene flow and the degree of genetic differentiation among subpopulations as measured by \( F \) statistics. For example, under the island model in which a population is subdivided into a large number of local populations of size \( N \) with a proportion \( m \) of each population dispersing at random over all local populations, the expected \( F_{ST} \) value (the ratio of the observed variance of allele frequencies across local populations to the theoretical maximum variance) is:

\[
F_{ST} = \frac{1}{4Nm + 1} \tag{1}
\]

From eqn 1, \( F_{ST} \) quickly approaches 0 as \( Nm \) (the effective number of migrants) increases. This means that once \( Nm \) exceeds a value of 4 or 5, there is little effect on \( F_{ST} \) values even with large changes in \( Nm \). Worse, if the estimated \( F_{ST} \) is not significantly different from 0, the nonlinearity from eqn 1 ensures that not even the order of magnitude of \( Nm \) could be estimated. As an example, consider a restriction site genetic survey of the alcohol dehydrogenase locus (Adh) of 39 lines of *Drosophila melanogaster* sampled at four localities in the eastern half of the USA (Aquadro et al. 1986). The algorithm of Davis et al. (1990) was used to estimate \( F_{ST} \) as 0.030, which was not significantly different from zero. The only statistical conclusion one could make about \( Nm \) is that it was greater than 2 with a probability of 95%.

Note that in the Adh example, \( Nm \) could be 2, 20, 200, or 200 000 000, and there is no way the traditional \( F_{ST} \) measurement can distinguish among these alternatives. However, a haplotype tree can be estimated from the restriction site variation found at the Adh locus (Aquadro et al. 1986), and this haplotype tree can be converted into a nested series of clades (branches) by using the nesting rules given in Templeton et al. (1987) and Templeton & Sing (1993). Basically, these nesting rules start at the tips of the haplotype network and move one mutational step into the interior, uniting all haplotypes that are connecting by this procedure into a ‘1-step clade.’ After pruning off the initial 1-step clades from the tips, this procedure is then repeated on the more interior portions of the haplotype network if needed until all haplotypes have been placed into 1-step clades. The next level of nesting uses the 1-step clades as its units, rather than individual haplotypes. The nesting rules are the same, but result in ‘2-step clades’. This nesting procedure is repeated until a nesting level is reached such that the next higher nesting level would result in only a single category spanning the entire original haplotype network. The resulting nested clades are designated by ‘C–N’ where ‘C’ is the nesting level of the clade and ‘N’ is the number of a particular clade at a given nesting level. Some special nesting rules are needed to deal with symmetries and ambiguities in the estimated haplotype network (Templeton & Sing 1993).

The resulting nested set of clades for the Adh haplotype tree is shown in Fig. 1, along with the geographical distributions of the various haplotypes found in the survey. These nested series of branches constitute an evolutionary-based statistical design that was originally used for investigating the relationship between genotype and phenotype (Adh activity in this case; Templeton et al. 1987). This nested design can also be used to look for geographical associations in two ways (Templeton et al. 1995). Only the more simple of the two approaches is considered in this section. The simple procedure is a nested contingency analysis (Templeton & Sing 1993) in which each geographical location is regarded as a categorical variable. Because geographical distance is ignored, this approach is somewhat of an analogue to the island model. To see how the nested contingency analysis is implemented, consider nesting clade 2–1. Clade 2–1 contains two nested clades within it: 1–1 and 1–2 (Fig. 1). Clade 1–1 consists of one haplotype (haplotype 1, Fig. 1) found in only one line sampled from Kansas and a second haplotype (haplotype 2, Fig. 1) found in two lines collected in Wisconsin. Clade 1–2 includes six lines from Rhode Island and one from Wisconsin. An exact 2×3 permutational contingency test (2 clades vs. 3 geographical locations) is performed to test the null hypothesis of no association of clades with geographical location. In this case the null hypothesis is rejected with an exact probability level of 0.033. Similarly, the nested contingency analysis of nesting clade 3–1 (Fig. 1) which contains clades 2–1, 2–2, and 2–3 is significant with a null probability of 0.028. These contingency tests are repeated on all nesting clades containing more than one nested clade and that were found at more than one location. No other nested contingency tests were significant at the 5% level. Hence, a geographical association is detectable when the haplotype

Given that a significant geographical association has been detected, and assuming for the moment that the association is due to restricted gene flow, it is now meaningful to estimate $N_m$, the effective number of migrants per generation among the geographical locations. An estimation procedure that makes use of the gene tree is given by Slatkin & Maddison (1989). Their procedure does not test whether there is a statistically significant geographical association. Consequently, the Slatkin & Maddison (1989) procedure should only be used when a significant geographical association has been detected; otherwise, the resulting estimator is of dubious statistical validity (Templeton et al. 1995). Moreover, the Slatkin & Maddison (1989) estimation procedure should only be used when the cause of the geographical association is inferred to be due to restricted gene flow; otherwise, the resulting estimator is biologically misleading. More on this point will be given in the next section. For now the assumption is made that the significant associations detected by the nested clade analyses are due to restricted gene flow. Applying the Slatkin & Maddison (1989) algorithm to the tree given in Fig. 1 yields an estimate of $N_m$ of 5.4, with a 95% confidence interval of 2.0–19.6. Recall that the insensitivity of the traditional $F_{ST}$ statistic to large values of $N_m$ precluded any inference on the possible upper bound of the $N_m$ value. Hence, the phylogenetic approach has resulted in far greater precision in estimating $N_m$ than is possible from a traditional $F_{ST}$ analysis.

