Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics

Neil Davies, Francis X. Villablanca and George K. Roderick

Do salmon caught in the open ocean belong to an endangered stock? How can we determine the ethnicity of a corpse? Is this an endangered wolf or a domestic dog? Are medflies trapped in California resident or introduced; if introduced, where from? Answering these questions requires a robust means of assigning individuals to a particular source population – no easy task when there are multiple, genetically similar potential sources that were only recently established themselves. For example, genetic researchers seeking to describe the historical and contemporary population structure of invasive pest populations face both practical and theoretical challenges. Such species, that have recently occupied new regions in association with man, often lack genetic diversity as a result of the population bottleneck associated with colonization[1]. Even if this practical difficulty can be overcome, severe theoretical problems remain. First, new populations violate the assumptions of equilibrium integral to most population genetic theory (e.g. methods of estimating gene flow[2]). Second, phylogenetic analyses are inappropriate when gene trees represent evolutionary events that occurred in the ancestral range of a species, rather than its recent history in newly occupied areas. We will briefly review some of the new genetic markers that are providing adequate variation in previously intractable situations. We will then discuss how these new, as well as more traditional data, can be best analysed without using the equilibrium models of population genetics or the phylogenetic approaches of evolutionary biology.

The trials of youth

Mitochondrial DNA (mtDNA) is the most commonly used genetic marker in population studies, but the reduction in diversity associated with colonization bottlenecks is exacerbated for mitochondrial genes because they have only a quarter of the effective population size (N_e) of nuclear genes. In the medfly (*Ceratitis capitata*), for example, most New World populations are less than 100 years old and have a single high frequency and a single low frequency haplotype. In contrast, ancestral medfly populations in Africa display up to six haplotypes[3]. The low levels of diversity typical of invading medflies also limit the usefulness of allozymes[4,5]. As nuclear markers, allozymes retain more diversity through colonization bottlenecks

Recently founded populations represent an enormous challenge for genetic analysis: new populations are often genetically impoverished, making it hard to find sufficiently variable markers, and what little variation is present tends to be ancestral, rendering phylogenetic methods inappropriate. Recently, novel genetic markers and new statistical analyses have made multilocus genotyping an invaluable tool in the fledgling field of nonequilibrium population genetics. Such advances are not of mere academic interest but address questions of great economic, medical and conservation significance.
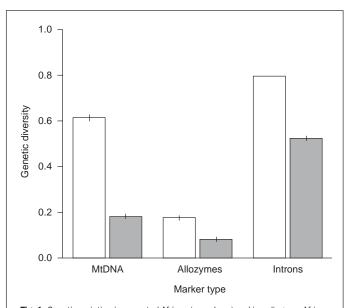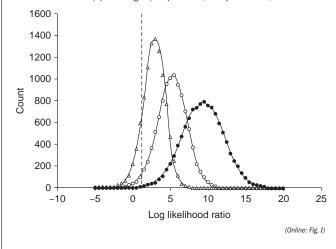
than mtDNA, but only a small amount of the underlying genetic variation at any given locus can be discerned using protein electrophoresis. Fortunately, markers are now available that assay nuclear DNA variation directly, such as microsatellites[6,7], introns[8], randomly amplified polymorphic DNA (RAPDs)[9,10] and restriction fragment length polymorphisms (RFLPs)[11]. These markers still reveal high levels of diversity when mtDNA and allozymes are relatively impoverished. For example, Baruffi *et al.*[12] compared RAPDs with allozymes in the medfly. Invading populations showed a fivefold decrease in allozyme variability compared with ancestral populations, whereas RAPDs showed only a twofold decrease. Similar results were reported by Villablanca *et al.*[13] for medfly introns (Fig. 1). In general, microsatellites also provide high expected heterozygosities, even in species or groups characterized by low levels of allozyme variation[14,15].

