Molecular Ecology (2004) doi: 10.1046/j.1365-294X.2003.02031.x

Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species

JODY HEY,* YONG-JIN WON,* ARJUN SIVASUNDAR,* RASMUS NIELSEN† and JEFFREY A. MARKERT‡

**Department of Genetics, Rutgers the State University of New Jersey, Piscataway, NJ 08854, †Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853–780, USA, ‡Department of Biological Sciences, University of Cincinnati, Cincinnati, OH 45221, USA*

Abstract

When populations or species have recently separated they often share genetic variation. However, it can be difficult to determine whether shared polymorphisms are the result of gene flow, the result of the persistence of variation in both populations since the time of common ancestry, or both of these factors. We have developed an empirical protocol for using loci that include unique nuclear DNA sequence haplotypes together with linked microsatellites or short tandem repeats (STRs). These 'HapSTRs' offer the potentially high resolution associated with the high mutation rate of STRs, together with the advantages of low homoplasy of unique sequence DNA. We also describe a new procedure for estimating the likelihood of HapSTR data under an Isolation with Migration model. An example using Cichlid fishes from Lake Malawi is described. The analysis suggests that the species have been exchanging genes since the time they began to diverge.

Keywords: cichlid, gene flow, Markov chain Monte Carlo, microsatellites, phylogeography, speciation, sympatric speciation

Received 10 July 2003; revision received 3 October 2003; accepted 3 October 2003

Introduction

Recently separated populations or species present a great challenge for population geneticists who wish to understand the role that gene flow plays in divergence. Such understanding is particularly important if we are to appreciate the effect of natural selection on divergence. If newly separated populations are completely allopatric, then of course there is no gene flow and divergence proceeds by genetic drift and the selective fixation of favoured mutations. Over time such fixations may have, as a by-product, the consequence that the hybrids are sterile should the populations come into contact. Under this standard Dobzhansky-Muller model of species formation (Dobzhansky 1936; Muller 1940; Orr 1996) natural selection plays an indirect role in forming the species barrier. Alternatively, if populations are not completely separated and gene flow is occurring at even moderate levels, then the populations will merge and divergence will not accrue. Thus natural selection that acts against gene flow, mediated either by low hybrid fitness or by environmental differences between the populations, can enable divergence. In short, to understand the role that natural selection plays in contributing to population divergence, we must be able to assess gene flow. In recent years there has been a resurgence of theoretical interest in parapatric and sympatric speciation (Dieckmann & Doebeli 1999; Higashi *et al*. 1999; Noor *et al*. 2001; Navarro & Barton 2003), and this has been accompanied by empirical findings that closely related species have been exchanging genes (Rieseberg *et al*. 1999; Machado *et al*. 2002; Shaw 2002).

Investigators of the demographic history of closely related populations or species often turn to microsatellite or short tandem repeat (STR) loci to address questions about divergence and gene flow. Common and often highly heterozygous, STR loci have become the focus of a large body of population genetic- and phylogeography-based research (Balloux & Lugon-Moulin 2002; Zhang & Hewitt 2003). Yet researchers using these allelic markers are confronted with two potential difficulties. First, it is fundamentally difficult, with any kind of data, to determine

Correspondence: J. Hey. E-mail: hey@biology.rutgers.edu

whether an observed level of differentiation between populations results from a balance between genetic drift and gene flow, or if it is simply a result of population history in the absence of present day gene flow (Broughton & Harrison 2003). Demographic events such as a recent bottleneck in one or both populations can transiently increase estimates of genetic divergence, even in the presence of substantial gene exchange, whereas vicariance events involving large subpopulations on either side of an impermeable barrier could falsely be interpreted as the signature of extensive migration (Zhivotovsky 2001). The second problem that applies especially to STR loci arises from their replication-slippage-based mutation process and their high mutation rates that together lead to high levels of homoplasy in samples drawn from natural populations. Sequencing studies have indicated significant levels of homoplasy for STR allele length (Jarne & Lagoda 1996; Brohede & Ellegren 1999; Colson & Goldstein 1999; Van Oppen *et al*. 2000; Estoup *et al*. 2002).

In practice, the severity of both problems is often moderated by the addition of supplementary information. An understanding of the biogeographic, demographic and morphological contexts may be incorporated into the analysis, or STR data may be supplemented with data derived from additional markers with different sensitivities to historical and demographic processes. For example, Bulgin *et al*. (2003) used knowledge about the population size of the Florida grasshopper sparrow to argue that retained ancestral polymorphisms, rather than high levels of gene flow, are responsible for low levels of mitochondrial and STR divergence between Florida and northern populations. Saint-Laurent *et al*. (2003) combined both STR and mitochondrial DNA sequence data with an analysis of morphological data to demonstrate the importance of selection in maintaining morphological divergence in the face of gene flow between two morphologically similar pairs of rainbow smelt. Surridge *et al*. (1999) supplemented an STR-based analysis of isolation by distance in European wild rabbits with other information on social stucture and historical bottlenecks caused by the rabbit disease myxomatosis to argue that historical patterns of population structuring are obscured in this species.

An alternative to using STRs to study recent population divergence is to focus solely on nuclear DNA sequence haplotypes. Such data often show low levels of homoplasy by mutation and it can be possible to assess the role of gene flow in the divergence process (Kliman *et al*. 2000; Machado & Hey 2003). However, nuclear gene mutation rates are very low, and such data are not expected to offer much resolution for populations that have diverged very recently.

The contrasting mutation rates and mutation processes of nuclear sequence and STR loci suggest the possibility that, if they could be used jointly, they may offset each other's limitations. If data from an STR locus can be considered jointly with the linked DNA sequence data, then together the two loci should permit a higher resolution of recent events. Here we introduce a series of protocols for the isolation and analysis of loci consisting of an STR and a linked sequence. We call such loci 'HapSTRs', a conjunction of 'haplotype' and 'STR', inspired by 'SNPSTR' which is a locus that includes an STR and one or more linked sites that have single nucleotide polymorphisms (SNPs) (Mountain *et al*. 2002). HapSTRs also share a rationale and some practical advantages of the use of pairs of tightly linked STRs or juxtaposed microsatellite systems (JMS) (Estoup *et al*. 1999; Estoup *et al*. 2000).