Any estimate of $N_m$ based upon a single locus should be regarded as preliminary because much evolutionary stochasticity is associated with the variation at any given locus (Ewens 1983), and because locus-specific forces (such as selection) can distort the apparent $F_{ST}$ or $N_m$ value (Lewontin & Krakauer 1973). Langley et al. (1988) studied the same lines of $D. melanogaster$ analysed above for a different locus, the duplicated amylase (Amy) locus. As with Adh, there is no significant $F_{ST}$ for the Amy locus. The Amy region has been subject to a nested contingency analysis for isozyme associations (Templeton & Sing 1993), and the same nested design is now used for geographical associations. Unlike the Adh locus, however,
Discriminating between recurrent gene flow and historical events

The major limitation of the above analysis was that it assumed that the significant association between haplotype variation and geographical location was due to restricted gene flow. However, suppose a species has been fragmented into two or more subpopulations that experience no gene flow at all. If they had a recent shared ancestry, the populations could still display some genetic similarity that would yield a traditional estimate of $F_{ST} < 1$, thereby erroneously implying nonzero gene flow. Alternatively, suppose the species recently expanded its range over a large area from some smaller subpopulation within the ancestral range. Then there would be much genetic similarity over this expanded range, leading to an overestimate of gene flow. Using $F$ statistics or an algorithm that assumes that all geographical associations are due to gene flow (e.g., Slatkin & Maddison 1989) can therefore yield an estimator of $N_{m}$ that is biologically misleading.

Fortunately, this potential confoundment of population structure with population history can be investigated by using haplotype trees. Indeed, the primary advantage of using the haplotype tree information is not the quantitative advantage of enhanced power and precision; rather, it is the qualitative advantage of discriminating among various biological explanations for any detected geographical association. To show that haplotype trees can discriminate among cases that appear identical to the nonhistorical $F$-statistic analyses, consider the study of Templeton & Georgiadis (1996) on mitochondrial DNA (mtDNA) restriction site variation in Eastern African populations of buffalo (*Syncerus caffer*) and impala (*Aepyceros melampus*). The $F$-statistic estimator of Davis et al. (1990) yields an $F_{ST}$ of 0.08 for the buffalo and 0.10 for the impala. Both of these $F_{ST}$ values are significantly different from zero, but they are not significantly different from each other. Moreover, in both species most of the geographical sites surveyed are relatively close together in Kenya and Tanzania, but one site (Chobe) is far to the south. In both species, the Chobe samples had many haplotypes not found in the other locations, and it was the Chobe samples that were primarily responsible for the significant $F_{ST}$ values in both cases. Hence, this $F$-statistic analysis implies that both species are equally subdivided, have comparable rates of gene flow, and display restricted gene flow primarily between the Chobe vs. Kenya/Tanzania localities.

Templeton & Georgiadis (1996) also estimated haplotype trees for the mtDNA, as shown here in Fig. 2. Figure 2 also indicates the haplotypes found only in Chobe in both species. Even a cursory glance at Fig. 2 reveals that the pattern of distribution of the Chobe-only haplotypes in the haplotype trees are completely different in these two species which appear indistinguishable by $F$-statistic analysis. In the buffalo, the Chobe haplotypes are scattered throughout the haplotype network; in the impala, the Chobe haplotypes are tightly clustered in the haplotype network. Although both species show the same degree of spatial subdivision as measured by $F_{ST}$, they have obviously achieved this degree of subdivision in very different fashions through time. Clearly, the use of haplotype networks allows a finer discrimination of biological pattern than an $F$-statistic analysis. The reason for this is straightforward; by using a haplotype network, one is examining a spatial/temporal pattern of genetic variation whereas with the $F$-statistic and other nonhistorical analysis one can only examine the current spatial pattern.

The scattered spatial/temporal pattern found in the buffalo (Fig. 2) indicates recurrent genetic interchange between Chobe and the more northerly populations throughout the time period from the coalescence of mtDNA to the present. The impala pattern is more difficult to interpret. Such a strong evolutionary clustering of haplotypes in a geographical region, particularly when the haplotype clusters are separated by a long branch length with missing intermediates, is often interpreted as evidence of a past fragmentation event (see Avise 1994 and references therein for examples). However, because impala are found in intermediate geographical locations that were not sampled, it is possible that this pattern arose from isolation-by-distance (i.e., a restricted gene flow model rather than an historical event) such that geographically intermediate populations would fill in the
missing haplotype nodes and show a gradual shift from one cluster of haplotypes to the other. Indeed, a rigorous quantitative analysis of these data (of the type to be discussed below) reveals that the sparseness of sampling prevents one from distinguishing between isolation-by-distance vs. fragmentation of the Chobe population from the Kenyan/Tanzanian populations (Templeton & Georgiadis 1996).

The analysis of Templeton & Georgiadis (1996) on the impala illustrates the dangers of making biological inferences simply by a visual inspection of how geography overlays upon a haplotype tree. Such visual inferences are commonplace in the phylogeographic literature (Avise 1994 and references therein), but they make no assessment of adequate sample sizes for statistical significance nor adequate sampling of geographical locations for distinguishing among potential causes of geographical associations. What is needed is an objective statistical analysis that first rejects the null hypothesis of no association between haplotype variation and geography, and then interprets the statistically significant patterns using explicit criteria that include an assessment of sampling adequacy. To accomplish this task, Templeton et al. (1995) have proposed a quantitative analysis of geographical data using the same nested design generated by the haplotype network that was used in the contingency analyses. As this technique and its inference criteria are discussed at length along with a detailed worked example in Templeton et al. (1995), only a brief summary will be given here.

