Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015

doi: 10.1111/2041-210X.12356

A discrete-beta model for testing gene flow after speciation

Junfeng Liu[1,2], De-Xing Zhang[1,3]* and Ziheng Yang[4,5]*

[1]*State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;* [2]*University of the Chinese Academy of Sciences, Beijing 100049, China;* [3]*National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China;* [4]*Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China; and* [5]*Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK*

Summary

1. Due to the polymorphism in the ancestral species and the stochastic fluctuation of the coalescent process, the divergence time between sequences from two closely related species varies throughout the genome. This variation is described by an exponential distribution. Gene flow between the species causes additional variation in the sequence divergence time, beyond the expected variation from the ancestral coalescent process. This prediction can be used to test for gene flow between species using genomic sequence data.

2. A previous implementation of the model used an empirical beta distribution to describe the variation in species divergence time across the genome, with numerical integration used to calculate the 3D integrals involved in the likelihood function. However, with a huge number of loci, the numerical integration is not accurate enough, so the false-positive rate of the likelihood ratio test may exceed the significance level.

3. In this study, we replace the beta model with a discretized version, so that the likelihood calculation involves 2D integrals. We implement the new discrete-beta model in the program 3s and also extend our previous implementation to accommodate loci with different data configurations. We use computer simulation to examine the false-positive rate, the power and the robustness of the likelihood ratio test.

4. The results show that the test had low false-positive rates and had even higher power in detecting gene flow than a likelihood ratio test based on a proper implementation of the isolation-with-migration model. It is, however, not robust to high levels of recombination within locus. Application of the test to two empirical data sets (from the hominoid genomes and from the Drosophila fruit flies) suggests the presence of gene flow for both data sets.

Key-words: coalescent, discrete-beta, gene flow, maximumlikelihood, speciation

Introduction

The mode and mechanism of speciation are fascinating questions in evolutionary biology since the time of Charles Darwin. Different theories of speciation make different predictions concerning the role of gene flow during species formation. Allopatric speciation considers complete isolation of the populations and lack of gene flow between them as prerequisites to formation of new species. Parapatric and sympatric speciation, in contrast, allows gene flow during speciation. Allopatric speciation has historically been considered the paramount mode of speciation (Futuyma & Mayer 1980), although theoretical modelling and empirical evidence appear to increasingly support the notion of speciation despite gene flow (see, e.g. Mallet *et al*. 2009; Smadja & Butlin 2011 for recent reviews). As gene flow causes different parts of the genome to diverge at different times between the two species, the mode of speciation may have left permanent marks in the genomes of modern species. Thus, the different theories of speciation may be tested using genomic sequence data.

Even under a model of complete isolation without gene flow (that is, even if gene flow ceases as soon as one species diverges into two), the time of sequence divergence between the two species will vary among different regions of the genome because of polymorphism and the coalescent process in the ancestral species. Indeed, the coalescent time in the ancestral population should follow an exponential distribution, with the mean to be $2N$ generations, where N is the effective population size of the ancestral species (Langley & Fitch 1974). However, gene flow after speciation introduces additional variation in sequence divergence among genomic regions, beyond the expected variation from the exponential distribution. Furthermore, since the information about the sequence divergence time comes from the accumulated mutations, another complicating factor is the possible variation in mutation rate along the genome, which can also cause variable sequence divergences among genomic regions. Thus, a test of gene flow should accommo-

*Correspondence authors. E-mails: dxzhang@ioz.ac.cn (De-Xing Zhang), z.yang@ucl.ac.uk (Ziheng Yang)

date the natural stochastic fluctuation of the coalescent process, as well as the accumulation of random mutations given the sequence divergence time, and should be relatively robust to possible mutation rate variation along the genome.

A likelihood ratio test (LRT) was implemented by Yang (2010), which uses a beta distribution to account for the variation in *species divergence time* between the two species caused by gene flow. Given the species divergence time, the coalescent waiting time has an exponential distribution. This is an extension of the test of Osada & Wu (2005; see also Wu & Ting 2004) for two species, which allows for two possible species divergence times among loci. Yang's (2010) implementation compares two species (1 and 2), with an out-group species 3 used to provide additional information and to improve the robustness of the test. With a pair of closely related species, the variation in sequence divergence among loci may be seriously confounded with the mutation rate variation (Yang 2010). Use of a third out-group species alleviates this problem to some extent. In the implementation of Yang (2010), the probability of data at each locus is calculated by summing over the gene tree topologies and integrating over the two coalescent times and over the beta distribution, so that the likelihood calculation involves 3D integrals, achieved by numerical integration (Gaussian quadrature). With K points used in each dimension in the quadrature, the computation is proportional to K^3. The simulation of Yang (2010) generating data of up to 1000 loci suggests that the test has good false-positive rate and good power. Nevertheless, it was found later that the numerical integration does not achieve sufficient accuracy: when a large number of loci (≥10 000 loci, say) are used, the false-positive rate may exceed the nominal 5%. Another limitation of the implementation of Yang (2010) is that it allows for only one type of locus: there must be three sequences, with one sequence from each species, at each locus.

