Genet. Res., Camb. (2007), **89**, *pp.* 475–477. © 2008 Cambridge University Press
doi:10.1017/S0016672308009683 Printed in the United Kingdom

Identity and coalescence in structured populations: a commentary on 'Inbreeding coefficients and coalescence times' by Montgomery Slatkin

N. H. BARTON*
IST Austria, Am Campus 1, A-3400, Klosterneuburg, Austria

The coalescent process describes the ancestry of a sample of genes as we trace them back through time: each pair of lineages coalesces in a common ancestor at a rate $1/2N_e$, where N_e is the effective size of the ancestral population through which the lineages were passing. This remarkably simple model for the ancestry of neutral genes, traced backwards in time, corresponds to the diffusion approximation, which describes how allele frequencies spread out as populations evolve forward in time.

The coalescent has come to dominate population genetics over the past 20 years, displacing the diffusion approximation that had been so important in the development of the neutral theory. The coalescent has proved so popular primarily because it gives a natural framework for analysing samples of DNA sequences and also because it allows efficient simulation, following samples of genes rather than the whole population.

Although the coalescent was described relatively recently (Kingman, 1982), the underlying ideas trace back to the beginning of population genetics. In the 1920s, Sewall Wright introduced F coefficients in order to quantify inbreeding, and Malécot (1948) clarified these using the concept of 'identity by descent', i.e. whether genes descend from the same gene in an ancestral reference population. Plainly, the probability of coalescence in any generation is just the difference between the probability of identity by descent in that generation and in the previous generation (eqn 1 in Slatkin, 1991). Thus, the classical idea of identity by descent is essentially the same as the modern concept of coalescence.

Slatkin's seminal paper set out the relation between the coalescent process and the classical theory, and used that relation to understand ancestry in spatially structured populations. Rather than working with the identity by descent, Slatkin used the distinct concept of identity in allelic state – the chance that two genes have not experienced any mutation since their divergence from a common ancestor. (Both 'identity by descent' and 'identity in state' are often referred to simply as 'identity', which makes the literature somewhat confusing). The identity in allelic state is $f=\sum_{t=0}^{\infty}(1-\mu)^{2t}P_t$, where P_t is the chance of coalescence at time t, and $(1-\mu)^{2t}$ is the chance that there was no mutation during the $2t$ generations during which the two genes have been diverging (eqn 21 in Slatkin, 1991). Approximating to continuous time, $f\approx\int_0^{\infty}\mathrm{e}^{-2\mu t}P(t)\mathrm{d}t$, which is just the Laplace transform of the distribution of coalescence times. Thus, there is a direct correspondence between the distribution of coalescence times and the identity in state: if we know one, we know the other. There is an extensive body of theory on identity in state in structured populations, initiated by Malécot, and elaborated by Maruyama and others. Slatkin (1991) showed how this carries over directly to give the distribution of coalescence times. (See Charlesworth *et al.*, 2003, for a review.)

Slatkin (1991) focuses on Wright's F_{ST} statistic, which measures the genetic diversity between populations, relative to the total diversity. It was defined by Wright (1951) as $(f_0-\bar{f})/(1-\bar{f})$, where f_0 is the identity between two genes within the same deme, and $\bar{f}$ is the identity between two randomly chosen genes. Here, 'identity' can refer to the identity by descent, relative to a reference population in the distant past, or identity in state, in the limit of low mutation rate; in either case, F_{ST} is independent of the time back to the reference population, or of the mutation rate. Moreover, Slatkin (1991) showed that F_{ST} can also be written in terms of the mean coalescence times between two genes from the same deme, $\bar{t}_0$, relative to the mean coalescence time for two randomly chosen genes, $\bar{t}_0$: $F_{ST}=(\bar{t}-\bar{t}_0)/\bar{t}$ (eqn 22 in Slatkin, 1991).

For an important class of models, the mean time since a pair of genes from within the same deme shared an ancestor is independent of population structure, and is just equal to the total number of genes in the whole population: $\bar{t}_0=2N_T$. This

* e-mail: n.barton@ed.ac.uk

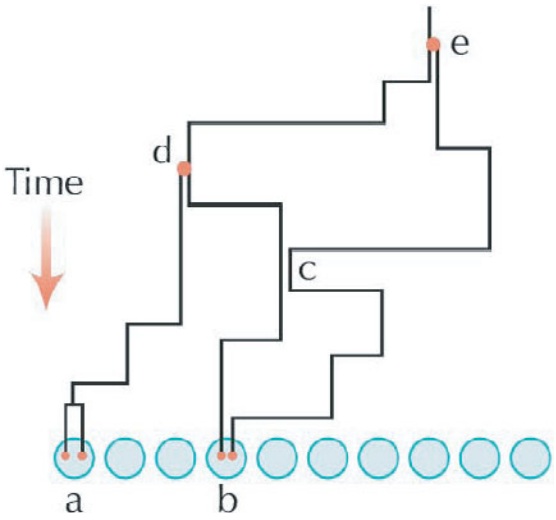

Fig. 1. Coalescence in the island model. Genes that are in the same deme (a) may trace back to a common ancestor within the same deme in the recent past. However, one or other lineage may escape (b) and wander about for a long time. Eventually, lineages come together in the same deme; they may then move apart again (c) or coalesce (d, e). From Barton *et al*. (2007).