The geographical data are quantified in two main fashions: the clade distance, $D_c$, which measures the geographical range of a particular clade; and the nested clade distance, $D_n$, which measures how a particular clade is geographically distributed relative to its closest evolutionary sister clades (i.e. clades in the same higher-level nesting category). In particular, the clade distance measures the average distance that an individual bearing a haplotype from the clade of interest lies from the geographical centre of all individuals bearing haplotypes from the same clade. The nested clade distance measures the average distance that an individual bearing a haplotype from the clade of interest lies from the geographical centre of all individuals bearing haplotypes from the next higher-level nesting clade that contains the clade of interest. Contrasts

in these distance measures between tip clades (clades that are not interior nodes in the haplotype tree) and the clades immediately interior to them in the cladogram are important in discriminating the potential causes of geographical structuring of the genetic variation (Templeton et al. 1995), as will be discussed later. The statistical significance of the different distance measures and the interior-tip contrasts are determined by random permutation testing which simulates the null hypothesis of a random geographical distribution for all clades within a nesting category given the marginal clade frequencies and sample sizes per locality.

If statistically significant patterns are detected, they then need to be interpreted biologically. Templeton et al. (1995) consider three major biological factors that can cause a significant spatial/temporal association of haplotype variation. The first factor is restricted gene flow, particularly gene flow restricted by isolation-by-distance (Wright 1943). Because restricted gene flow implies only limited movement by individuals during any given generation, it takes time for a newly arisen haplotype to spread geographically. Obviously, when a mutation first occurs, the resulting new haplotype is found only in its area of origin. With each passing generation, a haplotype lineage that persists has a greater chance of spreading to additional locations via restricted gene flow. Hence, the clade distances should increase with time under a model of restricted gene flow. If an outgroup can be successfully used to root the haplotype tree, any series of nested clades can be polarized temporally in an unambiguous fashion. However, often intraspecific haplotype trees cannot be rooted reliably by the outgroup method or other standard rooting procedures (Templeton 1993; Castelloe & Templeton 1994). Fortunately, in a nested series of clades, a nesting clade has to be as old or older than all the lower level clades nested within it. Hence, as nesting level increases, there is a nondecreasing age series even when the root is not known. Accordingly, the clade distances are expected to increase with increasing nesting level. This expected increase will continue until either the highest nesting level is reached or, if the gene flow is sufficiently high relative to the coalescent time of the haplotype tree, a nesting level will be reached in which the clades are uniformly distributed over the entire sampled geographical range, and all higher nesting levels will replicate that pattern. Another aspect of the expected patterns under restricted gene flow is that when a mutation occurs to create a new haplotype, that new haplotype obviously resides initially within the range of its ancestral haplotype. As the ancestral haplotype is older than its mutational offshoot, it should have a wider geographical distribution. Therefore, when the new haplotype starts spreading via gene flow, it will often remain within the geographical range of its ancestor for many generations, particularly under an isolation-by-distance model. Because there is a strong tendency for the ancestral haplotypes to be immediately interior to the derived haplotypes in terms of the topology of the haplotype network (Castelloe & Templeton 1994), this means that there will be a strong tendency under restricted gene flow for tip clades to have a geographical range smaller and often nested within the range of the clades that are immediately interior to them. Moreover, because the ancestral haplotype is expected to be most frequent near its site of geographical origin, most mutational derivatives of the ancestral haplotype will also occur near the ancestral site of geographical origin. This means that the geographical centres of all the clades nested together should be close; hence, the clade distances and nested clade distances should show similar patterns under restricted gene flow.

A second factor is past fragmentation events. When the nesting level reaches the temporal period at which the fragmentation event occurred, the clade distance cannot increase beyond the geographical ranges of the fragmented subpopulations, but the nested clade distances will generally show a marked increase when the fragmented clades are allopatric, as is typically the case. If the fragmentation event is an old one relative to the rate at which mutations accumulate, the branch lengths between the clades displaying large nested clade distances but plateaued clade distances will tend to be longer than the average branch length in the tree (due to the accumulation of mutations that differentiate the fragmented subpopulations).

Range expansion (including colonization) is the third factor that can create a geographical association with the haplotype network. When range expansion occurs, those haplotypes found in the ancestral population(s) that were the source of the range expansion will become geographically widespread (large clade distances), and the distinction between the relative geographical ranges of tip vs. interior clades expected under restricted gene flow breaks down or can even be reversed. Moreover, some of the haplotypes found in the expanding populations can become quite distant from some of the older haplotypes that are confined to the ancestral, pre-expansion area (large nested clade distances), particularly when long-distance colonization is involved. As mutations first start to accumulate in the colonizing population, they will be tips with large nested clade distances because the interior haplotypes from which they mutated will also be found in the ancestral range.

The procedure of Templeton et al. (1995) first limits inference to those clades showing statistically significant geographical associations. Next, the patterns displayed by these significant associations are evaluated relative to the above expectations. In order to make this pattern evaluation explicit and consistent, an inference key is provided as an appendix to Templeton et al. (1995) (reproduced here as

© 1998 Blackwell Science Ltd, Molecular Ecology, 7, 381–397
an appendix and hereafter referred to as the inference key). This quantitative geographical nested analysis and inference key should be used before using an estimation algorithm such as that of Slatkin & Maddison (1989) in order to ensure that the biological situation being examined corresponds to the assumptions of the estimation algorithm. Subjecting the *D. melanogaster* Adh data to such an analysis reveals a significantly small tip clade distance nested within clade 3–1, and application of the inference key led to the conclusion of restricted gene flow. Hence, it is meaningful to estimate *Nm* in this case, and the estimated *Nm* = 5.4 that was given earlier indicates that these eastern USA populations of *D. melanogaster* experience relatively high levels of recurrent gene flow. The two Amy subregions were also subjected to this quantitative distance analysis even though neither subregion displayed significant geographical associations in the categorical contingency analysis. No significant effects were detected in the right subregion, but significant effects were detected within nesting clade 1–1 and among clades 2–1 and 2–2 nested within the total haplotype network (see Fig. 3 in Templeton & Sing 1993). Application of the inference key given in the appendix led to the conclusion of restricted gene flow within 1–1, and an inconclusive outcome at the total network level. Hence, the same statistically significant qualitative conclusions were reached for both loci. However, for the low-resolution Amy haplotype tree, all the interior haplotypes are so widely distributed geographically that it is impossible to use the estimation procedure of Slatkin & Maddison (1989). All that one can conclude from the Amy locus is that the level of recurrent gene flow among these populations must be high even though somewhat restricted by isolation by distance, a conclusion also compatible with the Adh results.