Here in this paper, we replace the (continuous) beta model with a discretized beta distribution. The model is then called the discrete-beta model. The integration over the two coalescent times in the gene tree is then two-dimensional and the computation is proportional to K^2. Importantly, this formulation allows accurate calculation of the likelihood function, ensuring that the LRT has the false-positive rate under control. We also extend the implementation to other data configurations involving three sequences per locus. We conduct computer simulation to examine the false-positive rate and power of the LRT, as well as its robustness to violations of certain model assumptions. We then apply the test to two empirical data sets: one of genomic data from the human, chimpanzee and gorilla (Burgess & Yang 2008), and another of *Drosophila* fruit flies (Wang & Hey 2010).

Theory and methods

MODEL OF NO GENE FLOW (M0)

We first describe the model of speciation without gene flow to introduce the notation (Yang 2002, 2010). This is the complete isolation model, referred to as model M0 (no gene flow). Let the species tree be ((1, 2), 3) (Fig. 1a). The two ancestral species are referred to as 4 and 5. The focus of the analysis is on comparison of species 1 and 2, with species 3 used as an out-group. There are at most seven parameters in model M0, including two species divergence times: τ_0 and τ_1, and five population size parameters: $\theta_1 = 4N_1\mu$, $\theta_2 = 4N_2\mu$ and $\theta_3 = 4N_3\mu$ for the three modern species, and $\theta_4 = 4N_4\mu$ and $\theta_5 = 4N_5\mu$ for the two ancestral species. Here, μ is the mutation rate per site per generation, and the Ns are the effective population sizes. Parameter θ measures the population size or the genetic variation maintained in the population: for instance, $\theta = 0{\cdot}001$ means that two sequences sampled from the population have on average one difference per kilobases. Similarly, the species divergence times (τs) are scaled by the mutation rate and are measured by the expected number of mutations per site. Let $\Theta_0 = \{\tau_0, \tau_1, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$. We note that θ for a modern species is an estimable parameter only if there are at least two sequences from that species at some loci. We assume only one sequence from species 3 at each locus and θ_3 is not a parameter in the model. In this study, the term 'loci' refers to independent or loosely linked short segments of the genome, such that recombination within each segment is rare and negligible while different loci are so far apart that they are nearly free-recombining (Lohse & Barton 2011).

The data consist of DNA sequence alignments from multiple neutral loci. At each locus, there are two or three sequences from the modern species 1, 2 and 3. We use the index of the species from which the sequences are sampled to specify the data configuration at each locus.

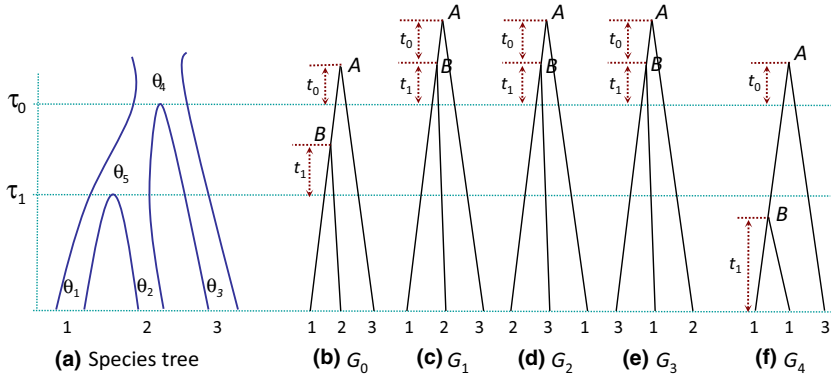

Fig. 1. The species tree ((1, 2), 3) for three species, showing the parameters in model M0: θ_4, θ_5, τ_0 and τ_1. The five possible gene trees for any locus are shown in (b–f). For loci of configuration 123, gene trees G_0–G_3 are possible. For loci of configuration 113, gene tree G_4 is also possible. The branch lengths in gene tree G_0 are defined as b_0 for AB and b_1 for branch $B1$, while those in other gene trees are defined similarly.

