Genealogies and geography

N. H. BARTON AND I. WILSON

Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, U.K.

SUMMARY

Any sample of genes traces back to a single common ancestor. Each gene also has other properties: its sequence, its geographic location and the phenotype and fitness of the organism that carries it. With sexual reproduction, different genes have different genealogies, which gives us much more information, but also greatly complicates population genetic analysis. We review the close relation between the distribution of genealogies and the classic theory of identity by descent in spatially structured populations, and develop a simple diffusion approximation to the distribution of coalescence times in a homogeneous two-dimensional habitat. This shows that when neighbourhood size is large (as in most populations) only a small fraction of pairs of genes are closely related, and only this fraction gives information about current rates of gene flow. The increase of spatial dispersion with lineage age is thus a poor estimator of gene flow. The bulk of the genealogy depends on the long-term history of the population; we discuss ways of inferring this history from the concordance between genealogies across loci.

INTRODUCTION

Until recently population genetics was based on allele frequencies or occasionally on genotype frequencies. In particular, the relative rates of gene flow and genetic drift have traditionally been estimated by using the standardized variance of allele frequencies across subpopulations (F_{st}, Wright *et al.* 1942). DNA sequencing now presents us with data that are more naturally represented by genealogies. Every sequence may be unique, in which case allele frequencies tell us nothing; all the information is contained in the genealogical relationship between sequences. If recombination is rare enough relative to mutation, the genealogy can be seen more or less directly, as for example in bacteria where recombination occurs via occasional transformation (Maynard Smith 1990) or in viruses where mutations are frequent (Sharp, this volume). Even if we cannot be sure of the relationships, the theory may still best be described in terms of the underlying genealogies. For example, if we can calculate the likelihood of parameters such as the population size or the rate of gene flow for a particular genealogy, the overall likelihood can be found as a weighted sum over the set of plausible trees (Felsenstein 1992). Even if our statistical inferences are based on allele frequencies, it may still give more insight to derive the theoretical predictions via a genealogical approach.

Here we shall assume that we have the full genealogy of each segment of the genome, including the times at which different lines of descent coalesce. How can we make inferences about the processes of evolution from such ideal data? How much more valuable is it to use information from the full set of genealogies, rather than from allele frequencies, or from pairwise relationships between genes? What is the best sampling scheme: is it better to have small genealogies across many loci, or genealogies for a few loci, but each containing many individuals? We concentrate on the specific question of how to analyse genealogies with geography and, in particular, how to estimate rates of gene flow and genetic drift. We first show the close relation between the classical theory of identity by descent and the genealogical structure, and then develop a simple diffusion approximation that describes how lines of descent coalesce. This approximation applies to a variety of local population structures, and so could be used to make robust inferences about effective population densities and rates of gene flow. However, we show that simple estimators based on the rate of dispersion of lineages over time can be misleading and that in a two-dimensional population the genealogical structure depends primarily on long-term history. Finally, we discuss possible ways of making inferences about this history. (A more extensive version of this paper is to be found in Barton & Wilson (1995), which includes more detailed derivations and a discussion of the effects of barriers to gene flow.)

This paper is to a large extent a synthesis of existing analyses: the classic work of Wright (1943) and Malecot (1948) on identity by descent; Slatkin's (1991) work on the structure of genealogies in stepping-stone models and its application to measuring gene flow (Slatkin & Maddison 1989; Slatkin & Maddison 1990; Hudson *et al.* 1992); Felsenstein's (1992) likelihood methods; and the use by Neigel, Ball and Avise (Ball *et al.* 1990; Neigel *et al.* 1991; Neigel & Avise 1993) of genealogies to infer population histories. The methods here aim primarily at estimating rates of gene flow and genetic drift rather than distinguishing qualitatively different population histories. Thus, they complement

Phil. Trans. R. Soc. Lond. B (1995) **349**, 49–59
Printed in Great Britain

the methods of Crandall and Templeton, which make qualitative distinctions between alternative hypotheses (Templeton 1992; Crandall & Templeton 1993).

THE COALESCENT PROCESS

First, consider a single panmictic population containing $2N$ genes. If a small fraction of these genes are sampled, then their relationship can be approximated very simply: there is a probability $1/2N$ that any two lines of descent will coalesce in a common ancestor in each generation (Kingman 1982). Thus, if there are k genes, there are $k(k-1)/2$ pairs that might coalesce, and the time back to the first coalescence is exponentially distributed with expectation $4N/k(k-1)$ generations. There are then $k-1$ lines of descent remaining, and so the expected time back to the previous coalescence is $4N/(k-1)(k-2)$ generations. The expected age of the whole genealogy (that is, the expected time back to the common ancestor) is thus $4N\{1/k(k-1)+1/(k-1)(k-2)\ldots+1/2\} = 4N(1-1/k)$ generations, which tends to $4N$ generations for a large sample (Felsenstein 1992).

This calculation shows that the structure of any genealogy is very variable. Looking back, the average time taken for a large number of lineages to coalesce to two ancestors is the same as the time these two remaining lineages take to merge (figure 1*a*). Since the latter time is exponentially distributed, and since any two randomly chosen genes have a chance of 1/3 of being related via the last common ancestor, the average divergence time between randomly chosen pairs of genes has a high variance, regardless of how large a sample is taken. Felsenstein (1992) uses this argument to show that pairwise statistics are much less efficient than estimators that include the genealogical structure. Essentially the same consideration applies with geographic structure, and so we summarize the argument here. The population size could be estimated from the average pairwise divergence time $\bar{t}$, by using the relation $\hat{N} = E(\bar{t})/2$. (In practice, $\bar{t}$ could itself be estimated from the average pairwise sequence divergence.) However, the variance of $\hat{N}$ tends to $2N^2/9$ for large samples, rather than to zero (Felsenstein 1992, equation 18). In contrast, the maximum likelihood estimator (MLE) is $\hat{N} = \sum_{j=2}^{k}(j-1)\,t_j/4(k-1)$, where t_j is the time for which there are j lines of descent. The variance of the MLE is $N^2/(k-1)$, which does tend to zero as sample size increases (Felsenstein 1992, equation 7). This basic argument suggests that genealogies will also give much more information about spatial structure than will pairwise measures such as Wright's F_{st} or the statistics proposed by Slatkin *et al.* (Slatkin & Hudson 1991; Hudson *et al.* 1992). However, it also suggests that information from any one genealogy may be misleading, so that reliable estimates may require data from many loci.

This theory of the coalescent process is equivalent to the classical theory of identity by descent. Two genes are said to be *identical by descent*, relative to an ancestral population t generations in the past, if they derive from the same gene in that population (Malecot 1948). Clearly, the probability that two genes coalesce at time t, f_t, is just the difference between the probability of identity by descent relative to ancestral populations at times t and $t-1$ ($f_t = \tilde{F}_t - \tilde{F}_{t-1}$, where $\tilde{F}_t$ denotes the probability of identity by descent via a population t generations back). It is important to realize that the probability of identity by descent is purely a description of the genealogy and does not depend on the allelic state of the genes. The latter is described by the probability of *identity in state*, which is the chance that two randomly chosen genes share the same allele. This depends on how alleles are identified and on the mutational process but is not defined relative to any base population. The probabilities of identity by descent and identity in state are closely related and indeed are often confused with each other. If there are infinitely many alleles and a rate μ per generation of mutation to a novel allele, then the probability of identity in state is

$$F = \sum_{t=1}^{\infty}(1-\mu)^{2t}(\tilde{F}_t - \tilde{F}_{t-1}) = \sum_{t=1}^{\infty}(1-\mu)^{2t} f_t. \qquad (1)$$

(Throughout, we assume discrete generations.) Equation (1) applies to any population structure and shows the close relation between identity in state ($F(\mu)$), identity by descent ($\tilde{F}_t$) and the distribution of coalescence times (f_t). Moreover, equation (1) shows that, if F is considered as a function of $z = (1-\mu)^2$, then $F(z)$ is the generating function for the distribution of coalescence times. Then, $\partial^t F/\partial z^t|_{z=0} = t!\,f_t$ and $\partial^k F/\partial z^k|_{z=1} = k!\,E(t^k)$. The wealth of existing results on identity in state in spatially structured populations therefore leads directly to the distribution of coalescence times.