Although only one cause of nonrandom spatial distributions of clades was inferred for the Adh and Amy examples, the nested analysis of Templeton et al. (1995) searches for multiple, overlaying patterns within the same data set. For example, in the analysis of mtDNA restriction site variation in the salamander *Ambystoma tigrinum* given in Templeton et al. (1995), an historical fragmentation event is inferred between two named subspecies followed by independent range expansion within each subspecies, overlaid upon a pattern of isolation by distance occurring within each subspecies. There is nothing about the evolutionary factors of restricted gene flow, fragmentation events, or range expansion events that make them mutually exclusive alternatives. One of the great strengths of this inference procedure is that it explicitly searches for the combination of factors that best explains the current distribution of genetic variation and does not make a priori assumptions that certain factors should be excluded or be regarded as unlikely. Moreover, by using the temporal polarity inherent in a nested design (or by outgroups when available), the various factors influencing current distributions of genetic variation are reconstructed as a dynamic process through time. For example, the analysis of the *A. tigrinum* mtDNA data reveals that the fragmentation event occurred prior to the expansion events (the expansion events were inferred in clades nested within the clade detecting the fragmentation event, as shown in Table 3 of Templeton et al. 1995). Moreover, the inference of restricted gene flow via isolation-by-distance is found in clades that nest and are nested within the clade leading to an inference of an expansion event, thereby implying that isolation-by-distance characterized the salamanders’ population structure both before and after the expansion event. Hence, this procedure does not merely identify and geographically localize the various factors influencing the spatial distribution of genetic variation, but rather it brings out the dynamic structure and temporal juxtaposition of these evolutionary factors.

The inference key also incorporates the types of pattern artifacts that can arise from inadequate sampling, thereby leading to no definitive biological inference. The ability of the key to yield an inconclusive outcome is a strength, not a weakness, because the deficiencies of the current sample for making unambiguous biological inference are identified. For example, the application of this TRP key to the buffalo mtDNA data set (Fig. 2A) yields a conclusion of gene flow between Chobe and Tanzania, but the absence of samples between these localities leave it ambiguous as to whether this gene flow is characterized by isolation-by-distance or by occasional but recurrent long-distance dispersal (Templeton & Georgiadis 1996; Table 13.2). For the impala, there is no discrimination between restricted gene flow, fragmentation, or range expansion as possible explanations for the Chobe/Tanzanian pattern because of the absence of samples between Chobe and Tanzania (Table 13.4 in Templeton & Georgiadis 1996). Obviously, in both cases, future sampling should be directed towards the geographical gap between Tanzania and Chobe. Thus, the inference key gives specific and detailed guidance for future sampling activities.

The buffalo/impala example also illustrates the difficulty of knowing what is an adequate sampling design a priori. Suppose that an investigator only wanted to know if the Tanzanian and Chobe populations were interconnected by recurrent gene flow (for example, in designing a conservation program). For the buffalo, the inference of recurrent gene flow has already been established with the samples given in Templeton & Georgiadis (1996), although the details of the nature of the gene flow remain hidden. On the other hand, the impala sample is inconclusive on this issue even though sampling occurred in the same areas and with comparable sample sizes. Each species has a potentially unique population structure and
history that shaped its evolution, so it is difficult to design
an optimal sampling scheme when sampling resources
are limited. It is better to use only a portion of the
sampling resources available to perform an initial analy-
sis, and then use the inference key as a guide in allocating
the remaining resources to obtain the most critical
samples needed for strong inference.

Validity of the criteria used to infer range expansion
The basic patterns associated with restricted gene flow
that were incorporated into the inference key are well
justified by recent work in coalescent theory and com-
puter simulations (Hudson et al. 1992; Slatkin &
Maddison 1990; Slatkin 1991, 1993; Nei & Takahata 1993;
Neigel et al. 1991; Neigel & Avise 1993; Takahata & Slatkin
1990; Takahata 1991). Similarly, the predictions under
fragmentation are straightforward and represent a quan-
titive rendering of the patterns commonly used to infer
fragmentation events (Avise 1994). The least theoretically
justified pattern is that associated with range expansion.
The range expansion expectations of widespread tip
clades with some ancestral haplotypes restricted to the
ancestral range were first described by Cann et al. (1987).
Although these expectations seem reasonable, they have
not been confirmed analytically or through extensive
computer simulations. Part of the problem is that range
expansion can arise in many different situations and can
interact with many different patterns of gene flow and/or
fragmentation events. Hence, the range of possible
assumptions that could be incorporated into an analytical
or simulation model is daunting, and it is not clear which
assumptions are most biologically realistic.

The best method of insuring biological realism is to
examine actual examples of range expansion. Fortunately,
there are many cases in which range expansion can be
inferred with much certainty without the use of genetic
data. Table 1 presents 13 data sets that have strong prior
evidence for range expansion, a genetic survey using
restriction site mapping or DNA sequencing (in all cases
involving mtDNA), and well-documented geographical
sampling with spatial frequency information on all haplo-
types. The first seven cases involve organisms with
current ranges that include areas that were not inhabi-
table during the Pleistocene; hence, post-Pleistocene range
expansion must have occurred. The remaining six cases
involve organisms whose ranges have been expanded by
human activities or range expansions by humans them-
selves. These natural examples of range expansion
provide an excellent vehicle for validating the expecta-
tions of Cann et al. (1987) and the inference key.