Thus, configuration '123' means three sequences at the locus, with one sequence from each species, configuration '113' means two sequences from species 1 and one sequence from species 3, and so on. In this paper, we consider only data configurations 123, 113 and 223, assuming three sequences at each locus. The program 3s has been extended to accommodate arbitrary data configurations with 2 or 3 sequences at each locus. The details of the implementation will be described elsewhere (Dalquen, Zhu and Yang, pers. comm.).

We use the JC69 mutation model (Jukes & Cantor 1969). As the sequences are expected to be very similar in such analysis and the role of the mutation model is to correct for multiple hits, JC69 appears to be adequate. With three sequences at any locus i, the data can be summarized as the counts of sites with five distinct site patterns: *xxx*, *xxy*, *yxx*, *xyx* and *xyz*, where *x*, *y* and *z* are any distinct nucleotides. Let these be $D_i = \{n_{i0}, n_{i1}, n_{i2}, n_{i3}, n_{i4}\}$. Also let $D = \{D_i\}$ be the data at all L loci.

Consider a locus with data configuration 123. The relationships among the three sequences are described by one of four possible gene trees: G_0, G_1, G_2 and G_3 (Fig. 1). Given the gene tree topology G_1: ((1, 2), 3); G_2: ((2, 3), 1); or G_3: ((3, 1), 2), and the branch lengths b_0 and b_1 (Fig. 1), the probability of observing the site pattern counts is given by the multinomial probability

$$\begin{aligned} P(D_i|G_1, b_0, b_1) &= p_0^{n_{i0}} p_1^{n_{i1}} p_2^{n_{i2}+n_{i3}} p_4^{n_{i4}}, \\ P(D_i|G_2, b_0, b_1) &= p_0^{n_{i0}} p_1^{n_{i2}} p_2^{n_{i3}+n_{i1}} p_4^{n_{i4}}, \\ P(D_i|G_3, b_0, b_1) &= p_0^{n_{i0}} p_1^{n_{i3}} p_2^{n_{i1}+n_{i2}} p_4^{n_{i4}}, \end{aligned} \qquad \text{eqn 1}$$

where the probabilities for the site patterns are (Saitou 1988; Yang 1994)

$$\begin{aligned} p_0(b_0, b_1) &= \text{prob}(xxx); = (1 + 3e^{-8b_1/3} + 6e^{-8(b_0+b_1)/3} \\ &\quad + 6e^{-(8b_0+12b_1)/3})/16, \\ p_1(b_0, b_1) &= \text{prob}(xxy) = (3 + 9e^{-8b_1/3} - 6e^{-8(b_0+b_1)/3} \\ &\quad - 6e^{-(8b_0+12b_1)/3})/16, \\ p_2(b_0, b_1) &= \text{prob}(yxx) = (3 - 3e^{-8b_1/3} + 6e^{-8(b_0+b_1)/3} \\ &\quad - 6e^{-(8b_0+12b_1)/3})/16, \\ p_3(b_0, b_1) &= \text{prob}(xyx) = p_2(b_0, b_1), \\ p_4(b_0, b_1) &= \text{prob}(xyz) = (6 - 6e^{-8b_1/3} - 12e^{-8(b_0+b_1)/3} \\ &\quad + 12e^{-(8b_0+12b_1)/3})/16. \end{aligned} \qquad \text{eqn 2}$$

Note that gene tree G_0 has the same topology as G_1, so that $P(D_i|G_0, b_0, b_1) = P(D_i|G_1, b_0, b_1)$.

The probability of data at the locus $f(D_i|\Theta_0)$ is a sum over the gene tree topologies and an integration over the two coalescent times (Yang 2002: eqn 8, 2010: eqn 3).

Note that the first term in the sum is for gene tree G_0 while the second term is for G_1–G_3, with $e^{-2(\tau_0-\tau_1)/\theta_5}$ to be the probability that sequences 1 and 2 do not coalesce in ancestral population 5.

For data configuration 113 (with two sequences sampled from species 1 and a third sequence from species 3), it is possible for the two sequences from species 1 to coalesce in species 1, in the interval $(0, \tau_1)$, before reaching the common ancestor 5. Thus, the probability of data at the locus involves one extra term corresponding to gene tree G_4 of Fig. 1. This is given in eqn 4 below, where the first two terms in the sum correspond to gene trees G_4 and G_0 of Fig. 1, respectively, while the third term is for G_1–G_3. Note that $e^{-2\tau_1/\theta_1}$ and $e^{-2(\tau_0-\tau_1)/\theta_5}$ are the probabilities that the two sequences from species 1 do not coalesce in population 1 and in ancestral population 5, respectively. Again G_4 has the same gene tree topology as G_1 so that $P(D_i|G_4, b_0, b_1)$ is given by equation (1) for G_1.