Fig. 1. Genetic variation in ancestral African (open bars) and invading non-African (filled bars) populations of medflies from mitochondrial DNA (mtDNA), allozymes and introns. *Reproduced, with permission, from Ref. 13.*

Box 1. Ethnic affiliation estimation

In human forensics, it is important to determine the ethnicity of corpses so that bone measurements can be used to calculate, for example, age at the time of death. Shriver *et al.*[28] used a maximum likelihood approach to assign ethnicity between Americans. Genotypes of population specific alleles (PSAs) with a high frequency in the Hispanic-American population were simulated on the basis of their frequency in Hispanic-Americans. The likelihood of each randomly generated genotype originating in the Hispanic population was then divided by its likelihood of being drawn from a population of European Americans. In the figure, the vertical line at the log likelihood ratio of 3 reflects a 1:1000 chance that the genotype is Hispanic. For five PSAs (filled circles), about half of the genotypes simulated from the Hispanic population were wrongly assigned as European-Americans. Power increased when ten (open circles) PSAs were used, but sufficient power to discriminate the two groups was considered to have been attained only when <95% of the simulated genotypes had <0.1% chance of occurring in another ethnic group; that is, when 20 PSAs were used (open triangles). *Reproduced, with permission, from Ref. 28.*

(Online: Fig. I)

The practical challenge of finding sufficiently variable markers might be solved, but the extreme youth of many populations still precludes certain types of analysis. In particular, traditional population-level approaches, such as F_{ST}, are of little use when assumptions of equilibrium are clearly violated; and phylogenetic methods, although the most powerful means of reconstructing the historical biogeography of relatively well established populations[16,17], cannot be used when most of the variation present in the populations of interest evolved in other regions (e.g. all mtDNA haplotypes of New World medflies are also found in ancestral Africa[3]). If the sorting of ancestral polymorphism prevents the phylogenetic analyses of mtDNA, then it will certainly be a problem in analyses of nuclear loci, which, because of their larger N_e, have longer coalescence times and hence are more likely to retain ancestral variation for longer periods[18]. In a phylogenetic approach[19,20] of medflies, for example, the intron gene trees of invading individuals represented allelic origin and evolution in Africa, rather than the history of invading populations[13]. It seems, therefore, that phylogenetic methods cannot be used to determine levels of gene flow among invading medfly populations or the origin of new outbreaks. Fortunately, hypervariable multilocus genotypes (multiple single locus markers) are spawning a new generation of statistical analyses that are well suited to recent, nonequilibrium situations.

Statistical approaches to multilocus genotypes

One of the first attempts to deduce origins using multilocus genotyping (mainly based on allozymes) was in fisheries management. Salmon catches often contain a mixture of fish from different sources and it is important to determine the species or populations being harvested. Salmon populations are often quite subdivided because of their loyalty to spawning in particular rivers. Consequently, variation between potential sources can sometimes be identified through allozyme electrophoresis. Statistical procedures, collectively known as mixed stock analysis (MSA), were developed to identify catch composition. These use a maximum likelihood approach to estimate the combination of potential source populations that best explain the allele frequencies observed in the catch: MSA does not assign individual fish to a particular source; instead, it estimates the most likely fraction of each source to yield the genotype frequencies observed in a catch (Utter and Ryman[21] have reviewed MSA in detail). MSA software, such as SPAM (Statistical Package for the Analysis of Mixtures, available from the Alaska Dept of Fish and Game), uses simulations to assess confidence.

Methods such as MSA that focus on populations are limited by the need to define those populations *a priori* and in doing so might miss individuals that have an unusual origin. For example, rare migrants will not always be detected in analyses based on allele frequencies because they represent only a tiny proportion of the resident population. An alternative approach is to use clustering algorithms based on the genetic distance between individuals. Human migration patterns have been studied in this way, with genetic distances calculated as the proportion of shared alleles (the shared-allele index) across RFLP (Ref. 22) or microsatellite[23] loci. Most individuals from the same region clustered together, although some individuals were outliers or clustered with individuals from a different region. The shared-allele index is useful for identifying individuals as potential hybrids and/or migrants. However, the main drawback of this approach is in evaluating statistical significance. Fortunately, the status of potential hybrids or migrants can be determined using an assignment test, where important developments in assessing statistical significance have been made.

Assignment tests are based on multilocus genetic data and use both individual genotypes and population level allele frequencies. The first assignment test was developed by Paetkau *et al.*[14] and simply calculated the likelihood of drawing a single multilocus genotype from several potential sources based on the observed allele frequencies at each locus in each source. Loci were assumed to be unlinked and independent. The main limitation of this early assignment test was its inability to assess the significance of the assignments. The only source of error considered was that the inclusion of the test genotype in any single population would bias the result in favor of assigning that population as the source. To remedy this problem, Paetkau *et al.*[14] added the test genotype to all populations. The correction simultaneously prevented alleles from having a frequency of zero: for diploid loci, nonobserved test alleles were taken to occur with a frequency of 1/2N+1 (where N is the number of individuals sampled). The weakness of this approach is that the frequency of nonobserved test alleles (for the purposes of the assignment test) is determined by sample size. Poorly sampled populations will appear to be the more likely source because nonobserved test alleles will be allocated a relatively high frequency in the test. An alternative approach that is less biased by differences in sample size is to assume that nonobserved test alleles have some extremely low frequency, in the order of 1×10^{-6}, irrespective of sample size (N. Davies *et al.*, unpublished).