Methods and materials

Field methods

Individuals of *Tropheops tropheops* and *T. gracilior* were collected from the waters adjacent to Otter Island and Harbour Island, near Cape Maclear, Malawi. The collection sites are separated by about 18 km along the coast. These species are just two of hundreds of 'mbuna' (rock-dwelling cichlids of Lake Malawi) species that dwell in rocky shoreline habitats. Following capture, using SCUBA gear and fine-mesh monofilament nets (Ribbink *et al*. 1983), fish were brought to the surface and anaesthetized in 40 mg/L Finquel (Argent Laboratories) buffered with an equal amount of Gram Pac phosphate buffer pH 7.41 (Fisher Scientific). A small (~0.5 cm^2) piece of the dorsal fin was removed from tranquillized fish using surgical scissors and immediately placed in a preservative solution containing 20% dimethyl sulphoxide, 0.25 M ethylenediaminetetraacetic acid (EDTA) and saturated NaCl (Seutin *et al*. 1991). Following recovery fish were returned to the lake near their original site of capture.

STR isolation and genotyping

A dinucleotide STR locus (U66814; GenBank accession no. AAU66814) developed from cichlids from Lake Victoria (Booton *et al*. 1996) was examined in this study. Because the original report included only a short region of sequence surrounding the STR, an inverse polymerase chain reaction (PCR) was used to isolate additional flanking DNA (Ochman *et al*. 1988).

Genomic DNAs were extracted from fish tissue (~20 mg) using the DNeasy Tissue Kit (Qiagen) according to the manufacturer's instructions. For genotyping the PCR was performed under the following conditions: 10–50 ng of template DNA, 2 μL 10× buffer (Promega), 1 μL 10× bovine serum albumin (1 mg/mL), 1.6 μL $MgCl_2$ (2.5 μM), 0.8 μL of a primer pair (10 μM stock concentration; U66814F-M13F: 5′-CAC GAC GTT GTA AAA CGA CTC ACT GAA

GGA CAA AGC AGG A-3′; U66814R: 5′-TTT TGC AGT AAT CCA CCG GA-3′), 0.7 μL M13 forward primer labelled with IRDye 800 (LiCor), 1 unit of *Taq* DNA polymerase (Promega), 1.6 μL of a 2-mM stock solution of dNTPs, and sterile H_2O to a final volume of 20 μL. PCR were carried out under the following conditions for 35 cycles on a thermocycler (MJ Research): 95 °C for 30 s, 52 °C for 30 s and 72 °C for 40 s, followed by a final extension at 72 °C for 2 min. PCR products were separated on a 7% denaturing polyacrylamide gel on a LiCor 4200 sequencer (LiCor), and were analysed using SAGA software (LiCor).

Sequencing of individual haplotypes

Heterozygosity for PCR fragment length, caused by differences in repeat number of the STR, was the basis for isolation of individual haplotypes (Blankenship *et al*. 2002). DNA fragments where amplified as above using a different left-side primer (U66814-LF-M13R: 5′-GGA TAA CAA TTT CAC ACA GGT GGA TGG TTT CCT TAC ACA GC-3′) together with U66814R (same as above). These primers span approximately 900 base pairs (bp) including about 650 bp of flanking region. The fragments were separated on 5% denaturing acrylamide gel for 8 h at 1200 V. Single-strand DNA bands were visualized using SYBR Gold (Molecular Probe) stain and the Dark Reader system (Clare Chemical Research). Bands were excised and incubated in 20 μL 1× TE buffer (10 mM Tris–HCl, 1 mM EDTA, pH = 8.0) buffer. This DNA was used in PCR containing 1 μL template DNA, 2 μL 10× buffer (supplied by the manufacturer), 0.4 μL $MgCl_2$ (2.5 μM), 0.4 μL of a primer pair (10 μM stock concentration; U66814-LF-M13R: same as above; U66814-L-M13F: 5′-CAC GAC GTT GTA AAA CGA CTC CTG CTT TGT CCT TCA GTG A-3′), 0.5 units of HotStar *Taq* DNA polymerase (Qiagen), 1.6 μL of a 2-mM stock solution of dNTPs, and sterile H_2O to final volume of 20 μL. After a 15-min activation period at 95 °C, amplifications were performed with the following steps: 32 cycles at 95 °C for 30 s, 50 °C for 30 s and 72 °C for 1 min, followed by a final extension at 72 °C for 7 min. These PCR products were used directly as template for bi-directional sequencing on a LiCor 4200 sequencer using dye-labelled M13 forward and reverse primers (LiCor) according to SEQUITHERM EXCEL II sequencing protocols (Epicentre).

Theory

Under the basic Isolation with Migration (IM) model (Takahata & Slatkin 1990; Wakeley & Hey 1998; Nielsen & Wakeley 2001), an ancestral population of size N_A splits into two descendant populations of sizes N_1 and N_2 at time t generations ago. Following separation, gene flow occurs at rates m_1 and m_2. Figure 1 depicts the model together with the basic parameters. Also shown are the parameters scaled by the mutation rate u, which are used in the model fitting. This is a fairly general model that includes several boundary cases that are also of interest. If the migration rates are zero, then this is a simple Isolation model (Takahata & Nei 1985; Hey 1994; Wakeley & Hey 1997) and may be appropriate for studying the divergence of populations under allopatry (Won *et al*. 2003). If migration rates are not zero and the time of splitting was a very long time ago, then this is a simple two-island model (Wright 1931) and can be used to study the countervailing forces of genetic drift and gene flow, and the equilibrium between them. If one of the descendant populations has a size of zero and there is no gene flow then the model becomes one of population size change at t generations ago. Finally, if the time of splitting was so recent that, in effect, there has not been a separation then the model becomes one of a single population.

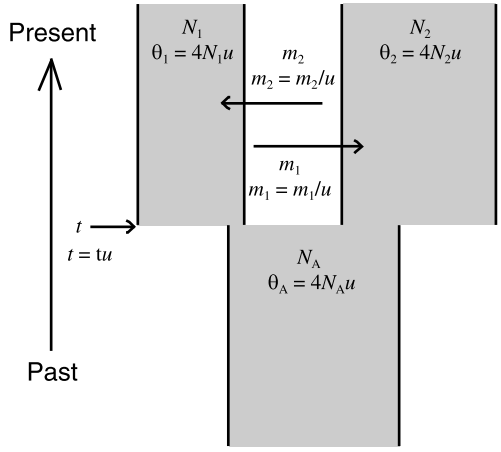

Fig. 1 The Isolation with Migration (IM) model is depicted with two parameter sets. The basic demographic parameters are constant effective population sizes (N_1, N_2 and N_A), gene flow rates per gene copy per generation (m_1 and m_2), and the time of population splitting at t generations in the past. The second set of parameters are all scaled by the neutral mutation rate u, and it is these parameters that are actually used in the model fitting.