Loci with data configuration 223 can be dealt with similarly to loci of data configuration 113.

We assume that the different loci are evolving independently, so that the log likelihood is a sum over the L loci

$$\ell(\Theta_0; D) = \sum_{i=1}^{L} \log f(D_i|\Theta_0). \qquad \text{eqn 5}$$

THE DISCRETE-BETA MODEL OF GENE FLOW (M1)

In the beta model, we use a beta distribution of τ_1 to describe the extra variation in the sequence divergence time among loci due to gene flow between species 1 and 2. This model does not describe the biological process of isolation with migration exactly. It is an empirical model, aimed at capturing the main features of the process, that is the extra variation in sequence distance among genomic regions caused by gene flow. In the case that gene flow is decreasing after species separation and migration rate is variable over time, this model may be a good approximation to the biological reality. Let $x = \tau_1/\tau_0$ for any locus have the beta distribution $x \sim \text{beta}(p, q)$, with density

$$f(x|p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}, \quad 0 < x < 1. \qquad \text{eqn 6}$$

This has the mean $E(x) = p/(p+q)$ and the variance $V(x) = pq/[(p+q)^2(p+q+1)]$. For a locus with the beta variable x, the two species divergence times will be τ_0 and $\tau_1 = x\tau_0$. Model M1 involves one more parameter than M0: instead of τ_0 and τ_1, we now have τ_0, p

$$\begin{aligned} f(D_i|\Theta_0) = &\int_0^\infty \int_0^{2(\tau_0-\tau_1)/\theta_5} P(D_i|G_0, \tau_0 - \tau_1 - \tfrac{1}{2}\theta_5 t_1 + \tfrac{1}{2}\theta_4 t_0, \tau_1 + \tfrac{1}{2}\theta_5 t_1) \times e^{-t_1} e^{-t_0} dt_1 dt_0 \\ &+ e^{-2(\tau_0-\tau_1)/\theta_5} \int_0^\infty \int_0^\infty \left[\sum_{k=1}^{3} P(D_i|G_k, \tfrac{1}{2}\theta_4 t_0, \tau_0 + \tfrac{1}{2}\theta_4 t_1)\right] \times e^{-3t_1} e^{-t_0} dt_1 dt_0 \end{aligned} \qquad \text{eqn 3}$$

$$\begin{aligned} f(D_i|\Theta_0) = &\int_0^\infty \int_0^{2\tau_1/\theta_1} P(D_i|G_4, \tau_0 - \tfrac{1}{2}\theta_1 t_1 + \tfrac{1}{2}\theta_4 t_0, \tfrac{1}{2}\theta_1 t_1) \times e^{-t_1} e^{-t_0} dt_1 dt_0 \\ &+ e^{-2\tau_1/\theta_1} \int_0^\infty \int_0^{2(\tau_0-\tau_1)/\theta_5} P(D_i|G_0, \tau_0 - \tau_1 - \tfrac{1}{2}\theta_5 t_1 + \tfrac{1}{2}\theta_4 t_0, \tau_1 + \tfrac{1}{2}\theta_5 t_1) \times e^{-t_1} e^{-t_0} dt_1 dt_0 \\ &+ e^{-2\tau_1/\theta_1} e^{-2(\tau_0-\tau_1)/\theta_5} \int_0^\infty \int_0^\infty \left[\sum_{k=1}^{3} P(D_i|G_k, \tfrac{1}{2}\theta_4 t_0, \tau_0 + \tfrac{1}{2}\theta_4 t_1)\right] \times e^{-3t_1} e^{-t_0} dt_1 dt_0. \end{aligned} \qquad \text{eqn 4}$$

and q. We parametrize M1 to have $\Theta_1 = \{\tau_0, \bar{\tau}_1, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, q\}$, with $p = q\frac{\bar{\tau}_1}{\tau_0 - \bar{\tau}_1}$. With this formulation, parameter q is inversely related to the variance in τ_1, and the null model M0 of constant τ_1 is represented by $q = \infty$ in M1.