Several exciting applications of the assignment test based on microsatellite data were recently described by Waser and Strobeck[24]. The test appears particularly useful for

determining the origin of an individual when there are multiple, genetically similar candidate sources. For example, the chinook salmon, *Oncorhynchus tshawytscha*, occurs as different temporal populations (runs) that spawn in the same river but at different times of the year. Runs are genetically very similar but have such distinct life histories that some are considered separate (and endangered) species. Microsatellites have been used to assign chinook salmon to particular runs[25] and it is hoped that catches might one day be monitored routinely by multilocus genotyping – fishing can be curtailed when an endangered run is caught in unacceptable numbers. As the examples presented by Waser and Strobeck[24] clearly demonstrate, multilocus genetic data provide a means of assigning individuals to particular sources and of assessing interpopulation dispersal. There are several important problems, however, with the assignment test originally proposed by Paetkau *et al.* In particular, how many loci are needed to assign confidently individuals to a source, and how can the statistical significance of those assignments be determined? Great progress has been made in this area as assignment tests have become more sophisticated. In the following sections, we discuss these developments and suggest directions for further improvements.

How many loci?

The benefits of using more loci have been demonstrated in hybridization studies. It is often important to determine whether some individuals are hybrids of two different species. For example, hybridization between endangered Ethiopian wolves (*Canis simensis*) and domestic dogs (*C. familiaris*) could eventually dilute the Ethiopian wolf genepool to extinction. Microsatellites and RFLPs have been used to evaluate this threat[26] and the ability to recognize hybrid individuals increases

---

Box 2. Determining the origin of a multilocus genotype

What is the likelihood that a genotype sampled in population A is an immigrant from population B?

In Rannala and Mountain's[29] computer program, IMMANC (available at http://mw511.biol.berkeley.edu/homepage.html) the log likelihood of the genotype being drawn from population B is subtracted from the log likelihood of it being drawn from population A. This yields a value (Ln Λ) that is negative if the individual is more likely to be an immigrant from B than a resident of A. Monte Carlo simulations estimate the significance of Ln Λ. G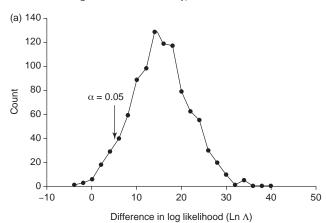enotypes are simulated at random from population A, conditioned on the allele frequencies observed in that population. Alleles are drawn from a (Dirichlet) distribution of possible allele frequencies – increased sample size narrows the distribution of possible allele frequencies. The Ln Λ values are calculated for simulated genotypes based on the observed allele frequencies in population A. In this way, one can visualize the distribution of Ln Λ (a). The test genotype can be considered an immigrant if Ln Λ is negative and <5% of the genotypes randomly generated from population A yield more negative Ln Λ (i.e. if it falls in the critical region to the left of the arrow).
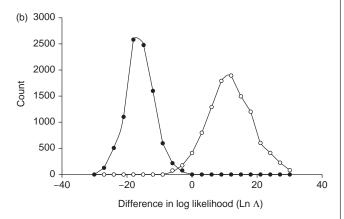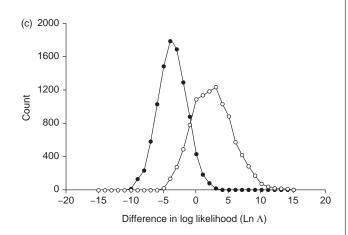
Power test for identifying immigrants

The relative likelihood of a genotype originating in population A rather than in population B (Ln Λ) is calculated for genotypes drawn randomly from the allele frequency distributions of the A (open circles) and B (filled circles) populations (b). As expected, genotypes from population A tend to yield positive Ln Λ values, indicating that they are more likely to come from population A. Conversely, genotypes simulated from population B yield negative Ln Λ values indicating that they are more likely to be immigrants from B than residents of A. Greater overlap of the two distributions demonstrates less power – that is, it is more difficult to diagnose the origin of any given genotype.