With six parameters, the IM model presents considerable challenges in application. For the simpler Isolation model with four parameters (i.e. no migration), fairly simple moment estimates are available and can be calculated using a few summary statistics calculated from DNA sequence data sets (Wakeley & Hey 1997). For the full IM model, there do not yet exist closed form expressions that can be used as estimators. Nielsen and Wakeley described a likehood/Bayesian approach that uses a Metropolis–Hastings, Markov Chain Monte Carlo algorithm (Nielsen & Wakeley 2001).

The method of Nielsen and Wakeley belongs to a family of newer methods that makes explicit use of the genealogical (i.e. gene tree) information in the data (Griffiths & Tavare 1994; Kuhner *et al*. 1995; Wilson & Balding 1998; Felsenstein *et al*. 1999). Unlike traditional phylogeographic approaches, these methods do not rely on one gene-tree estimate but rather consider all of the possible genealogies that are consistent with a data set. By doing so, these methods overcome the critical shortcoming of methods that are based on individual tree estimates, and that is the inherent stochastic variance that causes different genes to have widely varying histories even when they all come from the same samples of organisms (Broughton & Harrison 2003).

Consider a data set X collected from n individuals, and a set of model parameters Θ (e.g. a set of six numbers corresponding to the parameters of the IM model), and a gene tree G that is consistent with the data under a particular mutation model. If G is a full representation of both the topology and the branch lengths of a tree, then it is possible to calculate the probability of the data, $P(X \mid G,\Theta)$. The details of this calculation will depend on the mutation model. For example, under an infinite sites mutation model (Kimura 1969), this probability is a simple product of $n-1$ Poisson variables (i.e. one for each branch of the genealogy). It is also possible for many demographic models, including the IM model, to calculate the probability of a particular tree, given a set of parameter values, $P(G \mid \Theta)$. Then the posterior probability of the parameters, given the data, can be found using:

$$P(\Theta \mid X) = cf(\Theta)\int_G P(X \mid \Theta)P(G \mid \Theta)dG \qquad (1)$$

where c is a constant of proportionality to ensure that the total probability for all values of Θ sums to one and $f(\Theta)$ is the joint prior probability density of the parameters. In principle the prior distribution can be set to reflect actual prior information regarding Θ, however, for the present purposes $f(\Theta)$ is set to a constant value (i.e. uniform) for all sets of parameter values, where the range for each parameter is a prescribed interval set by the investigator. By setting this prior distribution to be uniform, $P(\Theta \mid X)$ is proportional to the likelihood of the parameters, given the data. Thus for example, if equation (1) can be evaluated, then the mode of the posterior distribution provides a maximum likelihood estimate of Θ.

The use of the integral form in (1) simply means that all genealogies are considered. Clearly G is quite complex because of variation in both branch lengths and topology, and equation (1) cannot be solved analytically for all but the smallest of sample sizes. However, the expression can be estimated using a Markov chain simulation in which the transition rates are specified by the Metropolis–Hastings criterion (Metropolis *et al*. 1953; Hastings 1970). Nielsen & Wakeley (2001) developed this approach for the IM model. Their method is suitable for DNA sequence data that fits the infinite sites mutation model, and for which all polymorphism can be fitted to a gene tree without invoking recombination.

Here we describe two important additions to the method: the inclusion of a stepwise mutation model and a procedure for considering loci that include both an infinite sites component and a stepwise component. For the stepwise model we applied the method of Wilson & Balding (1998) in which the genealogy, G, includes not only the topology and the branch lengths, but also the allelic state of all ancestral nodes. For example, a data set consisting of a total of n gene copies can be represented as n integers that are the observed allele sizes (i.e. the numbers of repeat units for the STR). These values then reside at the termini of the branches of each of the possible genealogies for those data. Under the infinite sites model, updating of G during the simulation involves the removal and repositioning of individual branches in a process that over many iterations explores the space of possible genealogies (Nielsen & Wakeley 2001). Under the stepwise mutation model this same procedure is used but in addition at each step of the simulation when G is being updated, the allelic states of all the ancestral states are updated according to a Metropolis–Hastings criterion. These updates require that a proposal distribution be specified for new allelic states. We used a geometric distribution with expectation set to the average of the difference between the current value and that of each of the three connecting nodes, weighted by the branch lengths. The probability of the update, under the geometric distribution is included in the Metropolis–Hastings criterion to ensure that the prior distribution for ancestral allele states is flat over a wide range of allele sizes. For calculation of $P(X \mid G,\Theta)$, expression (3) of Wilson and Balding was used (Wilson & Balding 1998).

To include the stepwise (SW) and infinite sites (IS) mutation models simultaneously for the analysis of haplotypes that include both a DNA sequence and an allele length for the linked STR locus, we modify equation (1) as follows. Let these two linked data sets be denoted by X_{SW} and X_{IS}. For any given genealogy, G we can assess both $P(X_{SW} \mid G,\Theta)$ and $P(X_{IS} \mid G,\Theta)$. Thus the posterior density is given by:

$$P(\Theta \mid X_{SW}, X_{IS}) = cf(\Theta)\int_G P(X_{SW} \mid G,\Theta)P(X_{IS} \mid G,\Theta)P(G \mid \Theta)dG \qquad (2)$$

Under the infinite sites model, the mutation rate u cannot be estimated independently of the model parameters (Fig. 1). With STR and sequence data we need to allow for two mutation rates. Let u_{SW} be a scalar such that $u \times u_{SW}$ is the mutation rate for the STR and let u_{IS} be a scalar such that $u \times u_{IS}$ is the mutation rate for the sequence portion

of the data. Finally the relationship between u_{IS} and u_{SW} is constrained so that $u_{IS} = 1/u_{SW}$ and a change in one scalar during the simulation is accompanied by a reciprocal change in the other scalar. In the course of the simulation, updates to u_{IS} are drawn from a log-scale uniform distribution centred on 1. In this way u is the geometric mean of the two mutation rates, and we can obtain locus-specific values for any parameter estimate by multiplying the parameter estimate by the estimate of the locus-specific mutation scalar (e.g. $\theta_{1,SW} = \theta_1 \times u_{SW}$). Because the prior distribution for θ_i is uniform and the priors for u_{SW} and u_{IS} are log uniformly distributed, the prior distribution for $\theta_{i,SW}$ and $\theta_{i,IS}$ are also log uniformly distributed.

The computer program of Nielsen & Wakeley (1901) estimates the posterior densities of IM model parameters assuming no recombination and the infinite sites mutation model (Kimura 1969). The stepwise and the joint infinite-sites/stepwise models were added to this program. The program was checked by comparing results with expectations based on coalescent simulations and by comparison of results under a simplified model of a single constant size population model (obtained by setting the time parameter, t, to zero) to results obtained using other computer programs that implement the stepwise mutation model (Wilson & Balding 1998; Beaumont 1999).