The probability of data at the locus is given as

$$f(D_i|\Theta_1) = \int_0^1 f(D_i|\tau_0, x\tau_0, \theta_1, \theta_2, \theta_4, \theta_5) f(x|p, q)\,\mathrm{d}x, \qquad \text{eqn 7}$$

where $f(D_i|\tau_0, x\tau_0, \theta_1, \theta_2, \theta_4, \theta_5)$ is $f(D_i|\Theta_0)$ of equation (3) for a locus of configuration 123 or equation (4) for a locus of configuration 113. The likelihood function is given again by equation (5). Note that equation (7) involves 3D integrals, over the two coalescent times (t_1 and t_0) and over x. Yang (2010) used Gaussian quadrature to calculate those 3D integrals.

Here we replace the (continuous) beta model with a discretized version. We cut the distribution into B bins, each with probability $1/B$. Let the cutting points be $y_0 = 0, y_1, y_2, \ldots, y_B = 1$, where y_j is the $100j/B$-th percentile of the distribution. The percentiles are calculated by a linear search using standard algorithms for the calculation of the cumulative distribution function (CDF) $F(x, p, q)$ for beta (p, q) (Steinbrecher & Shaw 2008). We use the mean as the representative for all x values in each bin. In other words,

$$\begin{aligned}\bar{x}_j &= B \times \int_{y_{j-1}}^{y_j} x f(x|p, q)\,\mathrm{d}x \\ &= B \times \frac{p}{p+q} \times \frac{\Gamma(p+1+q)}{\Gamma(p+1)\Gamma(q)} \int_{y_{j-1}}^{y_j} x^{p+1-1}(1-x)^{q-1}\,\mathrm{d}x \qquad \text{eqn 8} \\ &= B \times \frac{p}{p+q}\left[F(y_j; p+1, q) - F(y_{j-1}; p+1, q)\right]\end{aligned}$$

for $j = 1, 2, \ldots, B$.

The species divergence time in the ith bin is thus $\tau_1 = \bar{x}_j\tau_0$. The integral of equation (7) is then replaced by an average over the B bins, so that the probability of data at the locus becomes

$$f(D_i|\Theta_1) \approx \frac{1}{B}\sum_{j=1}^{B} f(D_i|\tau_0, \bar{x}_j\tau_0, \theta_1, \theta_2, \theta_4, \theta_5). \qquad \text{eqn 9}$$

Thus, we have to calculate 2D integrals at every locus (equations 3 and 4), so that the computation involved in Gaussian quadrature is proportional to BK^2. We note that the number of bins B is not a parameter in the model, and use a fixed value $B = 5$ in our implementation.

Figure 2 illustrates the discretization of beta (1·408, 0·394), using $B = 5$ bins. Those parameter values are maximum likelihood estimates obtained from the Drosophila data, analysed later. In the analysis of the simulated and real data sets, we used $B = 5$ bins in the discrete-beta model, and $K = 16$ or 32 in the numerical integration by Gaussian quadrature.

THE LIKELIHOOD RATIO TEST

As model M1 (discrete-beta) reduces to model M0 (no gene flow) when $q = \infty$, the two models are nested and can be compared using a LRT. The test statistic will be $2\Delta\ell = (\ell_1 - \ell_0)$, where ℓ_0 and ℓ_1 are the log likelihood values under models M0 and M1, respectively. However, as the value ∞ for q is at the boundary of the parameter space under M1, we cannot use χ_1^2 (the chi-square distribution with one degree of freedom) to conduct the test. Instead, the null distribution is a 50 : 50 mixture of the point mass at 0 and χ_1^2 (Self & Liang 1987). The critical values are 2·71 at 5% and 5·41 at 1%, as opposed to 3·84 for 5% and 6·63 for 1% according to χ_1^2 (Yang 2010). We refer to this LRT comparing models M0 (no gene flow) and M1 (discrete-beta) as test 1.

For comparison, we also applied the LRT of Zhu & Yang (2012), which compares models M0 (no gene flow) and M2 (SIM3s). M2 is a proper implementation of the isolation-with-migration (IM) model. The test is referred to as test 2, and we use the χ_2^2 distribution to conduct the test (Zhu & Yang 2012).