Reduced power to distinguish individuals with hybrid ancestry

The relative likelihood that a genotype originated in population A rather than being the F1 hybrid offspring individuals from population A and B (Ln Λ) is calculated for genotypes drawn randomly from the allele frequency distributions of the A (open circles) and hybrid (filled circles) populations (c). In F1 hybrids, one allele at each locus is randomly drawn from population A and the other allele at each locus is randomly drawn from population B. Clearly, this is a more difficult problem than distinguishing the two parental types (b) and the reduced statistical power is indicated by the large degree of overlap between the two distributions – that is, a large proportion of genotypes drawn from population A appear to be hybrids and *vice versa*.

Figures redrawn, with permission, from Ref. 29.

(Online: Fig. I)

as the number of loci used and the differences in allele frequencies that exist among the parental populations increase. Boecklen and Howard[27] determined the number of markers needed to distinguish nonhybrids, F1 hybrids and backcrosses, assuming that fixed genetic differences can be identified between the two species. Calculating the probability that a given proportion of the loci are heterozygous for each hybrid class, they found that five markers were usually sufficient for coarse classification, whereas recognizing advanced backcrosses required as many as 70 loci.

Assignment tests are generally employed at the population level, and fixed differences might not exist between recently isolated populations or those united by relatively high levels of gene flow. In such cases, it is hard to assign individuals that have yet to breed, let alone to recognize individuals with hybrid ancestry. For example, in human forensics, the number of nonfixed markers needed to assign individuals to different ethnic groups successfully was explored by Shriver et al.[28] (Box 1). A set of markers that differed in frequency between pairs of populations by at least 50% was selected. Twenty such markers were needed to differentiate genetically similar European- and Hispanic-Americans, whereas only five were needed to distinguish more distinct African- and European-Americans. Rannala and Mountain[29] used their assignment test (Box 2) to calculate the likelihood that an individual had immigrant ancestors d generations in the past. Comparing genetically distinct human populations (from Japan and Senegal), they found that 60 loci were needed to recognize individuals with a foreign grandparent ($d = 2$).

Power and significance in assignment tests

Knowing when one has a sufficient number of suitably variable markers relies on some means of assessing the significance of assignments. There are two major sources of error associated with the population level data used in an assignment test. First, observed allele frequencies are estimates, so sampling error must be considered. Second, differences in genetic diversity between potential source populations can cause a bias because the likelihood of drawing any genotype is inversely correlated with the diversity of the population from which it is drawn. If an individual originated in a very diverse source, there is a risk that it will be wrongly assigned to an alternative, more homogeneous population.

A good example of this was noted by Bogdanowicz et al.[30] North American populations of gypsy moth, *Lymantria dispar*, have recently been invaded by conspecific moths from Asia. The Asian moths are readily distinguished by mtDNA, but the introgression of Asian alleles cannot be determined solely on this basis because mtDNA ignores paternal gene flow. Bogdanowicz et al.[30], therefore, assayed nuclear variation using microsatellites. They found that several moths collected in Asia were assigned to North America even though they were unlikely to be immigrants. The result was probably an artifact of moths in North America having only a subset of the alleles found in Asia. Therefore, some genuine Asian genotypes have a higher probability of being assigned to North America than their native Asia.

Rannala and Mountain's[29] assignment test (Box 2) takes into account the sampling error associated with estimating allele frequencies and the differences in diversity between two potential sources. Other sources of error – in particular those associated with genotyping individuals, the existence of more than two candidate sources, and the proportion of those candidates sampled – also need to be considered and included in future assignment tests. These issues have received the most attention in kinship studies.

There are many similarities between the population level assignment tests that are the focus of this article and likelihood-based paternity tests. An excellent review of recent developments in paternity testing is provided by Marshall et al.[31] who also present a computer program, CERVUS (available at http//helios.bto.ed.ac.uk/evolgen/), that assesses the significance of paternity tests based on multilocus genetic data. Marshall et al. consider two sources of error that are also likely to be important in population level assignment tests. First, there can be experimental errors associated with the scoring of individual genotypes. Mismatches between offspring and candidate males rule out those males as the father, but these mismatches will sometimes result from laboratory scoring error. In paternity tests, this error rate can be estimated by the frequency of mismatches between mother and offspring (i.e. when they have no alleles in common). The error rate for allozymes is thought to be around 1% (Ref. 32). Marshall et al.[31] used simulations to assess the effect of typing errors in a paternity test. Assuming the genetic data to be error free always led to an overestimation of confidence. Including an error rate parameter sometimes prevented the exclusion of the true father, but it also meant that fewer nonfathers were excluded. A second source of bias considered by Marshall et al. was the number of potential sources (i.e. candidate males) and the proportion that have been sampled. Perhaps unsurprisingly, the fewer candidate males available, the higher the success rate in paternity assignment. Success rates also increased with the proportion of candidates sampled.