GENE FLOW AND THE COALESCENT PROCESS

The pattern of neutral genetic variation can be found by superimposing random changes in allelic state on the genealogy. This is possible because, by definition, neutral mutations do not affect reproduction and so are independent of the genealogy. It is tempting to treat the movement of genes in an analogous way: just as genes change their allelic state through random mutation, so their location changes as the organisms that carry them disperse. This idea was introduced by Cavalli Sforza & Edwards (1964), as the 'Brownian/Yule process'; Edwards (1970) discusses the joint inference of ancestral locations and times. (The process was originally conceived as a model of the whole population, run forwards in time, in contrast to the present application.) Here we use this model to derive a simple expression for the joint distribution of geographic location and genealogical structure. However, we shall see that this expression cannot describe natural populations because it implicitly ignores local regulation of population density (Felsenstein 1975). Nevertheless, it leads us to a diffusion approximation that does apply to a variety of actual population structures.

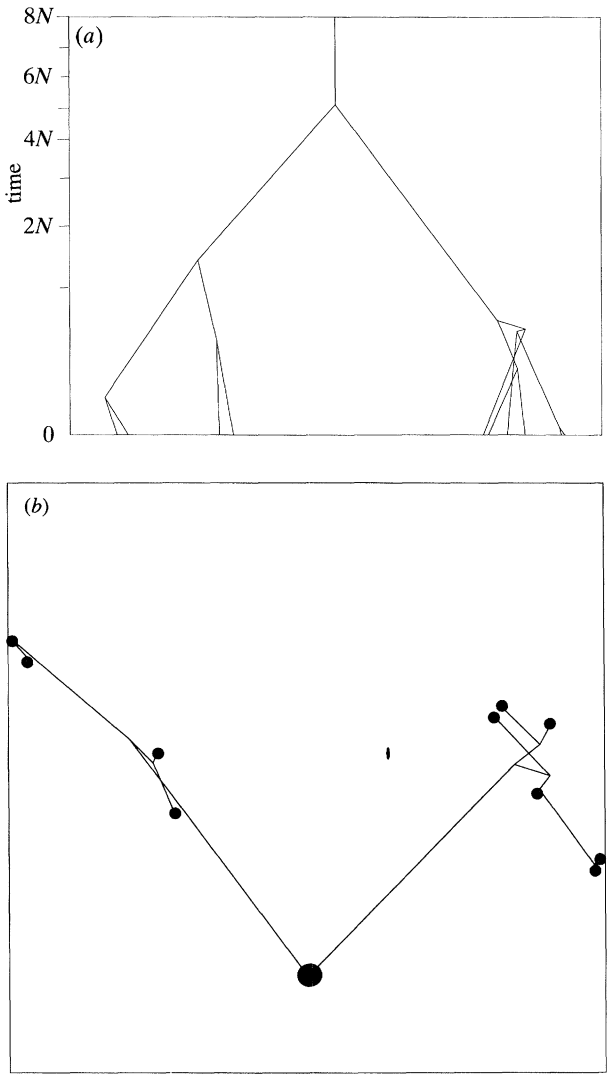

Figure 1. (*a*) A genealogy connecting ten genes. Time is plotted on one axis, and one spatial axis on the other. Random reproduction, with no local density regulation, is assumed. Time is drawn on a square-root scale. (*b*) The same tree in two dimensions, showing locations in space (circles), with lines to indicate the relations. Small circles indicate genes and the large circle the common ancestor. The species' range is assumed to be infinite, corresponding to zero density and zero neighbourhood size. With finite range, and hence non-zero density and neighbourhood size, the plots would be folded around, confusing the relation between geography and genealogy for large times.

Consider a large population of $2N$ genes, which are distributed over a two-dimensional habitat with area A. A sample of genes is taken at random from the population, without regard to their location. If we make the apparently innocent assumption that reproduction is independent of the position of the genes in the sample (or their ancestors), then their relationship should have the same distribution as in a single panmictic population, the intervals between coalescence times t_j being exponentially distributed with expectation $4N/j(j-1)$. If we now assume that genes move in a gaussian random walk, with variance σ^2 per generation along each of the two axes, we can write the joint distribution of coalescence times and locations. For example, the probability that two randomly chosen genes are related through a common ancestor t generations ago and are at positions x, x' is

$$\psi(\underline{x},\underline{x}',t)\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' = \frac{1}{2t\,Nb}\frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2}\left(1-\frac{1}{2N}\right)^t \exp\left(-\frac{|\underline{x}-\underline{x}'|^2}{4\sigma^2 t}\right), \quad (2)$$

where $Nb = 4\pi\rho\sigma^2$ is Wright's (1943) neighbourhood size and $\rho = N/A$ is the population density. Neighbourhood size plays a crucial role in determining the relative rates of gene flow and genetic drift in a two-dimensional population; roughly speaking, it is the number of individuals within one generation's dispersal range. The expression for a set of k genes would involve a similar product of the exponential distribution of coalescence times, with the gaussian distribution of locations, given those times.

In fact, genes are sampled from particular locations. We therefore require the distribution of coalescence times conditional on location, $\psi(t|\underline{x},\underline{x}') = \psi(\underline{x},\underline{x}',t)/\psi(\underline{x},\underline{x}')$, where $\psi(\underline{x},\underline{x}')$ is the chance that genes will be found at $\underline{x}$, $\underline{x}'$. By summing over time,

$$\psi(\underline{x},\underline{x})\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' = \frac{\log\sqrt{2N}}{Nb}\frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2}$$

$$\text{for } |\underline{x}-\underline{x}'| = 0, 2N \gg 1, \quad (3a)$$

$$\psi(\underline{x},\underline{x}')\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' = \frac{1}{Nb}K_0\left(\frac{|\underline{x}-\underline{x}'|}{\sqrt{2N\sigma^2}}\right)\frac{\mathrm{d}x\,\mathrm{d}\underline{x}'}{A^2}$$

$$\text{for } |\underline{x}-\underline{x}'| \gg \sigma \quad (3b)$$

$$\approx \frac{\log\left[\sigma|(2N)/|\underline{x}-\underline{x}'|\right]}{Nb}\frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2}$$

$$\text{for } \sigma\sqrt{2N} \gg |\underline{x}-\underline{x}'| \gg \sigma, \quad (3c)$$

where K_0 is the modified Bessel function.

This expression raises two difficulties. First, it implies extreme clumping: equation (3*a*) shows that the density near to a randomly chosen individual is increased by a factor of about $\log\sqrt{2N}$ above the average. This tendency for randomly dispersing and reproducing populations to become clumped was first emphasized by Felsenstein (1975) as a criticism of classical models of identity by descent. It can be seen as a consequence of the heterogeneous structure of random genealogies discussed above. On average, any large genealogy takes $2N$ generations to coalesce into two clades and then a further $2N$ generations for the remaining two lineages to meet. Hence, the population tends to evolve into distantly related and widely scattered clusters.

The second difficulty is that we have not properly accounted for the finite range of the species. At equilibrium, a random set of genes will have spread over an area of *ca.* $8N\sigma^2 = (2Nb/\pi)A$. (Recall that σ^2 is the variance of distance moved along each of the two axes.) Thus, unless neighbourhood size is very small, most sets of genes will be related by lineages that have crossed the species' range several times. This is true however large the range, because, for given neighbourhood size, divergence time increases in proportion