The program was run on the U66814 HapSTR data as well as on the DNA sequence data alone (i.e. not including the STR). *Tropheops gracilior* was designated as species 1 and *T. tropheops* as species 2, although this is arbitrary and has no effect on the analyses. Initial runs were performed using very wide prior distributions for model parameters so as to ensure that the complete posterior distribution could be obtained if possible. Because the method is Bayesian, by using initial very wide prior distributions, we were in effect recognizing that we had no prior beliefs on parameter values. These runs of 10^7 steps were also performed repeatedly to ensure that the Markov chain was mixing adequately for all parameters. As a general matter it can be difficult to ensure that Markov chain simulations such as these are adequately exploring the parameter space (Gelman & Rubin 1992; Geyer 1992). In practice our criteria for adequate mixing has been when all model parameters have update rates that are greater than 2%, and when multiple independent runs of 10^7 steps yield similar distributions. Following these preliminary runs, primary runs were carried out with a burn-in period of 10^5 steps following by a run of over 10^8 steps in the Markov chain simulation. These final longer runs required several hours on a fast personal computer.

Results

The U66814 HapSTR has a full length of about 900 bp and includes a dinucleotide STR (Booton *et al*. 1996). Eleven

a

```
AAGAAAAGCA TTACTGAGTC CTCACTTATG ATAATTGTAA GTGCACTGCT ATGTTGCTGGC
AAAACACAAG GCTGGACAGC AAACCAAAAG CAGAGCAGAT TATCATCAAA ATGATCTGTTT
ACTGATATGT GATCTCACCC TATTCCCTCA ACCGTAGCCT CCTTTAAGTG GATGTAGGAAG
CAGGGAAAAT GCCCTGAAGA AAAAGAAAAA TTTCATAGTG ATCAGTAAAT TACACACACAC
ATACACACCA AGACAATGAT GATAATATAG CACTATTGCC CATACCTTT GTGATTTCCTG
CGTAGAGCAT AGCCCCTATA CCAGCCTAAG AAGACATAAA TTACCCAACA AAAACGTTACA
CTGAGGGATA GAGCGAGATA GGAGTGATGT CAGGAGGACA ATGTAATAGA CTCATTTACAT
TACTATGCAT AACACACTGA GTGCCAGTTA GTTCAACATC TCATTTAAGT TCCTGTGTCAG
CTTGCTTTAG ATTAGGGGAA AGAGGGATGT GTCGGGGCCT GCTGTGTGCT TATGCTTGCTC
ACCTTCAAAT TTCTCAAGTA TGTGCACCGT GTCCCCCACT TGTAGACACA GCTCCTGTTCG
CCACTGGCAT CATAATTGTA GA
```

b

Fig. 2 (a) The DNA sequence of 632 bases of the U66814 flanking region. The positions of the two polymorphic sites are denoted in boxes. (b) The U66814 HapSTR is shown with the locations of PCR primers. The full length of the region is approximately 900 bp.

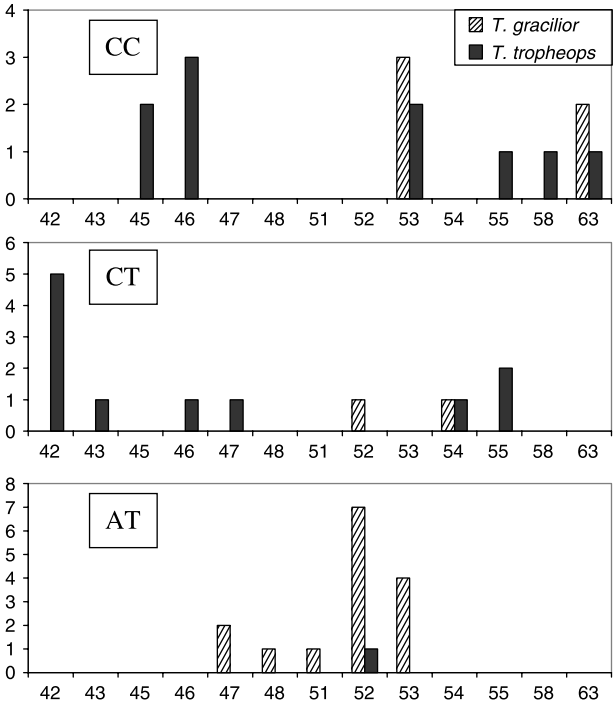

Fig. 3 Histograms of STR allele counts for each of the three DNA sequence haplotypes, designated CC, CT and AT based on the base values at polymorphic positions (Fig. 2).

individuals were genotyped and sequenced from each species, for a total sample size of 44 chromosomes. The sequenced region of 632 bp revealed two polymorphic sites in the sample (Fig. 2) and three sequence haplotypes. Figure 3 shows allele counts for each species and each sequence haplotype. All three sequence haplotypes and most STR alleles were shared by the two species, and the species share four joint haplotypes (CC-53, CC-63, CT-54 and AT-52).

Figure 4 shows the marginal posterior distributions for the mutation rate scalars u_{IS} and u_{SW}. The peaks are at

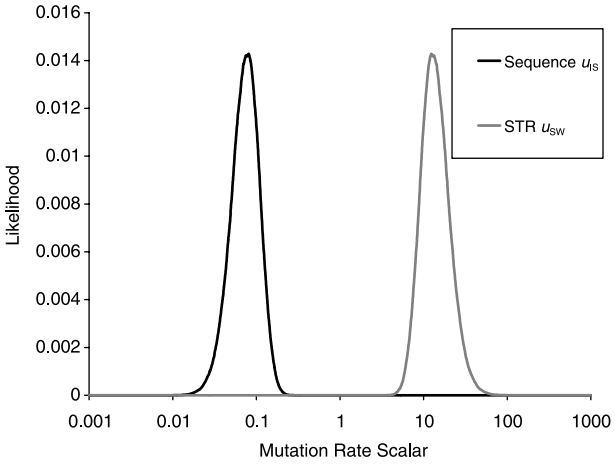

Fig. 4 The marginal likelihood surfaces for the mutation rate scalars for the U66814 flanking sequence (u_{IS}) and the STR (u_{SW}), obtained by integrating the full likelihood surface over all of the other model parameters. During the running of the Markov chain, the two mutation rate scalars are not independent, but rather are constrained to be reciprocals of each other (see text).

approximately 0.076 and 13.1, respectively, for a mutation rate ratio of 172. Given that the mutation rate for the sequence portion was determined over 632 bases, the estimate of the rate of stepwise mutations is about 100 000 times the point mutation rate per base pair in the flanking sequence. For example, if the point mutation rate per base pair is 10^{-9} per generation, then the stepwise rate for the STR would be about 0.0001 per generation.