Conclusions and future directions

Genetic markers are now available that provide ample variation even in the impoverished populations that often characterize recent invasions. Phylogenetic approaches to assign individuals to particular populations are often inappropriate when the candidate sources are themselves only recently founded. Alternative methods have been developed, the most promising of which is the assignment test. A major limitation of early assignment tests was their inability to assess statistical significance. We have shown that the combination of two statistical methods developed by Rannala and Mountain[29] and Marshall et al.[31] would provide an assignment test that takes into account most of the sources of error associated with assigning individuals to various source populations. Neither of these tests, however, considers the effect of genetic drift – the drift of source populations can be ignored when migration has occurred from well established sources very recently. However, drift might become important if historical as well as contemporary populations are potential sources. For example, medflies caught this year in California could be derived from existing populations in Latin America, or from previous medfly outbreaks in California. The drift of historical Californian populations must be considered when assessing their likelihood as a source.

Assignment tests provide a means of identifying migrants and recent gene flow without relying on assumptions of equilibrium that are clearly violated in many recently founded invading populations. To date, assignment tests have assumed that loci are independent, but the consideration of linked loci might provide more power to distinguish populations. For example, the combined mutation rate of several linked microsatellites could provide resolution over

tens of generations; although not all microsatellite mutations yield novel alleles – linked microsatellite loci might generate novel genotypes (S.R. Palumbi, pers. commun.). Ideally, one would like to have DNA (or RNA) sequences that evolve fast enough for unique genetic signatures to arise in populations after only a few generations of isolation. One possibility might be to use extragenomic markers (EGMs), such as viruses. Studies in molecular epidemiology have avoided some of the problems discussed here because of the speed at which viruses evolve: 50 years for an RNA virus can equal some 50 million organism years[33]. Importantly, viruses sometimes reflect the history of their hosts; for example, human papillomavirus type 16 follows human ethnicity rather than geography[34], and a close relationship between Melanesian and Australian aboriginals is suggested by the unusual human T-cell lymphotropic virus type I that they share[35]. The challenge for the future will be to develop approaches that harness the power of phylogenetic analyses to address ecological questions in nonequilibrium situations.

Acknowledgements

References

1 Nei, M., Maruyama, T. and Chakraborty, R. (1975) **The bottleneck effect and genetic variability in populations,** *Evolution* 29, 1–10

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

3 Gasparich, G.E. *et al.* (1997) **Population genetic structure of Mediterranean fruit fly (Diptera: Tephritidae) and implications for worldwide colonization patterns,** *Ann. Entomol. Soc. Am.* 90, 790–797

4 Huettel, M.D. *et al.* (1980) **Genetic effects of multiple population bottlenecks in the Mediterranean fruit fly (*Ceratitis capitata*),** *Genetics* 94, S47–S48

5 Roderick, G.K. (1996) **Population genetic studies of tephritid flies of economic importance,** in *Fruit Fly Pests. A World Assessment of Their Biology and Management* (McPheron, B.A. and Steck, G.J., eds), pp. 267–271, St Lucie Press

6 Weber, J.L. and May, P.E. (1989) **Abundant class of human DNA polymorphism which can be typed using the polymerase chain reaction,** *Am. J. Hum. Genet.* 44, 388–396

7 Queller, D.C., Strassmann, J.E. and Hughes, C.R. (1993) **Microsatellites and kinship,** *Trends Ecol. Evol.* 8, 285–288

8 Palumbi, S.R. and Baker, C.S. (1994) **Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales,** *Mol. Biol. Evol.* 11, 426–435

9 Williams, J.G.K. *et al.* (1990) **DNA polymorphisms amplified by arbitrary primers are useful as genetic markers,** *Nucleic Acids Res.* 18, 6531–6535

10 Welsh, J. and McClelland, M. (1990) **Fingerprinting genomes using PCR with arbitrary primers,** *Nucleic Acids Res.* 18, 7213–7218