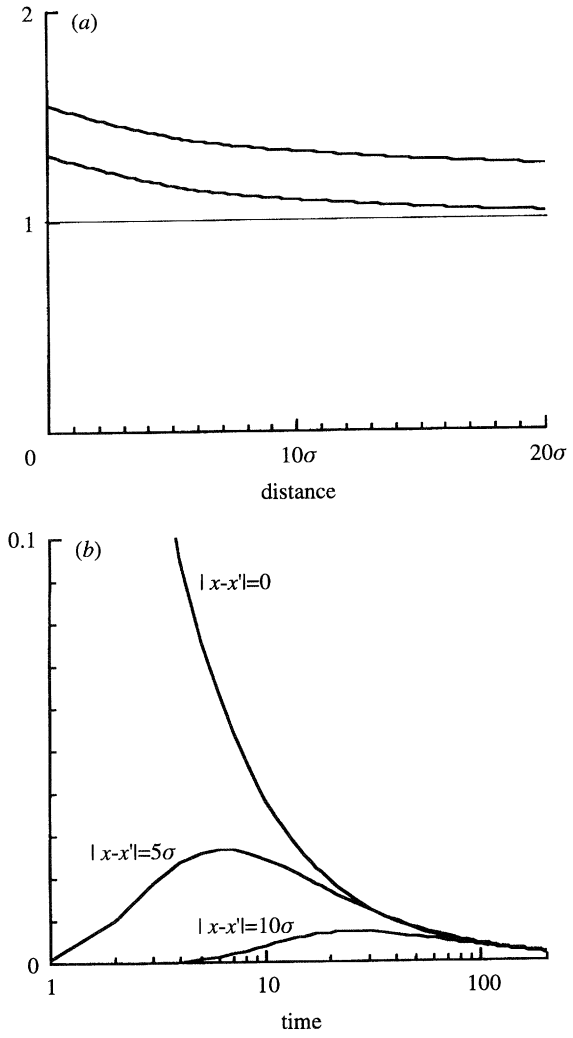

Figure 2. (*a*) The correlation in density as a function of distance, with random reproduction ($A^2\psi(\underline{x},\underline{x}')$, from equations (5)). This gives the density at a point, conditional on the presence of a gene a distance $|\underline{x}-\underline{x}'|$ away. Neighbourhood size is $Nb = 10$. The upper curve is for a species' range of $A = 10^6\,\sigma^2$, while the lower curve is for $A = 10^4\,\sigma^2$. (*b*) Conditional distribution of coalescence times for genes separated by $|\underline{x}-\underline{x}'| = 0,\ 5\sigma,\ 10\sigma$ (from the ratio of equations (4) and (5)).

to area. Hence to be consistent we must allow for either the finite spatial range or the finite age of the species.

It is traditional to allow for finite range by assuming that the habitat is wrapped over the surface of a torus of size $D \times D = A$, so that a movement $\underline{x}$ is equivalent to a movement of $(\underline{x}+\underline{j}D)$, where $\underline{j}$ represents a pair of integers (j_1, j_2). Now, equation (2) must be replaced by a sum over all equivalent locations,

$$\psi(\underline{x},\underline{x}',t)\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' = \frac{1}{2tNb}\frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2}\left(1-\frac{1}{2N}\right)^t \sum_{\underline{j}} \exp\left(-\frac{|\underline{x}-\underline{x}'+\underline{j}D|^2}{4\sigma^2 t}\right). \quad (4)$$

Assumption of a toroidal habitat should be seen as a convenient approximation to more realistic models, with complicated boundaries, and where individuals near the edge might be reflected back into the range. A square habitat with reflecting boundaries would be described in essentially the same way, folding over the square, rather than wrapping around the torus.

The spatial distribution is found by summing over time, and applying the same approximations that led to equations (3),

$$\psi(\underline{x},\underline{x})\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' = \frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{NbA^2}\left(\log\sqrt{2N} + \sum_{\underline{j}\neq 0} K_0\left(\sqrt{\frac{|\underline{j}|^2 2\pi}{Nb}}\right)\right) \quad \text{for } |\underline{x}-\underline{x}'| = 0,\ 2N \gg 1, \quad (5a)$$

$$\psi(\underline{x},\underline{x}')\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' \approx \frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{NbA^2}\left(\log\left(\frac{\sigma\sqrt{2N}}{|\underline{x}-\underline{x}'|}\right) + \sum_{\underline{j}\neq 0} K_0\left(\sqrt{\frac{|\underline{j}|^2 2\pi}{Nb}}\right)\right) \quad \text{for } \sigma\sqrt{2N} \gg |\underline{x}-\underline{x}'| \gg \sigma. \quad (5b)$$

For large neighbourhood size, the sum over Bessel functions can be approximated by an integral and tends to 1. Figure 2 shows examples of (*a*) the clumping produced by random reproduction, and (*b*) the distribution of coalescence times, conditional on location. Since the degree of clumping is small for large neighbourhood size, the conditional distribution does not differ much from the raw distribution.

Sedentary organisms may not have had time to diffuse over the species' range over the time since they occupied it. For example, the flightless alpine grasshopper *Podisma pedestris* has a dispersal range of $\sigma \approx 20$ m generation$^{-\frac{1}{2}}$ (Barton & Hewitt 1982) and an annual life cycle. It has occupied its present range since the glaciers retreated ($T < 10^4$ generations). In that time, genes would diffuse only *ca.* 2 km. In such cases ($\sigma^2 T \ll A$) clustering is limited by finite age rather than finite range. In general, we should consider a sporadic series of extinctions and recolonizations over the whole history of the organism; for simplicity, we suppose that the range was occupied by N diploid individuals T generations ago, distributed uniformly without regard to their relationship. The joint distribution of locations and genealogies is then given by equation (2) (or its multigene extension) for $t \ll T$, and by a uniform spatial distribution for more ancient relationships ($t > T$). For two genes,

$$\psi(\underline{x},\underline{x}',t)\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' = \frac{1}{2tNb}\frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2}\left(1-\frac{1}{2N}\right)^t \exp\left(\frac{-|\underline{x}-\underline{x}'|^2}{4\sigma^2 t}\right) \quad t < T, \quad (6a)$$

$$\psi(\underline{x},\underline{x}',t)\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' = g(t)\frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2} \quad t \geqslant T;\ \sum_{t=T}^{\infty} g(t) = \left(1-\frac{1}{2N}\right)^T, \quad (6b)$$

where $g(t)$ is the arbitrary distribution of relationships among pairs of colonists. By summing over time to obtain the distribution of spatial locations,

$$\psi(\underline{x},\underline{x}')\,\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}' = \frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2}\left(\left(1-\frac{1}{2N}\right)^T + \frac{1}{2Nb}\sum_{t=1}^{T}\frac{1}{t}\left(1-\frac{1}{2N}\right)^t \exp\left(-\frac{|\underline{x}-\underline{x}'|^2}{4\sigma^2 t}\right)\right) \quad (7a)$$

$$\approx \frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2}\left(1+\frac{\log(T)+\gamma}{2Nb}\right) \underline{x} = \underline{x}',\ T \ll 2N \quad (7b)$$

$$\approx \frac{\mathrm{d}\underline{x}\,\mathrm{d}\underline{x}'}{A^2}\left(1+\frac{\log(4\sigma^2 T/|\underline{x}-\underline{x}'|^2)+\gamma}{2Nb}\right)$$

$$\underline{x}=\underline{x}',\ T \ll 2N, \quad (7c)$$

where $\gamma = 0.5772\ldots$ is Euler's constant. Hence, the conditional distribution of coalescence times is

$$\psi(t\,|\,\underline{x},\underline{x}) \approx [t(2Nb+\log T+\gamma)]^{-1}\ T \ll 2N, \quad (8a)$$

$$\psi(t\,|\,\underline{x},\underline{x}') \approx \frac{\exp(-|\underline{x}-\underline{x}'|^2/4\sigma^2 t)}{t[2Nb+\log(4\sigma^2 T/|\underline{x}-\underline{x}'|^2)+\gamma]}$$

$$\sigma \ll |\underline{x}-\underline{x}'| \ll \sigma\sqrt{T}. \quad (8b)$$

These models of random reproduction give a consistent conclusion for times that are short compared with the age of the population or the time taken for genes to diffuse across the species' range ($t \ll T, A/\sigma^2$). The distribution of coalescence times is then given by $\psi(t\,|\,\underline{x},\underline{x}') \approx \exp(-|\underline{x}-\underline{x}'|^2/4\sigma^2 t)/2t\tilde{N}b$, where $\tilde{N}b(|\underline{x}-\underline{x}'|)$ is the effective neighbourhood size, which is increased somewhat above $4\pi\rho\sigma^2$ by the clustering that occurs in the absence of local density regulation. While we have given explicit results only for pairs of genes, it is easy to generate the joint distribution of genealogies and locations for arbitrary numbers of genes, or to draw random realizations of this distribution (figure 1).