The most common statistically significant inference in
these data sets is restricted gene flow, which is inferred in

<table>
<thead>
<tr>
<th>Organism</th>
<th>N I</th>
<th>N H</th>
<th>A priori reason for range expansion</th>
<th>Statistically significant historical events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambystoma tigrinum</td>
<td>13</td>
<td>6</td>
<td>Current range includes areas climatically uninhabitable in the Pleistocene</td>
<td>Northwest range expansion into the northern Great Plains</td>
</tr>
<tr>
<td>A. t. mavortium</td>
<td>40</td>
<td>16</td>
<td>Current range includes areas climatically uninhabitable in the Pleistocene</td>
<td>Northeast contiguous range expansion into the northern Great Plains</td>
</tr>
<tr>
<td>Etheostoma blennioides</td>
<td>11</td>
<td>4</td>
<td>Current range includes areas under glaciers in the Pleistocene</td>
<td>Range expansion from southern Ohio drainage rivers to rivers in formerly glaciated areas</td>
</tr>
<tr>
<td>E. b. pholidotum</td>
<td>15</td>
<td>5</td>
<td>Current range includes areas under glaciers in the Pleistocene</td>
<td>Range expansion from northern Ozark rivers draining into the Missouri vs. Mississippi Rivers</td>
</tr>
<tr>
<td>E. b. pholidotum</td>
<td>11</td>
<td>4</td>
<td>Current range includes areas under glaciers in the Pleistocene</td>
<td>Range expansion from northern Ozark rivers draining into the Missouri vs. Mississippi Rivers</td>
</tr>
</tbody>
</table>

Table 1 Summary of the nested geographical analyses of 13 mitochondrial DNA data sets with a priori evidence for range expansion. N I refers to the number of individuals in the study, N H to the number of haplotypes detected in the genetic survey.
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Organism</th>
<th>$N_L$</th>
<th>$N_I$</th>
<th>$N_H$</th>
<th>A priori reason for range expansion</th>
<th>Statistically significant historical events</th>
<th>Reference*</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Trimerotropis saxatalis</em></td>
<td>62</td>
<td>613</td>
<td>41</td>
<td>Ozarks climatically uninhabitable prior to the xerothermic maximum</td>
<td>Range expansion into the Ozarks from the southwest followed by fragmentation</td>
<td>Gerber (1994)*</td>
</tr>
<tr>
<td><em>Geomys bursarius</em></td>
<td>13</td>
<td>159</td>
<td>19</td>
<td>Current range includes areas under glaciers in the Pleistocene</td>
<td>Northwest range expansion from Texas to New Mexico&lt;br&gt;Range expansion from Texas to Minnesota&lt;br&gt;Colonization event east of Mississippi River</td>
<td>Davis (1986)</td>
</tr>
<tr>
<td><em>Galaxias truttaceus</em></td>
<td>16</td>
<td>211</td>
<td>58</td>
<td>Current range includes lakes created by melting Pleistocene glaciers in central Tasmania</td>
<td>Range expansion to central lakes from coast followed by fragmentation&lt;br&gt;Range expansion to north coast</td>
<td>Ovenden &amp; White (1990)</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>18</td>
<td>144</td>
<td>23</td>
<td>Current global distribution due to dispersal by human activities; native to Africa</td>
<td>Both contiguous and long-distance range expansion on a global basis</td>
<td>Hale &amp; Singh (1991)</td>
</tr>
<tr>
<td><em>Drosophila buzzatii</em></td>
<td>15</td>
<td>283</td>
<td>26</td>
<td>Introduction to Europe from South America by human transport of host plant</td>
<td>No population-level historical events detected</td>
<td>Rossi <em>et al.</em> (1996)</td>
</tr>
<tr>
<td><em>Canis latrans</em></td>
<td>25</td>
<td>327</td>
<td>32</td>
<td>Historical range expansion since 1900</td>
<td>Range expansion from southern and western North America to north and east coast</td>
<td>Lehman &amp; Wayne (1991)</td>
</tr>
<tr>
<td><em>Macaca fascicularis</em></td>
<td>3</td>
<td>52</td>
<td>17</td>
<td>Introduced to the Island of Mauritius in the 1500s by Portuguese sailors</td>
<td>Colonization event to Mauritius from the Philippines and/or Indonesia</td>
<td>Lawler <em>et al.</em> (1995)</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>14</td>
<td>345</td>
<td>127</td>
<td>Human settlement of remote Pacific Islands</td>
<td>Contiguous range expansion to nearby islands&lt;br&gt;Colonization events to distant islands</td>
<td>Sykes <em>et al.</em> (1995)</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>21</td>
<td>532</td>
<td>126</td>
<td>Human settlement of Siberia and the Americas</td>
<td>Contiguous range expansion in Siberia&lt;br&gt;1 or 2 colonization events in America followed by fragmentation between America and Siberia&lt;br&gt;Contiguous range expansion in the Americas</td>
<td>Torroni <em>et al.</em> (1993a,b)</td>
</tr>
</tbody>
</table>

*Reference includes the nested cladistic analysis.*
all data sets except the darter Etheostoma blennioides pholidotum, the macaque monkey, and the fish Galaxias truttaceus. For the darter, the failure to detect restricted gene flow may be due to low genetic resolution (only five haplotypes, Table 1), and in the macaque, too few localities (only three, Table 1). In G. truttaceus there is genetic variation but homogeneity across most sampling locations at all clade levels, thereby indicating much gene flow among most of the sampled localities in this species. This inference is consistent with the an isozyme survey (Ovenden & White 1990). In the 10 examples showing restricted gene flow, it is most commonly observed through isolation-by-distance when the sampling was sufficient to discriminate the type of restricted gene flow. Only the two human examples showed some recurrent long-distance exchange, but even in humans gene flow is restricted primarily through isolation-by-distance.