remarkable invariance applies when migration is conservative (i.e. does not alter overall allele frequency; Strobeck, 1987; Hey, 1991). The average coalescence time between genes in different demes will be longer than that between genes in the same deme (so that $\bar{t} > \bar{t}_0 = 2N_T$); that is, population subdivision slows down the loss of genetic diversity by keeping it in separate, partially isolated, demes. However, in more realistic models, where population sizes fluctuate so that migration is not conservative, coalescence times within demes can be substantially reduced. In general, an unstable population structure is likely to reduce overall genetic diversity (Whitlock & Barton, 1997).

If we assume a stable population, with conservative migration, then we know that the mean coalescence time for two genes in the same deme is $\bar{t}_0 = 2N_T$, and so to find F_{ST}, we just need to find $\bar{t}_1$, the average time to coalescence between two genes that are in different demes. This is just the sum of the time taken for two lineages in separate demes to come together into the same deme, and the time for them to subsequently coalesce, $\bar{t}_0$; the first component depends only on the pattern of migration, and not on the process of coalescence within demes. The simplest case is the island model, in which each of d demes exchanges genes with a common gene pool at a rate m (Fig. 1). For this model, the extra time taken for genes in different demes to come together in the same deme is just $(d-1)/(2m)$ (eqn 12 in Slatkin, 1991). Slatkin uses the classical theory of identity in state to give the corresponding results for populations that are spread over one and two dimensions. This way of looking at the ancestry of neutral genes within a structured population emphasizes the separation between coalescence within demes (which may be fast, over a timescale equal to the number of genes within a local deme) and the movement of ancestral lineages among demes (which may be slow if there are very many demes). This separation of timescales has been exploited very fruitfully to gain a detailed understanding of coalescence in structured populations (Wakeley, 2008).

The use of genetic data to make inferences about population structure has a long history, going back to the collaboration between Wright and Dobzhansky (Lewontin *et al*., 1981). With the current flood of genetic data, this is now a large and thriving field. Slatkin's (1991) paper was seminal, in laying out the relation between old and new theory, and in showing how this theory could be used in practical inference. Nevertheless, there is still a divide in the field, between qualitative inferences made from genealogies at one or a few loci ('phylogeography'), and quantitative estimation, fitting models to data from multiple loci. There are problems with both approaches: on the one hand, genealogies are drawn from a highly random process, while on the other, specific models do not capture the messy, and largely unknown, reality of actual populations (Hey & Machado, 2003). So, it is essential to build an intuitive understanding of how structured populations evolve, which allows the consequences of a wide variety of processes to be understood. Slatkin (1991) remains an excellent starting point for this task.

References

Barton, N. H., Briggs, D. E. G., Eisen, J. A., Goldstein, D. B. & Patel, N. H. (2007). *Evolution*, Chapter 16. New York: Cold Spring Harbor Laboratory Press.

Charlesworth, B., Charlesworth, D. & Barton, N. H. (2003). The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology and Systematics* **34**, 99–125.

Hey, J. (1991). A multi-dimensional coalescent process applied to multiallelic selection models and migration models. *Theoretical Population Biology* **39**, 30–48.

Hey, J. & Machado, C. A. (2003). The study of structured populations – new hope for a difficult and divided science. *Nature Reviews Genetics* **4**, 535–543.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Lewontin, R. C., Moore, J. A., Provine, W. B. & Wallace, B. (1981). The scientific work of Th. Dobzhansky. In *Dobzhansky's 'Genetics of Natural Populations' I–XLIII* (ed. R. C. Lewontin, J. A. Moore, W. B. Provine & B. Wallace), pp. 93–114. New York: Columbia University Press.

Malécot, G. (1948). *Les Mathématiques de L'Hérédité*. Paris: Masson et Cie.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research* **58**, 167–175.

Strobeck, C. (1987). Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**, 149–153.
Wakeley, J. (2008) *Coalescent Theory: An Introduction*. Englewood, CO: Roberts and Company.
Whitlock, M. C. & Barton, N. H. (1997). The effective size of a subdivided population. *Genetics* **146**, 427–441.
Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.

https://doi.org/10.1017/S0016672308009683 Published online by Cambridge University Press