Figure 5 shows the likelihood distributions for θ_1 and θ_2 (i.e. $4Nu$ for *Tropheops gracilior* and *T. tropheops*, respectively) for the joint U66814 HapSTR, as well as for the sequence data alone and the STR data alone. In addition Fig. 5 shows the HapSTR distribution of θ_1 and θ_2, in each case multiplied by the estimate of u_{IS} and u_{SW} ($\theta_1 \times \hat{u}_{IS}$ and $\theta_1 \times \hat{u}_{SW}$). It is interesting that the curves for each type of data for both θ_1 and θ_2, considered in isolation, are closer together than are the curves for $\theta \times \hat{u}_{IS}$ and $\theta \times \hat{u}_{SW}$ based on the joint HapSTR data. In effect, the joint HapSTR data suggest a lower value of the effective population size for each species than does the sequence data alone, and a higher value than does the STR data alone. The curves for θ_2 are positioned to the right of those for θ_1 suggesting that *T. tropheops* has had an effective population size two to three times that of *T. gracilior*.

The curves for θ_A (for the ancestral population) vary widely between that for the complete HapSTR and those for each of the individual subsets, the sequence data and the STR data, considered in isolation (Fig. 6). The joint HapSTR revealed a single peak at a θ_A with a value of 11.8, a value that is over 20 times that at which the HapSTR peak for θ_1 is located, suggesting a much larger size for the ancestral population prior to the onset of divergence. For the sequence data alone, the curve is almost completely flat, showing almost zero resolution. In contrast, the curve for the STR data alone has two peaks, one at the lower end of the distribution and a second higher peak at a value of 155.

In the case of the scaled time parameter, t, both the sequence data alone and the STR data alone generated a flat marginal likelihood curve (Fig. 7). In contrast the HapSTR data generated a sharply resolved peak at $t = 0.088$. Because $t = tu$ and $\theta_1 = 4N_1u$ (Fig. 1), we can obtain an estimate of time in units of $2N_1$ generations by dividing the value of t at the peak height by one half of the estimate of θ_1 (from Fig. 4, the peak of the curve for θ for the HapSTR data is at 0.53). This leads to an estimate of time of $0.088/(0.53/2) = 0.332$, in units of $2N$ generations. For related species of mbuna, the effective population size for territorial males has been estimated to range from 2500 to 18 000 (Van Oppen *et al*. 1997). If we double these values to account for both sexes and if the average generation time is 3 years, then the corresponding time estimates range from 5000 to 36 000 years.

The curves for the migration parameters are shown in Fig. 8. Migration, in both directions is suggested by the HapSTR and sequence data sets, however, once again the HapSTR curves are more sharply peaked. These curves are for the scaled migration rates m_1/u and m_2/u. To convert to a more readily interpreted effective population size scale, we can use the relationships $2N_1m_1 = 4N_1u \times 1/2 \times m_1/u$ and $2N_2m_2 = 4N_2u \times 1/2 \times m_2/u$. Substituting the estimates of these parameters from the peak heights, we estimate that the population rate of gene flow into *T. gracilior*, since divergence began has been $0.53 \times 1.2/2 = 0.32$ gene copies per generation and that the rate of gene flow into *T. tropheophs* has been $1.59 \times 0.79 = 0.63$ gene copies per generation. These values suggest low but nontrivial gene flow. They are less than, but near to the critical value of $2Nm = 1$ below which appreciably population divergence can accrue (Wright 1931).

In contrast to the HapSTR and sequence-based curves for the migration parameters, the estimated distributions for just the STR data had sharp peaks. The peak for m_1 is at the lower end of the distribution, while that for m_2 is at a value of 0.135. The shapes of these curves suggest that we can have higher confidence in the migration parameter estimates for the STR data than for the HapSTR data and for the sequence data alone. However it is important to keep in mind that these STR-based curves are integrated over a surface that revealed ambiguous or flat curves for both θ_A and t. It is as if the STR data suggest low gene flow, but tell us very little about other key details of the IM model.

Discussion

Phylogeographic and population genetic questions regarding the divergence of populations can be difficult,