Because the focus in this section is on historical events that influenced the spatial distribution of haplotype variation rather than gene flow, Table 1 only gives the events that were inferred using the inference key. If the geographical sampling is adequate, the inference key discriminates between contiguous range expansion (a gradual, moving front of range expansion) vs. colonization (an abrupt establishment of a population in a new geographical region), and these inferred discriminations are indicated in Table 1. However, in many cases the inference key yields an inference of range expansion, but the sample is inadequate to discriminate the details of the nature of the expansion. These events are simply referred to as ‘range expansion’ in Table 1.

The first five cases in Table 1 (two subspecies of tiger salamanders, two subspecies of a darter, and the lichen grasshopper) have already been analysed with a nested geographical analysis and the inference key. The remaining eight cases all represent new nested analyses of previously published data. The analysis of the first of these, the gopher Geomys bursarius (Davis 1986), is given below, but space limitations preclude giving the details of the other analyses. However, the details of any or all of these analyses are available upon request to the author.

Davis (1986) surveyed mtDNA restriction site variation in the gopher, G. bursarius (note, this is part of a complex of gopher taxa whose status is debated: this analysis excludes many of the controversial taxa and restricts the analysis only to the widespread populations of this species in the central part of its range). These gophers are found in a mid-continental belt in North America that extends from southern locations that were never glaciated, such as Texas, to northern locations such as Minnesota that were under Pleistocene ice sheets. Figure 3 represents the mtDNA haplotype network estimated from the data of Davis (1986) using the algorithm of Templeton et al. (1992), along with the resulting nested design. In this case, outgroup data can root the tree, and the only effect the outgroup has on the nested design is to designate clade 4–1 as an interior clade at the highest level of nesting. Figure 4 presents the results of the nested analysis of clade and nested clade distances, along with the inferences reached by using the inference key. As can be seen from Fig. 4, there is much evidence for range expansion in this species. The inferences of range expansion found within clades 1–1, 2–1 and 4–1 all involve populations in Texas and New Mexico, indicating a north-west range expansion from Texas. These three clades define a continuous nested series of clades (Fig. 4, note that clade 3–1 is the same as clade 2–1; such ‘degeneracy’ arises when there are internal nodes in the haplotype tree that are not represented by haplotypes actually present in the sample). This indicates that this range expansion occurred gradually on the time scale marked by mtDNA coalescent events. The range expansion within clade 3–4 is specifically inferred to be due to a colonization event, and involves the only population of this species on the east bank of the Mississippi River (in Illinois), indicating a transfer event of gophers to Illinois from the west. The range expansion detected within clade 4–2 involves populations from Kansas to Minnesota, and this event in turn is nested within a range expansion at the highest clade level involving populations from Texas and New Mexico expanding into the northern states from Kansas through Minnesota. The nesting of these two range expansion events both geographically and within the clade structure of the haplotype tree indicates that this was also a range expansion that occurred gradually on the timescale marked by mtDNA coalescent events. All of these inferences are consistent with prior expectations (Davis 1986).

Discussion

The results given and summarized in this study clearly demonstrate that nested analyses of haplotype trees with geographical data provide greater statistical power and precision than traditional F-statistic analysis for detecting genetic/geographical associations. More importantly, the nested haplotype tree approach can reveal extremely different patterns of association even in those cases that appear indistinguishable to nonhistorical analyses. The fact that different spatial/temporal patterns can be detected with the nested analysis opens up the potential for discriminating among various evolutionary causes for associations arising between genetic and spatial variation. Restricted gene flow, fragmentation events, and range expansion events (including both contiguous range expansion and long-distance colonization events) can all create genetic/spatial associations. Of these, the patterns associated with range expansions are the most controversial and
least studied. This study therefore examined the ability of the nested analysis to infer range expansion events. Accordingly, 13 data sets with strong prior evidence of range expansions were analysed, with the results summarized in Table 1. As can be seen, range expansions were inferred in 12 of the 13 cases.

One potential explanation for this high success rate is that the criteria for range expansion given in the inference key are so broad that range expansion events will be commonly inferred. If this were true, a large number of false positives would be expected as well. Fortunately, this does not seem to be the case. A total of 99 nesting clades had significant deviations from the null hypothesis in the 13 data sets analysed, but only a subset of 35 led to the inference of range expansion. The most common inference was restricted gene flow, and a few fragmentation events were also inferred (Table 1). Moreover, of the 35 nesting clades associated with a significant range expansion pattern, 34 were consistent with prior knowledge (many range expansion events influenced multiple clades, as shown by the gopher example). Only one inferred range expansion event was not expected a priori, and that is the expansion of *Galaxias truttaceus* from the south-eastern coastal rivers to the north in Tasmania. Even this inferred range expansion event is not necessarily a false positive because a land bridge existed between Tasmania and Australia 10 000–20 000 years ago (Ovenden & White 1990) which may have prevented the south-eastern fish from reaching the northern streams until relatively recently. However, even if this case is regarded as a false positive, the fact that 34 of the 35 clades inferring range expansion were compatible with prior knowledge out of a
A total of 99 significant clades implies that the inference key does not lead to frequent false positives.

