Heredity (2012) 108, 362–363

www.nature.com/hdy

Heredity

NEWS AND COMMENTARY

Gems from the *Heredity* archive

Estimating the migration rate from genetic variation data

M Yamamichi and H Innan

Heredity (2012) **108,** 362–363; doi:10.1038/hdy.2011.83; published online 21 September 2011

In nature, there would be no panmictic populations; any population is at least partially structured into subpopulations, which should be in different environments. Migration allows subpopulations to share genetic variation, which contributes to the maintenance of genetic variation within each subpopulation. Migration also enhances adaptation to local environments because some alleles could be adaptive in certain environments but not in others. Thus, to understand the evolutionary dynamics of a population, it is very important to quantify the level of migration. As it is a challenging task to directly estimate the migration rate in wild populations, it has been a common approach to use genetic variation data including microsatellites and single-nucleotide polymorphisms (SNPs) (Slatkin, 1985a; Neigel, 1997; Broquet and Petit, 2009).

Classically, Wright (1951) introduced F_{ST}, a summary statistic of population differentiation. F_{ST} measures the difference in heterozygosity among populations, which can be easily computed for any kind of polymorphism data. It is well known that the expectation of F_{ST} is given by $1/(1+4Nm)$ in the island model with equal effective sizes of subpopulations (N) and uniform migration rates among them (m). When Nm is large, F_{ST} is small because there is little difference in heterozygosity between subpopulations, while F_{ST} is large when Nm is small. Given the simple relationship between F_{ST} and Nm, F_{ST} is very frequently used for estimating Nm for various species (Holsinger and Weir, 2009, but see Whitlock and McCauley, 1999).

Slatkin (1981, 1985b) proposed an alternative idea to estimate Nm by using private alleles, which are defined as alleles that appear in the sample from only one subpopulation (Neel, 1973). Figure 1 illustrates a hypothetical situation with three subpopulations, where there are four private alleles (presented in black). In this highly cited gem from the Heredity archive, Barton and Slatkin (1986) obtained the analytical relationship between the average frequency of private alleles and Nm, suggesting that they are roughly in a linear correlation for a reasonable range of Nm. This simple relationship allowed the private allele frequencies to be a simple estimator of Nm, which has been commonly used for decades.

What is the difference between the two methods for estimating Nm? As both of them are summary statistics, they reflect only part of the data. In the ideal situation (that is, sampling with no errors from equilibrium populations under neutrality), the expectations of the estimates of Nm by the two methods would be the same, but when some assumptions are violated they would be different. The direction and extent of the bias caused by such violations have not been fully explored, but we can have some intuitive understanding. For example, private allele-based estimates of Nm should be most sensitive to recent migration because most private alleles are relatively rare (Slatkin and Takahata, 1985). Note that rare alleles are expected to be young. In contrast, F_{ST} is a summary statistic based on heterozygosity, which is largely determined by the frequencies of common alleles. Because common alleles are usually old, F_{ST} should reflect migration over a relatively longer time span. Because both estimators assume neutrality, any kind of selection will lead to bias. This bias could be stronger for one measure than for the other. For example, F_{ST} should be more sensitive to local adaptation because it causes a major shift in the common alleles frequencies. See Slatkin and Barton (1989) for more technical discussions on the difference between the two estimators.

Given the obvious importance of understanding population dynamics and evolution, these two simple methods for estimating Nm made significant contributions in ecology and evolution especially since the 1990s. They were applied to genetic variation data from a wide range of species, partly because the two methods are incorporated in the GENEPOP software (Raymond and Rousset, 1995). Thanks to dramatic improvement in

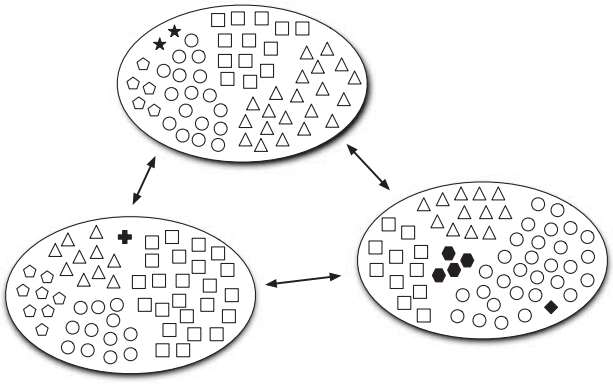

Figure 1 An illustration of the spatial distribution of shared (white) and private (black) alleles in a three-subpopulation model.