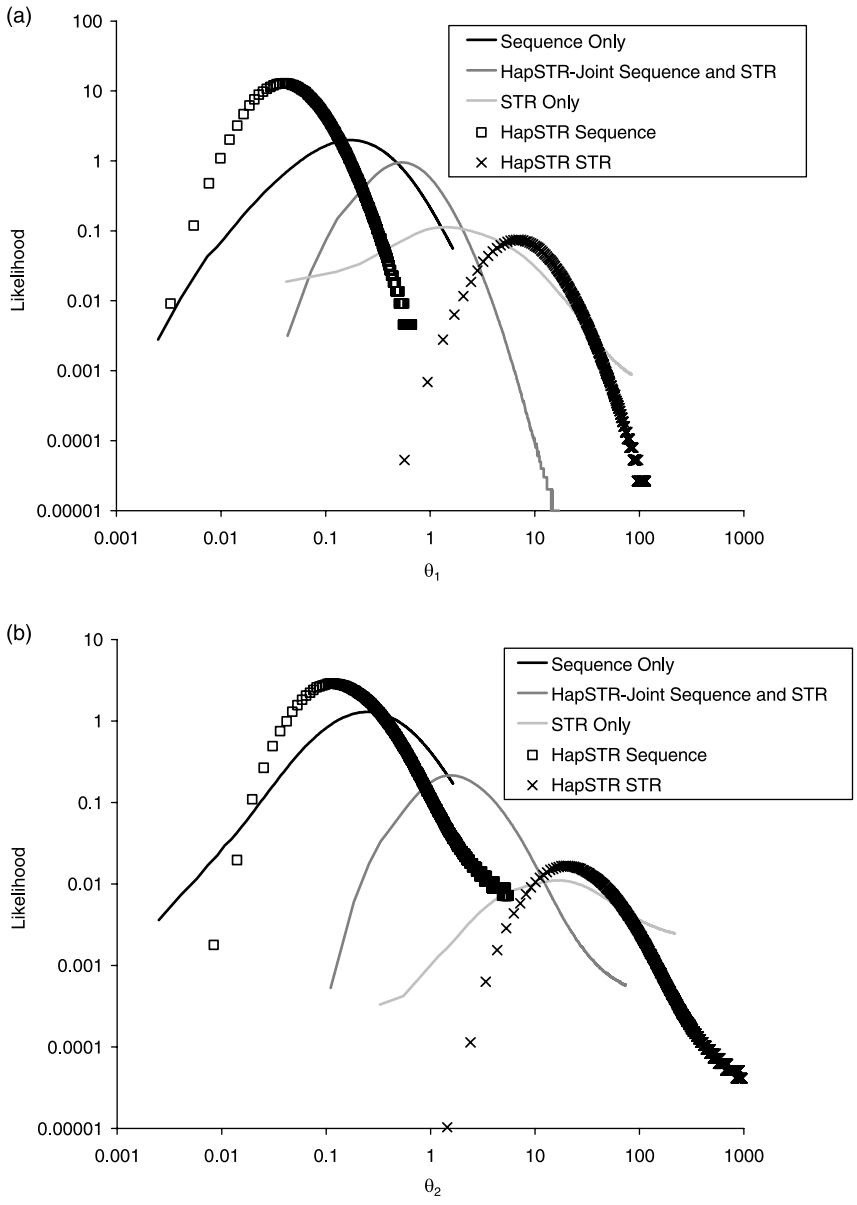

Fig. 5 (a) The marginal likelihood surface for θ_1 obtained by integrating the full likelihood surface over all of the other model parameters. Surfaces are shown for model fitting with the sequence data alone (i.e. without including the STR data), for the STR data alone, for the complete HapSTR data, as well as for the product of θ_1, based on the HapSTR data, and the estimates of the mutation rate scalars for each portion of the HapSTR, u_{IS} (HapSTR sequence) and u_{SW} (HapSTR STR). Surfaces are shown on log–log plots because of the wide range of values for different data sets. (b) The same for θ_2.

particularly if divergence has been recent and the presence of gene flow is uncertain. Here we present a series of laboratory and analytical protocols that should improve our ability to study gene flow when populations have recently separated. The protocols centre on the use of two different mutation models considered jointly for data from loci that include both an STR and its immediately flanking DNA sequence.

Given DNA samples from multiple individuals from two populations or closely related species, the first step in obtaining HapSTR data is to determine STR genotypes using PCR primers immediately flanking the STR. Second, PCR is performed using more widely spaced primers that span the STR and the flanking DNA. Finally, these larger fragments are separated on a denaturing acrylamide gel and individual bands are excised and sequenced. Because most STRs will be heterozygous in most individuals, this protocol leads directly to complete haplotypes (including STR and flanking sequence) from diploid loci without cloning individual gene copies. For locus U66814, we found that some individuals were homozygous for the STR allele or were heterozygous for alleles that differed by only two bases in length. In these cases, individual PCR products cannot be separated. In such cases three strategies can be anticipated. First, it can be useful to excise and sequence the products jointly as there may be a good chance that they are homozygous for the flanking DNA, given their similarity at the STR. This proved to be the case

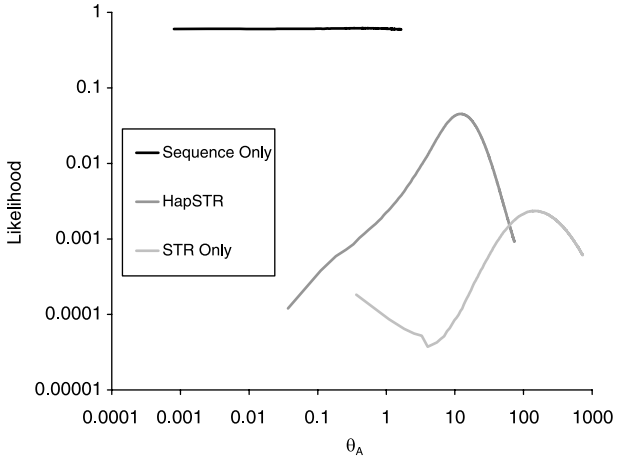

Fig. 6 The marginal likelihood surface for the ancestral population parameter, θ_A, obtained by integrating the full likelihood surface over all of the other model parameters.

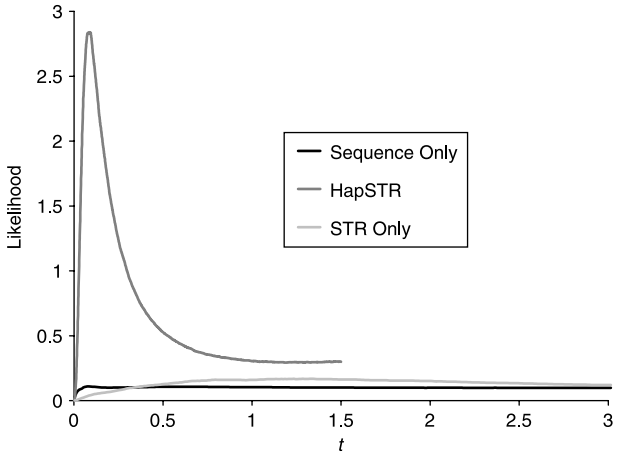

Fig. 7 The marginal likelihood surface for t obtained by integrating the full likelihood surface over all of the other model parameters. The full curve for the sequence data alone extended flatly over a much longer distance than shown which is why the curve has a low overall height.

for the subset of our samples that were homozygous or had similar alleles at the STR. Second, if the number of individuals in which PCR products cannot be separated because of homozygosity or very similar length makes up only a small proportion of the total number of individuals, then it may be acceptable to exclude them. The reason is that because their STR lengths are already known it should still be possible to sample haplotype data from other heterozygous individuals in proportion to the frequencies of observed STR alleles. Third, when an individual is homozygous for the STR and all gene copies are required, one can clone and sequence each flanking sequence (Blankenship *et al*. 2002).

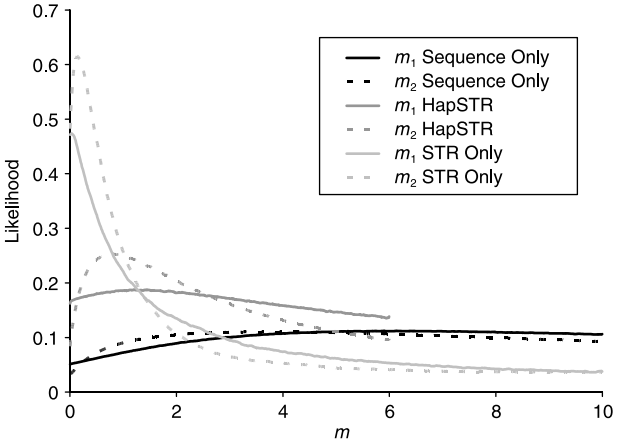

Fig. 8 The marginal likelihood surface for the migration parameters m_1 and m_2 obtained by integrating the full likelihood surface over all of the other model parameters.