A DIFFUSION APPROXIMATION TO THE COALESCENT PROCESS IN TWO DIMENSIONS

The model developed in the previous section is simple but unrealistic: in nature, fitness must decrease with local density. Densities may vary from place to place, but in response to local carrying capacity rather than to the unchecked accumulation of demographic fluctuations. The model of a locally unregulated population might describe the random evolution of morphology and would give an explanation for the clustering of asexual phenotypes into species (Higgs & Derrida 1991). However, it is implausible as a description of population dynamics in two spatial dimensions.

There have been extensive treatments of the extreme case where individuals are grouped into demes whose density is absolutely regulated. Analyses of such stepping-stone models have dealt primarily with identity in state (Wright 1943; Malecot 1948; Kimura & Weiss 1964; Maruyama 1972; Nagylaki 1974; Felsenstein 1976; Nagylaki 1986). However, since the identity in state is the generating function for the distribution of coalescence times (equation (1)), it leads immediately to the distribution of coalescence times (see below). Slatkin (1991) and Nei & Takahata (1993) have derived the mean coalescence time. However, the results would be cumbersome to extend to higher moments. Here, we develop a simple diffusion approximation that applies to all but local scales and that extends to whole genealogies.

The approximation is based on Wright's (1943) argument that ancestors can be considered as being drawn from a neighbourhood whose size increases with time into the past. Thus, the probability that two nearby genes are identical by descent in the previous generation is (by definition) $1/2Nb$; in two dimensions, the pool of ancestors is spread over an area that increases linearly with time, and so the probability of identity by descent t generations back is $1/2tNb$.

Let $f(t,\underline{z}\,|\,\underline{x},\underline{x}')$ be the probability that genes at $\underline{x}$ and $\underline{x}'$ are identical by descent via an ancestor t generations back who lived at $\underline{z}$. (With two genes, we shall not need to keep track of the position of ancestors. However, this would be necessary to extend the argument to more genes. See Barton & Wilson (1995).) Let $g_1(\underline{y},\underline{x})$ be the chance that a gene at $\underline{x}$ derived from an ancestor at $\underline{y}$ in the previous generation. We assume that g_1 depends only on the distance between parent and offspring, has zero mean and has variance σ^2 along each of the two axes. Let $f(1,\underline{y}\,|\,\underline{x},\underline{x}') = h(\underline{y},\underline{x},\underline{x}')/2Nb$ be the chance that two genes at $\underline{x}$, $\underline{x}'$ were identical by descent through an ancestor at $\underline{y}$ in the previous generation. There is a chance $g_1(\underline{y},\underline{x})\,g_1(\underline{y},\underline{x}')\,\delta\underline{y}^2$ that two lines of descent come from some small area $\delta\underline{y}$ in the previous generation; if they do, there is a chance $(1/2\rho\delta\underline{y})$ that they will be identical and a chance $(1-1/2\rho\delta\underline{y})$ that they are not but instead are identical via some more distant ancestor. Hence

$$f(1,\underline{y}\,|\,\underline{x},\underline{x}') \equiv \frac{1}{2Nb}h_1(\underline{y},\underline{x},\underline{x}') = \frac{g_1(\underline{y},\underline{x})\,g_1(\underline{y},\underline{x}')}{2\rho}, \quad (9a)$$

$$f(t,\underline{z}\,|\,\underline{x},\underline{x}') = \int f(t-1,\underline{z}\,|\,\underline{y},\underline{y}')\,g_1(\underline{y},\underline{x})\,g_1(\underline{y}',\underline{x}')\,\mathrm{d}\underline{y}\,\mathrm{d}\underline{y}' - \frac{1}{2Nb}\int f(t-1,\underline{x}\,|\,\underline{y},\underline{y})\,h_1(\underline{y},\underline{x},\underline{x}')\,\mathrm{d}\underline{y}. \quad (9b)$$

This argument requires several approximations. It is assumed that the lines descend independently, so that the joint probability of movement of two genes is $g_1(\underline{y},\underline{x})\,g_1(\underline{y}',\underline{x}')$. It is assumed that an area $\delta\underline{y}$ can be chosen large enough that $1/2\rho\delta\underline{y} < 1$ but small enough that f is approximately constant within it. These assumptions hold for a demic structure with strict density regulation, in which case equations (9) give the coalescence times exactly. They are approximations to models of truly continuous populations; in the cases we consider, the approximation is remarkably good (see figure 3).

Equations (9) give a recursion across one generation that relates the probability of coalescence at time t to that at time $t-1$. It can be rewritten as a recursion across many generations, which leads naturally to the diffusion approximation. Let $g_t(\underline{y},\underline{x})$ be the chance that a gene at $\underline{x}$ descended from an ancestor at $\underline{y}$, t generations back. Extend the definition of h_t to

$$(1/2Nb)\,h_t(\underline{y},\underline{x},\underline{x}') = g_t(\underline{y},\underline{x})\,g_t(\underline{y},\underline{x}')/2\rho t. \quad (10a)$$

This is the chance that two genes both descend from an ancestor at y, t generations back, the possibility of more recent coancestry being ignored. By applying equations (9) recursively,

$$f(t,\underline{z}\,|\,\underline{x},\underline{x}') = \frac{1}{2Nb}\left(\frac{h_t(\underline{y},\underline{x},\underline{x}')}{t} - \sum_{i=1}^{t-1}\int f(t-i,\underline{z}\,|\,\underline{y},\underline{y})\frac{h_i(\underline{y},\underline{x},\underline{x}')}{i}\,\mathrm{d}\underline{y}\right). \quad (10b)$$

This is the chance that the two genes descend from a common ancestor at $\underline{z}$, after subtraction of the

probabilities that they were identical by descent in any of the intervening generations.

By the central limit theorem, $g_t(\underline{y},\underline{x})$ tends to a gaussian with variance $\sigma^2 t$ for large t; h_t tends to a gaussian with variance $\sigma^2 t/2$, being the distribution of locations of the common ancestor. If we average over the location of the ancestors ($f(t|\underline{x},\underline{x}')$) and use the fact that $f(t|\underline{x},\underline{x})$ is independent of $\underline{x}$, equations (10) simplify to

$$f(t|\underline{x},\underline{x}) = \frac{1}{2Nb}\left(\frac{1}{t} - \sum_{i=1}^{t-1}\frac{f(t-i|\underline{x},\underline{x})}{i}\right), \quad (11a)$$

$$f(1|\underline{x},\underline{x}) = 1/2Nb, \quad (11b)$$

$$f(2|\underline{x},\underline{x}) = (1/2Nb)\,(\tfrac{1}{2} - 1/2Nb), \quad (11c)$$

$$f(3|\underline{x},\underline{x}) = \frac{1}{2Nb}\left[\frac{1}{3} - \frac{1}{2Nb.2} - \frac{1}{2Nb}\left(\frac{1}{2} - \frac{1}{2Nb}\right)\right]$$

$$= \frac{1}{2Nb}\left(\frac{1}{3} - \frac{1}{2Nb} + \frac{1}{(2Nb)^2}\right). \quad (11d)$$

The distributions of coalescence times predicted by this gaussian approximation are shown in figure 3, together with simulation results. Agreement is close over all but short times and nearby genes. In general F_{st} is low (< 0.10, say (Slatkin 1987)), implying that neighbourhood size is large. In this limit, the distribution of coalescence times in this model of absolute density regulation converges to that developed in the previous section for no local regulation.