This conclusion is reinforced by contrasting these analyses to other nested analyses done on species or sets of populations where there was no prior expectation of range expansion. There was no prior expectation of range expansion in the *Drosophila melanogaster* populations from the eastern half of the USA, and none was inferred for either the Adh or Amy loci even though both loci detected significant geographical associations due to gene flow. There was also no prior expectation of range expansions in the three African bovid species analysed by Templeton & Georgiadis (1996), and only one was inferred for the impala (which in hindsight is biologically reasonable, Templeton & Georgiadis 1996). Another recently published example of an organism with no prior expectation of range expansion is provided by the work of Williams & Benzie (1997) on the high-dispersal starfish *Linckia laevigata* in the Indo-West Pacific. The mtDNA haplotype tree given in that paper was subjected to the nested distance analysis. Out of 23 nested clades, six led to the rejection of the null hypotheses of no geographical associations. Using the inference key, four of these associations were inferred to be due to gene flow constrained by isolation by distance.
and two were inconclusive. These conclusions are consistent with the interpretations given in Williams & Benzie (1997). If the inference key is an accurate guide, there should be an excess of inferences of range expansion in the data sets with prior knowledge of range expansion. This is indeed the case: 12 of the data sets with prior knowledge of range expansion led to a statistically significant inference of range expansion (s) and one did not; of the six data sets without prior knowledge of range expansion discussed above, one led to a significant inference of range expansion and five did not. This difference in the frequency of range expansion inference is significant (the two-tailed Fisher’s exact test P value is 0.003). Even if the *D. melanogaster* Adh and Amy results are regarded as a single test rather than two, the difference is still highly significant (the two-tailed Fisher’s exact test P value is 0.008). The difference between these two data sets is also seen when the Fisher’s exact test is applied to nesting clades as the unit of analysis rather than species of gene region: for the 13 data sets with prior knowledge of range expansion, 35 significant nesting clades led to inferences of range expansion while 64 did not; for the six data sets with no prior knowledge of range expansion, one significant nesting clade led to an inference of range expansion and 23 did not. This difference has a two-tailed P value of 0.002. These results reinforce the conclusion that the inference key is not prone to false positives for range expansion.

However, as the *Drosophila buzzatii* example reveals, the inference key is not infallible (Table 1). The patterns described by Cann et al. (1987) which were the basis of the criteria incorporated into the inference key require that the expanding populations carry along with them only a subset of the haplotype variation found in the ancestral geographical range, and places great importance upon tip clades found in the expanded area. There are two ways in which tip clades can be found in the expanded area. First, one or more tip clades could be carried over from the ancestral population into the expanding population. Second, after expansion occurs, restricted gene flow (or fragmentation) between the colonized region and the ancestral region would allow the mutational process to create new tip clades that are found primarily or exclusively in the colonized region. Hence, the criteria for range expansion in the inference key will not be satisfied if the range expansion took place by a colonization event associated with an extreme founder effect such that no tip haplotypes were included in the original colonizing population (or at least failed to survive to the present time in the descendants of the colonists), and if the colonization event was sufficiently recent such that no new mutations have arisen in the colonized area.

This is apparently what happened in the case of *D. buzzatii*. Rossi et al. (1996) studied mtDNA restriction site variation in Argentinean and Iberian Peninsula popula-

At the beginning of this article, the question was raised ‘can this new temporal information be used to shed more light upon the spatial distribution of current allelic variation?’ In light of the results presented and summarized in this article, the answer is clearly yes. Haplotype trees allow more power in testing for genetic/geographical associations and more precision in estimating gene flow parameters. Nested analyses of haplotype trees solve one of the major problems of interpreting spatial patterns of genetic variation, i.e. separating the effects of population structure from population history. Moreover, unlike drawing inferences from pictorial overlays of haplotype networks upon geography, the nested geographical analysis coupled with the inference key provides an assessment of statistical significance, explicit inference criteria, and guidance to the researcher for how to collect future samples to make sound biological inference. Finally, the nested analyses allow a dynamic, temporal reconstruction of how population structure and historical events have been interwoven to shape the present-day composition of the population under study. For these reasons, haplotype trees represent a powerful tool that quantitatively and qualitatively enhances the ability to study population structure and recent evolutionary history.

One major limitation of this approach is that it is basically a single-locus analysis. As a result, both evolutionary stochasticity and locus-specific evolutionary forces such as natural selection may either erode power or even mislead the investigator. One way to circumvent this problem would be to perform separate haplotype tree analyses on multiple loci surveyed in the same individuals. This approach is exemplified by the nested geographical distance analyses of the Adh and Amy loci surveyed in the same stocks of D. melanogaster in the eastern USA. Both analyses detected significant associations, and in both cases the inference key led to the inference of restricted (but still high) gene flow as the cause. The compatibility of the results across loci is reassuring, but in the future it would be desirable to go beyond an assessment of compatibility. Just as many loci can be pooled together to yield a single estimate of $F_{ST}$ in the traditional, nonhistorical analyses of population structure, methods need to be developed for pooling the results across loci into an integrated analysis of population structure and history. However, even an assessment of compatibility across loci could represent a powerful tool for investigating evolutionary forces. Lewontin & Krakauer (1973) suggested that loci that yield $F_{ST}$ values which are discrepant with the $F_{ST}$ values estimated from the majority of loci are good candidates for having been influenced by natural selection. Given that the results of this study indicate that nested haplotype tree analyses are more powerful and detailed than nonhistorical $F_{ST}$ analyses, it may prove that haplotype tree analyses will be much more powerful in detecting discrepant DNA regions that have been subjected to locus-specific evolutionary forces. Thus, a second need for future development would be to integrate these haplotype tree analyses of geographical distribution with haplotype tree analyses of natural selection (e.g. Templeton 1996b). With such an integrated geographical/selectional analyses, it would be possible to test directly the hypothesis that outlier DNA regions have been subjected to natural selection. This would result not only in cleaner and harder inferences about population structure and history, but would also provide a potentially powerful tool for studying natural selection. This is indeed an exciting prospect.