We have also developed a procedure for fitting the IM model to HapSTR data. The method builds directly on the method originally described for the infinite sites mutation model (Nielsen & Wakeley 2001). However, it is important to recognize that a full likelihood assessment of the IM model, given real or simulated data, presents a number of challenges. Long runs of the computer program are necessary to generate Markov chains that have thoroughly sampled the parameter space, and it can be difficult to know whether the space has been adequately sampled (Gelman & Rubin 1992; Geyer 1992; Gilks *et al*. 1996). It is also difficult to take advantage of the full multidimensional posterior distribution, as it cannot be visualized or easily examined for the locations of peaks. Also, we cannot generally have a great deal of confidence that analyses will yield findings that are similar to what is expected, based either on intuition or on simpler preliminary analyses. The complexity that is allowed by the high dimensionality of the process will often mean that there are unanticipated models (i.e. sets of parameter values) that have high likelihoods.

One of the key assumptions is that mutations at the STR conform to the stepwise mutation model. However, studies of the actual mutational spectrum have revealed multistep mutations, overall length constraints, and non-stepwise mutations that create alleles that are not multiples of the basic motif (Estoup *et al*. 1995; Lehman *et al*. 1996; Di Rienzo *et al*. 1998; Colson & Goldstein 1999; Renwick *et al*. 2001). In the future, it may be possible to incorporate parameters of the mutational model to accommodate multistep mutations and length constraints (Nielsen 1997), however, the method is probably not suitable for STRs that show considerable numbers of nonstepwise variant alleles.

The present application describes the use of just a single HapSTR locus. However, individual loci will vary widely in

their histories, and in particular if gene flow has occurred at low levels, then some loci may show evidence of gene flow while others may not. Thus most applications should make use of multiple independently segregating loci. Recently the Nielsen and Wakeley method has been extended to multiple loci (Hey & Nielsen, submitted for publication). The computer program implementing these methods is available upon request to J.H. or R.N.

The cichlids of Lake Malawi

Lake Malawi is home to hundreds of described species in the family Cichlidae, most of which arose within the two million years that the lake has been in existence. There are active debates on the actual mode of speciation, and the kinds of divergent natural selection that may have caused speciation (Kornfield & Smith, 2000; Danley & Kocher 2001). Recent work on the mbuna cichlids of Lake Malawi has revealed that these species share much of their genetic variation (Kornfield 1978; McKaye *et al*. 1982; McKaye & Gray 1984; Moran & Kornfield 1993; Kornfield & Parker 1997), as expected if they have recently diverged but also expected if they have been exchanging genes. Our limited analysis on a small data set from two species is consistent both with recent divergence and with gene flow between species since the time that divergence began. If this finding holds up with more data and analysis, it will have implications for our appreciation of how there came to be so many cichlid species in Lake Malawi, and will shift discussions away from models that require allopatry (or microallopatry) to discussions that include gene flow, including sympatric, parapatric and ecological speciation models (Schluter 1998; Danley & Kocher 2001).

Acknowledgements

We are grateful to Matt Arnegard for helpful discussions and for help with collecting samples. Richard Zatha, Daniel Phiri and David Mwafulirwe also provided valuable field assistance. Thanks to Roy Bhima and the Lake Malawi National Park for a collection permit. We also thank Aggrey Ambali of the Molecular Biology and Ecology Research Unit of the University of Malawi for helping to facilitate the collecting trip and for the use of the MBERU field station. This research was supported by a grant from the National Science Foundation to J.H.

References

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology*, **11**, 155–165.

Beaumont MA (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Blankenship SM, May B, Hedgecock D (2002) Evolution of a perfect simple sequence repeat locus in the context of its flanking sequence. *Molecular Biology and Evolution*, **19**, 1943–1951.

Booton GC, Kaufman L, Chandler M, Fuerst PA (1996) Use of DNA microsatellite loci to identify populations and species of Lake Victoria haplochromine cichlids. In: *Symposium Proceedings, International Congress on the Biology of Fishes* (eds Donaldson EM, MacKinlay DD), pp. 105–113. American Fisheries Society, Physiology Section, Vancouver, CA.

Brohede J, Ellegren H (1999) Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proceedings of the Royal Society, London, Series B*, **266**, 825–833.

Broughton RE, Harrison RG (2003) Nuclear gene genealogies reveal historical, demographic and selective factors associated with speciation in field crickets. *Genetics*, **163**, 1389–1401.

Bulgin NL, Gibbs HL, Vickery P, Baker AJ (2003) Ancestral polymorphisms in genetic markers obscure detection of evolutionarily distinct populations in the endangered Florida grasshopper sparrow (*Ammodramus savannarum floridanus*). *Molecular Ecology*, **12**, 831–844.

Colson I, Goldstein DB (1999) Evidence for complex mutations at microsatellite loci in *Drosophila*. *Genetics*, **152**, 617–627.

Danley PD, Kocher TD (2001) Speciation in rapidly diverging systems: lessons from Lake Malawi. *Molecular Ecology*, **10**, 1075–1086.

Di Rienzo A, Donnelly P, Toomajian C *et al*. (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics*, **148**, 1269–1284.

Dieckmann U, Doebeli M (1999) On the origin of species by sympatric speciation. *Nature*, **400**, 354–357.

Dobzhansky T (1936) Studies of hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics*, **21**, 113–135.

Estoup A, Cornuet JM, Rousset F, Guyomard R (1999) Juxtaposed microsatellite systems as diagnostic markers for admixture: theoretical aspects. *Molecular Biology and Evolution*, **16**, 898–908.

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, **11**, 1591–1604.

Estoup A, Largiader CR, Cornuet JM *et al*. (2000) Juxtaposed microsatellite systems as diagnostic markers for admixture: an empirical evaluation with brown trout (*Salmo trutta*) as model organism. *Molecular Ecology*, **9**, 1873–1886.

Estoup A, Tailliez C, Cornuet JM, Solignac M (1995) Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). *Molecular Biology and Evolution*, **12**, 1074–1084.

Felsenstein J, Kuhner MK, Yamato J, Beerli P (1999) Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In: *Statistics in Genetics and Molecular Biology* (ed. Seillier-Moiseiwitsch F). Institute of Mathematical Statistics and American Mathematical Society, Hayward, CA.

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Geyer CJ (1992) Practical Markov Chain Monte Carlo. *Statistical Science*, **7**, 473–511.

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Introducing Markov chain Monte Carlo. In: *Markov Chain Monte Carlo in Practice* (eds Gilks WR, Richardson S, Spiegelhalter DJ), pp. 1–20. Chapman & Hall, Boca Raton, FL.

Griffiths RC, Tavare S (1994) Simulating probability distributions in the coalescent. *Theoretical Population Biology*, **46**, 131–159.