11 Aquadro, C.F., Noon, W.A. and Begun, D.J. (1992) **RFLP analysis using heterologous probes,** in *Molecular Genetic Analysis of Populations* (Hoelzel, A.R., ed.), pp. 115–158, Oxford University Press

12 Baruffi, L. *et al.* (1995) **Polymorphism within and between populations of *Ceratitis capitata*: comparison between RAPD and multi-locus enzyme electrophoresis data,** *Heredity* 74, 425–437

13 Villablanca, F.X., Roderick, G.K. and Palumbi, S.R. (1998) **Invasion genetics of the Mediterranean fruit fly: variation in multiple nuclear introns,** *Mol. Ecol.* 7, 351–365

14 Paetkau, D. *et al.* (1995) **Microsatellite analysis of population structure in Canadian polar bears,** *Mol. Ecol.* 4, 347–354

15 Allendorf, F.W. *et al.* (1979) **Electrophoretic variation in large mammals. 1 The polar bear, *Thalarctos maritimus*,** *Hereditas* 91, 19–22

16 Avise, J.C. (1994) *Molecular Markers, Natural History, and Evolution*, Chapman & Hall

17 Roderick, G.K. (1996) **Geographic structure of insect populations: gene flow, phylogeography, and their uses,** *Annu. Rev. Entomol.* 41, 263–290

18 Moore, W.S. (1995) **Inferring phylogenies from mtDNA variation: mitochondrial gene trees versus nuclear gene trees,** *Evolution* 49, 718–726

19 Templeton, A.R., Routman, E. and Phillips, C.A. (1995) **Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*,** *Genetics* 140, 767–782

20 Templeton, A.R. (1998) **Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history,** *Mol. Ecol.* 7, 381–397

21 Utter, F. and Ryman, N. (1993) **Genetic markers and mixed stock fisheries,** *Fisheries* 18, 11–21

22 Mountain, J.L. and Cavalli-Sforza, L.L. (1997) **Multilocus genotypes, a tree of individuals, and human evolutionary history,** *Am. J. Hum. Genet.* 61, 705–718

23 Bowcock, A.M. *et al.* (1994) **High resolution of human evolutionary trees with polymorphic microsatellites,** *Nature* 368, 455–457

24 Waser, P.M. and Strobeck, C. (1998) **Genetic signatures of interpopulation dispersal,** *Trends Ecol. Evol.* 13, 43–44

25 Banks, M.A., Baldwin, B.A. and Hedgecock, D. (1996) **Research on chinook salmon (*Oncorhynchus tshawytscha*) stock structure using microsatellite data,** *Bull. Nat. Res. Inst. Aquacult.* (Suppl.) 2, 5–9

26 Gotelli, D. *et al.* (1994) **Molecular genetics of the most endangered canid: the Ethiopian wolf *Canis simensis*,** *Mol. Ecol.* 3, 301–312

27 Boecklen, W.J. and Howard, D.J. (1997) **Genetic analysis of hybrid zones: numbers of markers and power of resolution,** *Ecology* 78, 2611–2616

28 Shriver, M.D. *et al.* (1997) **Ethnic-affiliation estimation by use of population-specific DNA markers,** *Am. J. Hum. Genet.* 60, 957–964

29 Rannala, B. and Mountain, J.L. (1997) **Detecting immigration using multilocus genotypes,** *Proc. Natl. Acad. Sci. U. S. A.* 94, 9197–9201

30 Bogdanowicz, S.M. *et al.* (1997) **Microsatellite DNA variation among Asian and North American gypsy moths (Lepidoptera: Lymantriidae),** *Ann. Entomol. Soc. Am.* 90, 768–775

31 Marshall, T.C. *et al.* (1998) **Statistical confidence for likelihood-based paternity inference in natural populations,** *Mol. Ecol.* 7, 639–655

32 Lathrop, G.M. *et al.* (1983) **Evaluating pedigree data. I The estimation of pedigree error in the presence of marker mistyping,** *Am. J. Hum. Genet.* 35, 241–262

33 Holland, J. *et al.* (1982) **Rapid evolution of RNA genomes,** *Science* 215, 1577–1585

34 Ho, L. *et al.* (1993) **The genetic drift of human papillomavirus type 16 is a means of reconstruction prehistoric viral spread and the movement of ancient human populations,** *J. Virol.* 67, 6413–6423

35 Yanagihara, R. (1994) **Geographic-specific genotypes or topotypes of human T-cell lymphotropic virus type I as markers for early and recent migrations of human populations,** *Adv. Virus Res.* 43, 147–186