M Yamamichi and H Innan are at the Department of Evolutionary Studies of Biosystems, The Graduate University for Advanced Studies, Shonan village, Hayama, Kanagawa 240-0193, Japan.
E-mail: innan_hideki@soken.ac.jp

computational power, this field is shifting to depend more on computationally intensive methods using likelihood-based algorithms such as Markov chain Monte Carlo (MCMC) and approximate Bayesian computation (ABC) methods (Nielsen and Wakeley, 2001; Beaumont *et al*., 2002). Nevertheless, simple theoretical solutions for F_{ST} and private allele frequencies provide great intuitive understanding of migration and are useful in various situations. One interesting example is comparing estimates of *Nm* within a single genome, which gives significant insights into natural selection (Storz, 2005). If migration is defined as movements of individuals between populations, we should have similar estimates of *Nm* from different genomic regions, but this does not hold when selection is active. Consider two subpopulations, I and II, between which migration is allowed. Selection works on a particular biallelic locus with alleles A and B; A is favored in population I but disfavored in population II, and vise versa. Then, as B is preferentially selected out in subpopulation I and A is selected against in subpopulation II, those migrants are less likely to contribute to genetic admixture between the two subpopulations. In this situation, the migration rate is 'effectively' reduced because of less success in admixture. Other unlinked genomic regions are free from this selection, so that there would be no reduction in the effective migration rate, making a clear contrast to the selected locus. In genome-wide polymorphism data, thus, there could be heterogeneity in the 'effective' migration rate due to selection. There would also be cases where the effective migration rate is elevated at the selected gene. Suppose a new population emerges, in which A is assumed to be advantageous over B, then, there should be preferential migration of A into this new niche, resulting in an increased effective migration rate. This idea has been frequently used to scan a genome for evidence of selection, and there are many successful demonstrations of selection (for example Akey, 2009). For this kind of large-scale polymorphism data analysis, there are many situations where simple summary statistics are very useful and powerful. Thus, *Nm* tells not only about migration itself but also about the action of natural selection working on particular genomic regions, making it very important information in ecology and evolution.

Unfortunately, F_{ST} has been predominantly used as a summary statistic to describe the level of migration for a long time, but the amount of information we can obtain from F_{ST} alone is very limited. The proportion of private alleles is a useful second summary statistic. With dramatic improvement in computational power, the current trend is toward using as much information from data as possible. An example is the likelihood-based analysis under the isolation with migration (IM) models (Nielsen and Wakeley, 2001), in which the major focus is on the ratio of private to shared alleles. As more polymorphism data become available, this kind of computationally intensive method that does not fully rely on F_{ST} will have a central role.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Akey JM (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* **19**: 711–722.

Barton NH, Slatkin M (1986). A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity* **56**: 409–415.

Beaumont MA, Zhang W, Balding DJ (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

Broquet T, Petit EJ (2009). Molecular estimation of dispersal for ecology and population genetics. *Annu Rev Ecol Evol Syst* **40**: 193–216.

Holsinger KE, Weir BS (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat Rev Genet* **10**: 639–650.

Neel JV (1973). 'Private' genetic variants and the frequency of mutation among South American Indians. *Proc Natl Acad Sci USA* **70**: 3311–3315.

Neigel JE (1997). A comparison of alternative strategies for estimating gene flow from genetic markers. *Annu Rev Ecol Syst* **28**: 105–128.

Nielsen R, Wakeley J (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**: 885–896.

Raymond M, Rousset F (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Hered* **86**: 248–249.

Slatkin M (1981). Estimating levels of gene flow in natural populations. *Genetics* **99**: 323–335.

Slatkin M (1985a). Gene flow in natural populations. *Annu Rev Ecol Syst* **16**: 393–430.

Slatkin M (1985b). Rare alleles as indicators of gene flow. *Evolution* **39**: 53–65.

Slatkin M, Barton NH (1989). A comparison of three indirect methods for estimating average levels of gene flow. *Evolution* **43**: 1349–1368.

Slatkin M, Takahata N (1985). The average frequency of private alleles in a partially isolated population. *Theor Popul Biol* **28**: 314–331.

Storz JF (2005). Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol* **14**: 671–688.

Whitlock MC, McCauley DE (1999). Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4\,Nm+1)$. *Heredity* **82**: 117–125.

Wright S (1951). The genetical structure of populations. *Ann Eugen* **15**: 323–354.

Editor's suggested reading

Boulding EG, Culling M, Glebe B, Berg PR, Lien S, Moen T (2008). Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity* **101**: 381–391.

Lachish S, Miller KJ, Storfer A, Goldizen AW, Jones ME (2010). Evidence that disease-induced population decline changes genetic structure and alters dispersal patterns in the Tasmanian devil. *Heredity* **106**: 172–182.

Excoffier L, Hofer T, Foll M (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* **103**: 285–298.