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hey J (1994) Bridging phylogenetics and population genetics with gene tree models. In: *Molecular Approaches to Ecology and Evolution* (eds Schierwater B, Streit B, Wagner G, DeSalle R), pp. 435–449. Birkhäuser-Verlag, Basel.

Higashi M, Takimoto G, Yamamura N (1999) Sympatric speciation by sexual selection. *Nature*, **402**, 523–526.

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, **11**, 424–429.

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**, 893–903.

Kliman RM, Andolfatto P, Coyne JA *et al.* (2000) The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics*, **156**, 1913–1931.

Kornfield I (1978) Evidence for rapid speciation in African cichlid fishes. *Experientia*, **34**, 335–336.

Kornfield I, Parker A (1997) Molecular systematics of a rapidly evolving species flock: the mbuna of Lake Malawi and the search for phylogenetic signal. In: *Molecular Systematics of Fishes* (eds Kocher TD, Stepien CA), pp. 25–37. Academic Press, New York.

Kornfield I, Smith PF (2000) African cichlid fishes: model systems for evolutionary biology. *Annual Review of Ecology and Systematics*, **31**, 163–196.

Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.

Lehman T, Hawley WA, Collins FH (1996) An evaluation of evolutionary constraints on microsatellite loci using null alleles. *Genetics*, **144**, 1155–1163.

Machado C, Kliman RM, Markert JM, Hey J (2002) Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and its close relatives. *Molecular Biology and Evolution*, **19**, 472–488.

Machado CA, Hey J (2003) The causes of phylogenetic conflict in a classic *Drosophila* species group. *Proceedings of the Royal Society, London, Series B*, **270**, 1193–1202.

McKaye ER, Kocher T, Reinthal P, Kornfield I (1982) A sympatric species complex of *Petrotilapia trewavas* from Lake Malawi analysed by enzyme electrophoresis (Pisces, Cichlidae). *Zoological Journal of the Linnean Society*, **76**, 91–96.

McKaye KR, Gray WN (1984) Extrinsic barriers to gene flow in rock-dwelling cichlids of Lake Malawi: macrohabitat heterogeneity and reef colonization. In: *Evolution of Fish Species Flocks* (eds Echelle AA, Kornfield I). University of Maine Press, Orono.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Moran P, Kornfield I (1993) Retention of an ancestral polymorphism in the Mbuna species flock (Teleostei: Cichlidae) of Lake Malawi. *Molecular Biology and Evolution*, **10**, 1015–1029.

Mountain JL, Knight A, Jobin M *et al.* (2002) SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Research*, **12**, 1766–1772.

Muller HJ (1940) Bearings of the *Drosophila* work on systematics. In: *The New Systematics* (ed. Huxley J), pp. 185–268. Clarendon Press, Oxford.

Navarro A, Barton NH (2003) Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, **57**, 447–459.

Nielsen R (1997) A likelihood approach to population samples of microsatellite alleles. *Genetics*, **146**, 711–716.

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

Noor MA, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences U S A*, **98**, 12084–12088.

Ochman H, Gerber AS, Hartl DL (1988) Genetic applications of an inverse polymerase chain reaction. *Genetics*, **120**, 621–623.

Orr HA (1996) Dobzhansky, Bateson, and the Genetics of Speciation. *Genetics*, **144**, 1331–1335.

Renwick A, Davison L, Spratt H, King JP, Kimmel M (2001) DNA dinucleotide evolution in humans: fitting theory to facts. *Genetics*, **159**, 737–747.

Ribbink AJ, Marsh AC, Marsh BA, Sharp BJ (1983) A preliminary survey of the cichlid fishes of the rocky habitats in Lake Malawi. *South African Journal of Zoology*, **18**, 149–310.

Rieseberg LH, Whitton J, Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**, 713–727.

Saint-Laurent R, Legault M, Bernatchez L (2003) Divergent selection maintains adaptive differentiation despite high gene flow between sympatric rainbow smelt ecotypes (*Osmerus mordax* Mitchill). *Molecular Ecology*, **12**, 315–330.

Schluter D (1998) Ecological cuases of speciation. In: *Endless Forms: Species and Speciation* (eds Howard DJ, Berlocher SH), pp. 114–129. Oxford University Press, New York.

Seutin G, White BN, Boag PT (1991) Preservation of avian blood and tissue samples for DNA analyses. *Canadian Journal of Zoology*, **69**, 82–91.

Shaw KL (2002) Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proceedings of the National Academy of Sciences USA*, **99**, 16122–16127.

Surridge AK, Bell DJ, Ibrahim KM, Hewitt GM (1999) Population structure and genetic variation of European wild rabbits (*Oryctolagus cuniculus*) in East Anglia. *Heredity*, **82**, 479–487.

Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, **110**, 325–344.

Takahata N, Slatkin M (1990) Genealogy of neutral genes in two partially isolated populations. *Theoretical Population Biology*, **38**, 331–350.

Van Oppen M, Rico C, Turner G, Hewitt G (2000) Extensive Homoplasy, nonstepwise mutations, and shared ancestral polymorphism at a complex microsatellite locus in Lake Malawi cichlids. *Molecular Biology and Evolution*, **17**, 489–498.

Van Oppen MJH, Turner GF, Rico C *et al.* (1997) Unusually fine-scale genetic structuring found in rapidly speciating Malawi cichlid fishes. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **264**, 1803–1812.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

Wakeley J, Hey J (1998) Testing speciation models with DNA sequence data. In: *Molecular Approaches to Ecology and Evolution* (eds DeSalle R, Schierwater B), pp. 157–175. Birkhäuser Verlag, Basel.

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.

Won Y, Young CR, Lutz RA, Vrijenhoek RC (2003) Dispersal barriers and isolation among deep-sea mussel populations (Mytilidae: Bathymodiolus) from eastern Pacific hydrothermal vents. *Molecular Ecology*, **12**, 169–184.

Wright S (1931) Evolution in mendelian populations. *Genetics*, **16**, 97–159.

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563–584.

Zhivotovsky LA (2001) Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. *Molecular and Biological Evolution*, **18**, 700–709.

Jody Hey conducts empirical and theoretical genetic research on diverse problems in speciation and evolutionary genetics. Yong-Jin Won is a postdoctoral fellow working on the population genetics of diverging populations. Arjun Sivasundar is a graduate student with interests in the genetic structure of natural populations. Rasmus Nielsen is a faculty member at Cornell University who works on developing and applying analytical and statistical approaches to population genetic problems.

Jeffrey Markert is a postdoctoral fellow with long term interests in speciation, particularly the diversification of Cichlids of the African great lakes.