Acknowledgements

I would like to thank Keri Williamson, Christina Gardenal, Reinaldo Alves de Brito, and Rebecca Reichardt for their many valuable comments and suggestions on an earlier version of this article. I would also like to thank Robert Ricklefs, Eldredge Bermingham, and an anonymous reviewer for their excellent reviews which have resulted in a stronger article.

References


© 1998 Blackwell Science Ltd, Molecular Ecology, 7, 381–397


The work presented here is a continuation of the author’s research on data analysis through statistical overlays upon gene trees. The gene tree approach was first developed to study the association of genetic variation at candidate trees with clinical phenotypes related to coronary artery disease. The author has since extended this approach to a wide variety of applied and basic problems.
Appendix I: Inference key for the nested haplotype tree analysis of geographical distances

Start with haplotypes nested within a 1-step clade:

1. Are there any significant values for \( D_c, D_n \), or I-T within the clade?
   - **NO:** the null hypothesis of no geographical association of haplotypes cannot be rejected (either panmixia in sexual populations, extensive dispersal in nonsexual populations, small sample size, or inadequate geographical sampling). Move on to another clade at the same or higher level.
   - **YES:** go to step 2.

2. Are the \( D_c \) values for tip or some (but not all) interior clades significantly small or is the I-T \( D_c \) distance significantly large?
   - **NO:** go to step 11.
   - **YES:** go to step 3.
   - Tip/interior status cannot be determined: inconclusive outcome.

3. Are any \( D_n \) and/or I-T \( D_n \) values significantly reversed from the \( D_c \) values, and/or do one or more tip clades show significantly large \( D_n \) values or interior clades significantly small \( D_n \) values or I-T significantly small \( D_n \) with the corresponding \( D_c \) values being nonsignificant?
   - **NO:** go to step 4.
   - **YES:** go to step 5.

4. Do the clades (or two or more subsets of them) with restricted geographical distributions have ranges that are completely or mostly nonoverlapping with the other clades in the nested group (particularly interiors), and does the pattern of restricted ranges represent a break or reversal from lower level trends within the nested series (applicable to higher-level clades only)?
   - **NO:** go to step 6.
   - **YES:** go to step 15.

6. Do clades (or haplotypes within them) with significant reversals or significant \( D_n \) values without significant \( D_c \) values define geographically concordant subsets, or are they geographically concordant with other haplotypes/clades showing similar distance patterns?
   - **NO:** go to step 7.
   - **YES:** go to step 13.

7. Are the clades with significantly large \( D_n \) values (or tip clades in general when \( D_n \) for I-T is significantly small) separated from the other clades by intermediate geographical areas that were sampled?
   - **NO:** go to step 8.
   - **YES:** restricted gene flow/dispersal but with some long-distance dispersal.

8. Is the species absent in the nonsampled areas?
   - **NO:** sampling design inadequate to discriminate between isolation by distance (short distance movements) vs. long-distance dispersal
   - **YES:** restricted gene flow/dispersal but with some long-distance dispersal over intermediate areas not occupied by the species.

9. Are the different geographically concordant clade ranges separated by areas that have not been sampled?
   - **NO:** past fragmentation. (If inferred at a high clade level, additional confirmation occurs if the clades displaying restricted by, at least partially, nonoverlapping distributions are mutationally connected to one another by a larger than average number of steps.)
   - **YES:** go to step 10.

10. Is the species absent in the nonsampled areas?
    - **NO:** geographical sampling scheme inadequate to discriminate between fragmentation and isolation by distance.
    - **YES:** allopatric fragmentation. (If inferred at a high clade level, additional confirmation occurs if the clades displaying restricted by at least partially nonoverlapping distributions are mutationally connected to one another by a larger than average number of steps.)

11. Are the \( D_c \) values for some tip clades significantly large, and/or the \( D_c \) values for all interiors significantly small, and/or the I-T \( D_c \) significantly small?
• NO: go to step 17.
• YES: range expansion, go to step 12.

12. Are the $D_n$ and/or I-T $D_n$ values significantly reversed from the $D_c$ values?
• NO: contiguous range expansion.
• YES: go to step 13.

13. Are the clades with significantly large $D_n$ values (or tip clades in general when $D_n$ for I-T is significantly small) separated from the other clades by intermediate geographical areas that were sampled?
• NO: go to step 14.
• YES: long-distance colonization.

14. Is the species absent in the nonsampled areas?
• between contiguous range expansion and long-distance colonization.
• YES: long-distance colonization.

15. Are the different geographically concordant areas separated by areas that have not been sampled?
• NO: past fragmentation. (If inferred at a high clade level, additional confirmation occurs if the clades displaying restricted by, at least partially, nonoverlapping distributions are mutationally connected to one another by a larger than average number of steps.)
• YES: go to step 16.

16. Is the species absent in the nonsampled areas?
• NO: go to step 18.
• YES: allopatric fragmentation. (If inferred at a high clade level, additional confirmation occurs if the clades displaying restricted by, at least partially, nonoverlapping distributions are mutationally connected to one another by a larger than average number of steps.)

17. Are the $D_n$ values for tip or some (but not all) interior clades significantly small, or the $D_n$ for one or more interior clades significantly large, or is the I-T $D_n$ value significantly large?
• NO: inconclusive outcome.
• YES: go to step 4.

18. Are the clades found in the different geographical locations separated by a branch length with a larger than average number of mutational steps?
• NO: geographical sampling scheme inadequate to discriminate between fragmentation, range expansion, and isolation-by-distance.
• YES: geographical sampling scheme inadequate to discriminate between fragmentation and isolation-by-distance.