These recursions for the distribution of coalescence times lead to parallel recursions for the identity in state. By applying equation (1) to equations (9),

$$F(\underline{x},\underline{x}') = (1-\mu)^2 \int g_1(\underline{y},\underline{x})\, g_1(\underline{y}',\underline{x}')\left[F(\underline{y},\underline{y}') + \frac{1}{2\rho}(1-F(0))\,\delta(\underline{y}-\underline{y}')\right]\mathrm{d}\underline{y}\,\mathrm{d}\underline{y}', \quad (12)$$

where $\delta(\underline{y})$ is the Dirac delta function, and the probability that nearby genes are identical has been rewritten as $F(\underline{x},\underline{x}) = F(0)$ to emphasize that it is independent of location. Equation (12) is identical to Malecot's (1948) model if (as Malecot assumed) the dispersal distribution is gaussian. It applies exactly to stepping-stone models if the integrals are replaced by sums and is an approximation to continuous models that neglects the interactions between nearby genes caused by local density dependence. In the appendix of Barton & Wilson (1995), it is shown that for arbitrary dispersal distributions equation (12) is approximated by

$$F(0) = (1 + Nb/\log\{\sigma/\mathrm{K}\sqrt{[1-(1-\mu)^2]}\})^{-1}, \quad (13a)$$

$$F(\underline{x},\underline{x}') = \frac{(1-F(0))}{Nb} K_0\left(\frac{|\underline{x}-\underline{x}'|}{\sigma}\sqrt{[1-(1-\mu)^2]}\right)$$

$$\text{for } |\underline{x}-\underline{x}'| \gg \sigma, \mathrm{K} \text{ and } \mu \ll 1. \quad (13b)$$

Equation (13) gives the probability of identity in state, based on the assumption of infinitely many alleles; the same expression gives F_{st} measured from the variance in allele frequencies at loci with a finite number of alleles. K is a characteristic scale which depends on the local structure of the population. For a stepping-stone model with nearest-neighbour migration on a square grid, K = (deme spacing)/$\sqrt{32}$. For Malecot's model of gaussian dispersal, $\mathrm{K} = \sigma$. If mutation rates are low ($\mu \ll 1$), $[1-(1-\mu)^2] \approx 2\mu$, and equations (13) depend on the scale $\ell = \sigma/\sqrt{2\mu}$. There will be significant fluctuations over local scales ($\approx \sigma, \mathrm{K}$), but there will be correlations between allele frequencies over the much longer scale ℓ. Equations (13) are a close approximation to continuously distributed populations and break down only over local scales (figure 5 of Barton & Wilson 1995). Agreement is similarly close for stepping-stone models, even for neighbouring demes. The distribution of coalescence times can be calculated by differentiating equations (13) with respect to $z = (1-\mu)^2$. The breakdown of equation (13*b*) for small $|\underline{x}-\underline{x}'|/\sigma$ corresponds to the breakdown of the gaussian approximation to equation (9) for small coalescence times (figure 3).

ESTIMATING THE RATE OF GENE FLOW

The recursions developed in the preceding sections allow calculation of the distribution of coalescence times among genes sampled from a two-dimensional population. However, it would not be easy to use this distribution to make statistical estimates. Here, we outline possible methods for estimating the rate of gene flow (σ), assuming that genes diffuse through a stable and homogeneous population. The traditional method for inferring population structure from genetic data was introduced by Wright (1942, 1943). Geographic variation in allele frequency is generated by genetic drift and reduced by gene flow. In the island model, the balance between these processes is determined by Nm, the number of migrants exchanged between demes; the standardized variance of allele frequencies across demes is $F_{st} = 1/(1+4Nm)$ for large deme size and low migration rates. In two dimensions (though not in one), the relationship is similar, with F_{st} decreasing with neighbourhood size ($Nb = 4\pi\rho\sigma^2$; equations (13)). Almost all analyses of population structure infer Nm or Nb from F_{st}, or from some equivalent measure of variation in allele frequency, such as Slatkin's private allele method (Slatkin 1985; Barton & Slatkin 1986). However, the spatial pattern of allele frequencies also contains information. If (as is usually the case) F_{st} is small, allele frequencies will fluctuate rather little, and the whole distribution can be approximated by a multivariate gaussian, which is defined by its mean and covariance. This covariance is given by equations (13) and depends on two scales: the scale that describes local population structure and the scale $\ell = \sigma/\sqrt{2\mu}$, which describes the balance between mutation and gene flow. Thus, if the mutation rate is known, the rate of gene flow (σ) can be estimated. This approach was first used by Sokal & Wartenberg (1983) and has been explored more recently by Epperson (1989, 1993). It is important to realize that it gives estimates of both the number of migrants or neighbourhood size (Nm or $4\pi\rho\sigma^2$) and the proportion of migrants or rate of diffusion (m or σ): confusingly, both are referred to as the 'rate of gene flow'. This approach is discussed in more depth in Barton & Wilson (1995),

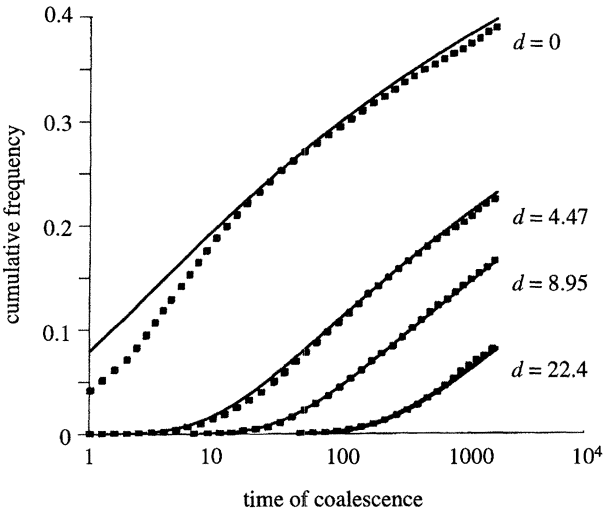

Figure 3. Comparison of simulated distributions of coalescence times from a two-dimensional stepping-stone with ten diploid individuals per deme and a migration rate of 0.05 in each direction (dotted lines), with a gaussian approximation for the distribution of times (equation (10), solid lines). Simulations were based on a grid of 100 by 100 demes, over 2000 generations. The simulated cumulative frequency is based on 8000 replicates for each distance. The distances are zero demes ($d = |\underline{x}-\underline{x}'| = 0$), one deme in both directions ($d = |\underline{x}-\underline{x}'| = 4.47\sigma$), two demes in both directions ($d = |\underline{x}-\underline{x}'| = 8.95\sigma$) and five demes in both directions ($d = |\underline{xx}'| = 22.2\sigma$). Neighbourhood size is $Nb = 6.28$.

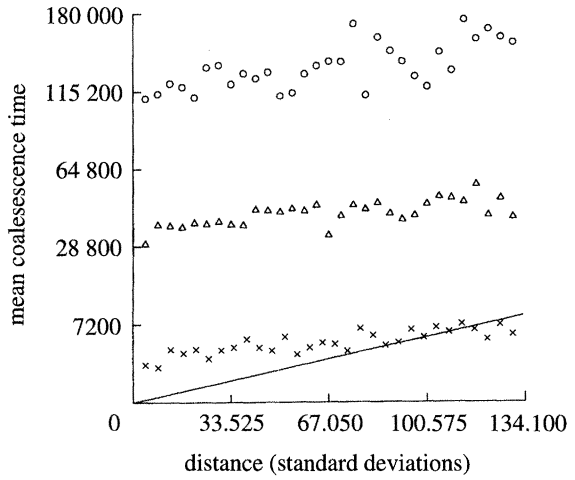

Figure 4. Simulation results from a stepping stone model compared with the results of Neigel & Avise (1993) (straight line). The simulated results are based on all coalescence times (circles), on times less than 100000 generations (triangles) and on times less than 10000 generations (crosses). The distances were taken from a linear array of demes in a stepping-stone model of 100 by 40 demes with 20 diploid individuals per deme and a migration rate of 0.05 in both x and y directions. Distances are given in standard deviations.

in which it is shown that σ could be estimated, but only if the population had been in equilibrium for about $1/\mu$ generations and the mutation rates were known. Even then large samples would be needed.

Neigel, Ball and Avise propose a straightforward method for estimating the long-term rate of gene flow from genealogies (Neigel *et al.* 1991; Neigel & Avise 1993). They suggest plotting the squared distance between genes against the time since they diverged, for all pairs in the sample, and use of the slope of this relation to estimate $4\sigma^2$ (figure 4). (In the notation used in this paper, the squared distance in two dimensions increases as $2\sigma^2$ per unit time; two genes that shared a common ancestor t generations back are separated by $2t$ generations.) Neigel *et al.* (1991) simulate a set of genes forward in time, assuming random dispersal over a square of size $(10000\,\sigma)^2$. Overall population size is regulated at 1000 individuals, so that the model corresponds to that described by equation (2) above. Statistics are based on samples taken from this set and support a linear relation between the variance of dispersal distance and lineage age, at least for closely related genes. Neigel & Avise (1993) run a wider range of simulations, some of which include local population regulation.

This approach is supported by the analytic results given here, which are also based on the idea that pairs of genes diffuse apart at a rate proportion to σ^2. However, there are several problems, which suggest some possible improvements. First, because the simulations are run forward in time, they must necessarily follow a small number of individuals, of which a relatively large proportion are sampled. Moreover, the neighbourhood size is much smaller than is typical of natural populations; for example, figure 5 of Neigel & Avise (1993) shows results from a population of $N = 10000$ genes dispersion over an area of $A = 10^6\,\sigma^2$. This corresponds to a neighbourhood size of $Nb = 4\pi\rho\sigma^2 = 4\pi N\sigma^2/A = 0.13$. Hence most lineages coalesce before they have had time to spread over the whole species range, and nearby genes are likely to be close relatives. Thus it is not clear how far their results would extend to small samples taken from a large population with moderate to large neighbourhood size. Simulations of the coalescent process for $Nb = 6.28$ show that although $|\underline{x}-\underline{x}'|^2$ does increase approximately linearly with time, at least initially, the slope of the relation is much smaller than $4\sigma^2$ and even nearby genes are separated by a long coalescence time (figure 4).

Second, Neigel *et al.* suggest regression of $|\underline{x}-\underline{x}'|^2$ against time. Now, the distribution of $|\underline{x}-\underline{x}'|^2$ is determined by the sampling geometry, and so $|\underline{x}-\underline{x}'|^2$ should be treated as the independent variable; the coefficient of regression of $|\underline{x}-\underline{x}'|^2$ against time will depend on the distribution of distances sampled, whereas the converse regression of time against $|\underline{x}-\underline{x}'|^2$ would not. Neigel *et al.* (1991, p. 425) avoid this difficulty by sampling uniformly over the species' range; however, this would not be easy in practice. Third, because the distribution of coalescence times decreases with $1/t$ in two-dimensional populations, the expected coalescence time at equilibrium is infinite, for any $|\underline{x}-\underline{x}'|$. (This can be shown by treating F in equations (13) as the generating function for $f(t)$; its differential with respect to $(1-\mu)^2$ at $\mu = 0$ is infinite.) In any actual sample, the mean must be finite; however, if the distribution has infinite moments, the results will be very variable. This paradoxical behaviour of the distribution of coalescence times only applies to an infinitely large and indefinitely old population. If the population in fact colonized its present range T generations back, it would be

reasonable to use only pairs that share a common ancestor before that time. However, the mean coalescence time would depend strongly on T (cf. figure 4) and would be very variable.

In a single panmictic population, pairwise estimates are inefficient because they include a disproportionately large contribution from a single event, the coalescence of two clades into the one common ancestor (Felsenstein 1992). This problem is less severe in samples from two dimensions, because the present rate of gene flow can be estimated only from closely related pairs of genes; the locations of more distantly related genes depend on the ancient history of the species. Unless neighbourhood size is unusually low, only a small proportion (*ca.* $1/Nb$) of pairs of genes will be closely related. Hence, most information will come from independent pairs of genes, rather than related clusters of genes. (Slatkin & Maddison 1990 make the same point, by showing that branch lengths are so long in two dimensions as to be uninformative.) This raises a difficulty, however, in that most of the genealogy does not tell us about the rate of diffusion, but rather about the more distant history of the species. Genealogical data are therefore an inefficient source of information of information about gene flow. The ideal solution to the problem would be to make a maximum likelihood estimate based on the structure of the whole tree, rather than on pairs of genes. Since only the recent part of the tree, which evolved after the species occupied its present range, can be used to infer σ^2, this method would only need to be applied to small clusters of genes. Nevertheless, it presents a daunting computational problem. In the next section, we discuss ways of analysing the bulk of the tree, which tells us about the long-term history of the species.

EXTINCTION AND RECOLONIZATION

Over long timescales, populations cannot be adequately described by the uniform diffusion of genes from place to place. We know that (at least in temperate regions) most species have suffered drastic range changes in the last few thousand generations as a result of climatic change. If genes diffused at the current dispersal rate, they would not have had time to spread to fill the present range of the species. If lineages coalesced at the slow rate implied by current population sizes, then genealogies would be much deeper than is seen, and neutral heterozygosity would be much higher. The above analysis shows that, unless neighbourhood size is very small, most of the information in a tree comes from distant relationships: only a small fraction (*ca.* $1/Nb$) of gene pairs are close relatives. The prime need in genealogical analysis is thus to find statistical methods for using this information to infer the distant history of the population and the processes responsible for that history. In particular, we must explain how genes spread faster, and become more closely related, than is possible by diffusion through a large and stable population.

If the whole range were rapidly colonized from a randomly mating source at time T, there would be no association between genealogy and geography before that time. At the other extreme, spatial relations might be preserved despite expansions and contractions of the range. This is plausible if populations are adapted to a climatic gradient and shift with that gradient (Coope 1979; Atkinson *et al.* 1987). Between these extreme possibilities, there might be expansion from a number of refugia, so that before time T genealogies would only reflect the source of the ancestral population and not more detailed spatial relations. We can imagine fitting data to a variety of such particular historical scenarios and indeed this is the usual approach to 'phylogeography' (Avise 1991). However, unless these scenarios can be constrained or corroborated by independent evidence, there is a danger of being able to explain too much. It is therefore attractive to seek ways of representing drastic changes as a statistical process, for example by supposing that there is a low rate of expansions, in which the population in some area A_1 is replaced by individuals drawn from a smaller area A_2. The way allele frequencies are affected by random extinctions and recolonizations has received considerable attention, though mainly for the simplest case of the island model (Slatkin 1977; Wade & McCauley 1988; Whitlock & McCauley 1990; McCauley 1991). However, this theory had not led to ways of distinguishing random drift from random extinction (Slatkin 1987; Slatkin & Maddison 1989, figure 9); the question is whether genealogical data may be more informative.

Changes in the species' distribution will cause older lineages to spread over larger areas than expected with diffusion alone and cause lineages to coalesce faster than expected from the current population density. One could make an *ad hoc* estimate of some effective diffusion rate, σ_e^2, and effective neighbourhood size, Nb_e, as a function of time, by adapting the methods discussed above. Naively, an increase in σ_e^2 would lead to an increase in the scale over which allele frequency fluctuations are correlated, or equivalently, an increase in the rate of dispersion of lineages with age. Genetic distances do indeed often increase over scales much larger than can be explained by simple isolation by distance, suggesting sporadic range expansions. For example, in the alpine grasshopper *Podisma pedestris*, allele frequencies are correlated over all scales from 50 m to 3 km, a much flatter relation than is consistent with isolation by distance with σ rate of 20 m year$^{-\frac{1}{2}}$ and $T \approx 8000$ years (figure 8 of Barton & Wilson 1995).

However, diffusion rates do not actually increase back into the past: what is needed is a model of the extinction/recolonization process itself. The crucial feature of this process is that it involves concerted movements, such that all the genes within an area tend to move together. This correlation across genes in turn generates correlations between the relations of genes at different loci. How does this affect genetic variation? First, consider allele frequencies. Isolation by distance alone (i.e. diffusive gene flow and sampling drift) causes fluctuations that are independent across loci; in contrast, range expansions tend to produce correlations between the patterns at different loci. This idea has been applied with particular success to the recon-

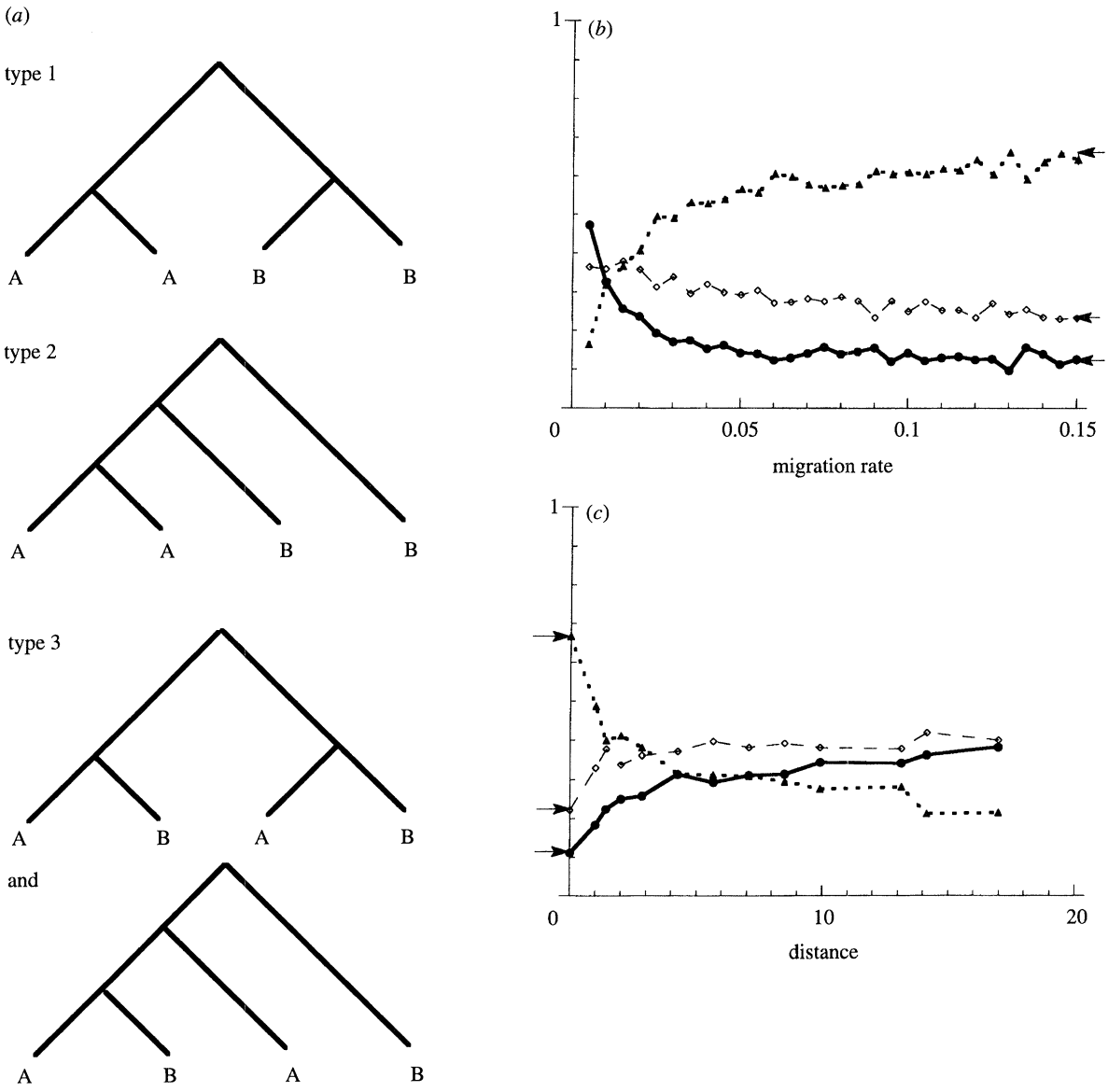

Figure 5. (*a*) The three possible concordances for two pairs of genes; with no population subdivision, the null probabilities of the trees are 1/9, 2/9, 2/3. (*b*) The frequencies of the three different types of concordance possible for two sets of two genes. Simulations are from a stepping-stone model with 20 diploid individuals per deme. Pairs of genes are sampled from demes a distance 2 apart in both x and y directions. Migration rates given are in both directions. (*c*) Simulation results for pairs of genes sampled from a stepping-stone model with ten diploid individuals per site and a migration rate of 0.06 per generation in x and y directions. The distance is given in standard deviations of distance moved (σ). For distance 0 the results are exact.

struction of the history of human populations in Europe. Consistent patterns across loci have shown that linguistic boundaries coincide with genetic boundaries (Sokal *et al.* 1990) and that genes from the originators of agriculture spread into the native European population with a degree of intermingling ('demic diffusion'; Ammermann & Cavalli-Sforza 1981; Sokal *et al.* 1991).

What degree of concordance is to be expected between genealogies at different loci? Even if dispersal is solely by the independent diffusion of genes, and even if linkage is loose, genealogies at different loci are expected to be to some degree parallel, because they will mirror the geographic location of the samples. In contrast, allele frequencies should be independent across loci if linkage disequilibria are negligible. To interpret genealogical data from multiple loci, we therefore need to understand the outcomes expected under the null hypothesis of isolation by distance. Slatkin & Maddison (1989) use the proportion of concordant genealogies to set a bound on the number of migrants between two demes, and give simulation results that allow estimates of neighbourhood size from two samples embedded in one- and two-dimensional stepping stone models (Slatkin & Maddison 1990).

Here, we give a simple illustration of this idea for a two-dimensional population. It is simplest to find the chance that a genealogy will be concordant with the geographic location of the samples, since this determines the concordance of genealogies derived from independent genes with each other. Figure 5*a* shows the three types of relation between two pairs of genes, drawn from two locations. The genealogy may perfectly match the geography (type 1), one pair may

match (type 2), or no pair may match (type 3). If the sample locations are very close, or if the neighbourhood size is very high, there will be no relation between geography and genealogy, and the three types of tree will be in the proportions 1:2:6. Figure 5*c* shows how the degree of concordance increases as the genes move further apart, while figure 5*b* shows how concordance falls with increasing flow. When gene flow is low and the samples are far apart the concordance between geography and genealogy reaches a plateau, which depends on neighbourhood size. This is because in two dimensions only a proportion of approximately $1/Nb$ of lineage pairs coalesce early enough to preserve spatial information. Thus, under isolation by distance, only a fraction of the more closely related genealogies will be concordant with geography. Spatial patterns and concordance across loci among more distantly related genes therefore indicate large-scale changes in population structure.

The power of analyses of allele frequencies comes from having data from many samples and many loci. The same may be true for genealogies. As noted above, strong concordance with geography over large scales and concordance between loci indicate the degree of large-scale population movement, as opposed to independent diffusion. However, discordance may arise for a variety of reasons, reducing the power of this approach. Contraction of the range into refuges will only leave a genetic trace if the populations are small enough, for long enough, for there to be appreciable coalescence. Otherwise, the only effect will be a randomization of ancestral locations. There are inevitably errors in estimating the tree: it is disturbing that, even when using mitochondrial DNA to estimate the relationships among ten major groups of vertebrates, at least 8000 contiguous bases of sequence are needed to give a 95% chance of inferring the correct tree (Cummings *et al.* 1995). Discordance may also arise through selection on particular loci. This may be a particular problem in using mitochondrial DNA for within-species analyses, since the genealogy can be distorted by selection on any of the genes it carries (Thorpe *et al.*, this volume) or on other elements that are inherited maternally, as in *Wolbacchia* (Turelli *et al.* 1992). *Drosophila* mitochondrial sequences show significant deviations from neutral expectations (Ballard & Kreitman 1994; Rand *et al.* 1994), and the frequent introgression of mitochondrial genomes across boundaries demarcated by nuclear alleles may also be a sign of selection on the mitochondrial genome (Harrison 1989). It remains to be seen whether the more detailed information that is contained in genealogies will compensate for the much smaller sample sizes, and whether the concordance between the genealogies for a few loci could allow similar inferences to those based on allele frequencies at many more loci.

CONCLUSIONS

The main purpose of this paper is to emphasize the close relation between genealogical descriptions of spatially structured populations and the classical theory of identity by descent. A naive model of an unregulated population leads to an explicit formula for the joint distribution of locations and relationships, but also leads to unreasonable clumping. For populations subject to strict density regulation, we develop a diffusion approximation for the relation between genes. Both approaches give approximately the same distribution of coalescence times, $f_t = \exp(-|x-x'|^2/4\sigma^2 t)/2Nb\,t$, though only for short times ($t \approx Nb$; see equations (11)). This mathematical complexity makes it hard to develop sound statistical estimators that make full use of genealogical information. For example, the suggestion by Neigel, Ball & Avise (1991) that the rate of dispersion of lineages with time gives the rate of gene flow fails for populations with large neighbourhood size.

These difficulties arise in part from the mathematical and computational complexities. However, there would be fundamental problems in making inferences about evolutionary processes from genealogies and geography, even if the ideal data were available. First, in two dimensions only a small fraction of gene pairs are likely to be closely related, and so most of the information in the tree is about sporadic events in the distant past. This contrasts with the simpler case, where the population is divided across a few islands, all of which can be sampled (cf. Slatkin & Maddison 1989). Inferences may then be tested against geological history (see, for example, Thorpe *et al.*, this volume). Second, sporadic events are hard to fit to any quantitative model, and so we are left with the difficult task of judging the relative merits of a multitude of possible histories, rather than estimating any well defined parameters. Third, genealogies derived from one or a few loci can only inform us of the history of the whole population if they are all affected by population structure in the same way. If species really consist of competing geographic races, which hardly recombine, then genealogies may well be largely concordant. However, whether this is so is at present obscure. Here, we have sketched some possible solutions to the simplest case of isolation by distance. There is an urgent need for a better theoretical and empirical understanding of the distribution of genealogies across multiple loci, and of the effects of large-scale population restructuring.

This work was supported by BBSRC grant GR/H/09928 and by a Scottish Office studentship. We thank A. W. F. Edwards and S. Otto for their helpful comments.

REFERENCES

Ammermann, L. & Cavalli-Sforza, L. L. 1981 *The neolithic transition and the genetics of populations in Europe*. New Jersey: Princeton University Press.

Atkinson, T. C., Briffa, K. R. & Coope, G. R. 1987 Seasonal temperatures in Britain during the past 22,000 years, reconstructed using beetle remains. *Nature, Lond.* **325**, 587–593.

Avise, J .C. 1991 Ten unorthodox perspectives on evolution prompted by comparative population genetic findings on mitochondrial DNA. *A. Rev. Genet.* **25**, 45–69.

Ball, R. M., Neigel, J. E. & Avise, J. C. 1990 Gene genealogies within the organismal pedigrees of random-mating populations. *Evolution* **44**, 360–370.

Ballard, W. O. & Kreitman, M. 1994 Unraveling selection

in the mitochondrial genome of *Drosophila*. *Genetics, Princeton* **138**, 757–772.

Barton, N. H. & Hewitt, G. M. 1982 A measurement of dispersal in the grasshopper *Podisma pedestris* (Orthoptera: Acrididae). *Heredity, Lond.* **48**, 237–249.

Barton, N. H. & Slatkin, M. 1986 A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity, Lond.* **56**, 409–416.

Barton, N. H. & Wilson, I. 1995 Genealogies and geography. In *New uses for new phylogenies* (ed. P. H. Harvey, A. J. Leigh Brown & J. Maynard Smith). Oxford University Press.

Cavalli-Sforza, L. L. J. & Edwards, A. W. F. 1964 Analysis of human evolution. *Proc. Int. Cong. Genet.* **3**, 923–933.

Coope, G. R. 1979 Late Cenozoic fossil Coleoptera: evolution, biogeography, and ecology. *A. Rev. Ecol. Syst.* **10**, 247–267.

Crandall, K. A. & Templeton, A. R. 1993 Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics, Princeton* **134**, 959–969.

Cummings, M. P., Otto, S. P. & Wakeley, J. 1995 Sampling properties of DNA sequence data in phylogenetic analysis. *Molec. Biol. Evol.* (In the press.)

Edwards, A. W. F. 1970 Estimation of the branch points of a branching diffusion process. *J. R. Stastist. Soc.* **32**, 155–174.

Epperson, B. K. 1993 Spatial and space–time correlations in systems of subpopulations with genetic drift and migration. *Genetics, Princeton* **133**, 711–727.

Epperson, B. K. & Allard, R. W. 1989 Spatial autocorrelation analysis of the distribution of genotypes within populations of lodgepole pine. *Genetics, Princeton* **121**, 369–377.

Felsenstein, J. 1975 A pain in the torus: some difficulties with the model of isolation by distance. *Am. Nat.* **109**, 359–368.

Felsenstein, J. 1976 The theoretical population genetics of variable selection and migration. *A. Rev. Genet.* **10**, 253–280.

Felsenstein, J. 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Gent. Res.* **59**, 139–147.

Harrison, R. G. 1989 Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends Ecol. Evol.* **4**, 6–12.

Higgs, P. G. & Derrida, B. 1991 Stochastic models for species formation in evolving populations. *J. Phys.* A **24**, 985–992.

Hudson, R. R., Slatkin, M. & Maddison, W. P. 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics, Princeton* **132**, 583–589.

Kimura, M. & Weiss, G. H. 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics, Princeton* **49**, 561–576.

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

Malecot, G. 1948 *Les mathématiques de l'hérédité*. Paris: Masson et Cie.

Maruyama, T. 1972 Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics, Princeton* **70**, 639–651.

Maynard Smith, J. 1990 The evolution of prokaryotes – does sex matter? *A. Rev. Ecol. Syst.* **21**, 1–12.

McCauley, D. E. 1991 Genetic consequences of extinction and recolonisation. *Trends Ecol. Evol.* **6**, 5–8.

Nagylaki, T. 1974 Continuous selective models with mutation and migration. *Theor. Popul. Biol.* **5**, 284–295.

Nagylaki, T. 1986 *Neutral models of geographic variation* (ed. P. Tautu), pp. 216–237. Berlin: Springer-Verlag.

Nei, M. & Takahata, N. 1993 Effective population size, genetic diversity and coalescence time in subdivided populations. *J. molec. Evol.* **37**, 240–244.

Neigel, J. C. & Avise, J. C. 1993 Application of a random walk model to geographic distributions of animal mtDNA variation. *Genetics, Princeton* **135**, 1209–1220.

Neigel, J. E., Ball, R. M. & Avise, J. C. 1991 Estimation of single-generation migration distances from geographic variation in animal mitochondrial DNA. *Evolution* **45**, 423–432.

Rand, D. M., Dorfsman, M. & Kann, L. M. 1994 Neutral and non-neutral evolution of *Drosophila* mitochondrial DNA. *Genetics, Princeton* **138**, 741–756.

Slatkin, M. 1977 Gene flow and genetic drift in a species subject to frequent local extinctions. *Theor. Popul. Biol.* **12**, 253–262.

Slatkin, M. 1985 Rare alleles as indicators of gene flow. *Evolution* **39**, 53–65.

Slatkin, M. 1987 Gene flow and the geographic structure of natural populations. *Science, Wash.* **236**, 787–792.

Slatkin, M. 1991 Inbreeding coefficients and coalescence times. *Genet. Res.* **58**, 167–175.

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

Slatkin, M. & Maddison, W. P. 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics, Princeton* **123**, 603–613.

Slatkin, M. & Maddison, W. P. 1990 Detecting isolation by distance using phylogenies of genes. *Genetics, Princeton* **126**, 249–260.

Sokal, R. R., Oden, N. L., Legendre, P., Foster, M.-J., Kim, J., Thomson, B. A., Vauder, A., Harding, R. M. & Barbujani, G. 1990 Genetics and language in European populations. *Am. Nat.* **135**, 157–175.

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

Sokal, R. R. & Wartenberg, D. E. 1983 A test of spatial autocorrelation analysis using an isolation by distance model. *Genetics, Princeton* **105**, 219–237.

Templeton, A. R. 1992 Human origins and analysis of mitochondrial DNA sequences. *Science, Wash.* **255**, 737–737.

Turelli, M., Hoffmann, A. A. & McKechnie, S. W. 1992 Dynamics of cytoplasmic incompatibility and mtDNA variation in natural *Drosophila simulans* populations. *Genetics, Princeton* **132**, 713–723.

Wade, M. J. & McCauley, D. E. 1988 Extinction and recolonisation: their effects on genetic differentiation of local populations. *Evolution* **42**, 995–1005.

Whitlock, M. C. & McCauley, D. E. 1990 Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. *Evolution* **44**, 1717–1724.

Wright, S. 1943 An analysis of local variability in flower color in *Linanthus parryae*. *Genetics, Princeton* **28**, 139–156.

Wright, S. 1943 Isolation by distance. *Genetics, Princeton* **28**, 114–138.

Wright, S., Dobzhansky, T. & Hovanitz, W. 1942 Genetics of natural populations. VII. The allelism of lethals in the third chromosome of *Drosophila pseudoobscura*. *Genetics, Princeton* **27**, 363–394.