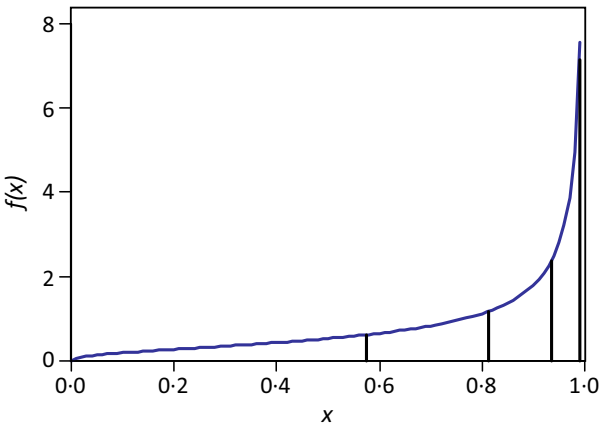

Fig. 2. Discretization of the beta distribution beta (1·408, 0·394), with estimates obtained from the Drosophila data. The four vertical lines are at x = 0·5750, 0·8112, 0·9351 and 0·9890, and are the 20th, 40th, 60th and 80th percentiles of the distribution and cut the density into five categories, each of proportion 1/5. The mean in the five categories is 0·3575, 0·7054, 0·8803, 0·9667 and 0·9969.

COMPUTER SIMULATION TO EXAMINE THE FALSE-POSITIVE RATE AND POWER OF THE TEST

We conducted computer simulations to examine the statistical properties of the LRT based on the discrete-beta model. To examine the false-positive rate, we simulated sequence data at multiple loci under model M0 (no gene flow), using the program MCcoal in the bpp package (Rannala & Yang 2003; Yang & Rannala 2010). Two sets of parameter values were used, roughly based on estimates from the hominoids (Burgess & Yang 2008) and the mangroves (Zhou *et al.* 2007). They are as follows: $\theta = 0{\cdot}005$ for all populations, $\tau_0 = 0{\cdot}006$ and $\tau_1 = 0{\cdot}004$ (hominoids); and $\theta = 0{\cdot}01$ for all populations, $\tau_0 = 0{\cdot}02$ and $\tau_1 = 0{\cdot}01$ (mangroves).

To examine the power of the test, we simulated the gene trees and sequence alignments under the IM model for three species, with $M_{12} = M_{21} = M$, using the program MCcoal. This is the symmetrical IM model for three species (M2: SIM3s) of Zhu & Yang (2012), and differs from the discrete-beta model that test 1 is based on. In MCcoal, the (scaled) migration rate from species i to j is defined as $M_{ij} = N_j m_{ij}$, where m_{ij} is the proportion of individuals in population j that are immigrants from population i. Thus, M_{ij} is the expected number of migrants from population i to population j per generation. In this study, we considered only the symmetrical case with $M_{12} = M_{21} = M$ and assume no migration to or from species 3. We used three values: $M = 0{\cdot}1$, 1 and 10. Data of multiple loci were simulated by first generating the gene trees and branch lengths and then generating sequence alignments given the gene trees. See Zhang *et al.* (2011) for a detailed description of the simulation strategy.

Our simulation considered different proportions of loci of three data configurations: 123, 113 and 223. For example, a total number of 1000 loci with the ratios 4 : 3 : 3 means that each data set consists of 400 loci of configuration 123, 300 loci of 113 and 300 loci of 223. There were 500 bp at every locus. The number of replicates was 1000. The data were analysed using the program 3s, which implements all three models: M0 (no gene flow), M1 (discrete-beta) and M2 (SIM3s). The JC69

mutation model (Jukes & Cantor 1969) was used in both the simulation and analysis of the data.

We conducted simulation to examine the robustness of the test to the assumption of no recombination within each locus. We considered two scenarios, related to the hominoid and Drosophila data sets analysed later. In the hominoid case, we used $\theta = 0{\cdot}005$ for all populations, $\tau_0 = 0{\cdot}006$ and $\tau_1 = 0{\cdot}004$. Rough estimates of the recombination rate for the human are about $r = 0{\cdot}37$ cM Mb^{-1} (with the 95% CI to be 0·27–0·47) (Arnheim, Calabrese & Tiemann-Boege 2007). With an effective population size of $N = 10\,000$, this gives $\rho = 4Nr = 0{\cdot}148$ per kilo base pairs or 0·074 for a 500 bp locus. We used a range of values for ρ. In the *Drosophila* case, we used $\theta = 0{\cdot}01$ for modern species and $\theta = 0{\cdot}02$ for ancestral species, and $\tau_0 = 0{\cdot}025$ and $\tau_1 = 0{\cdot}013$. Estimates of recombination rates in *Drosophila* are around $r = 2$ cM Mb^{-1} (Fiston-Lavier *et al*. 2010; Langley *et al*. 2012). With a population size of about 10^6 (Li, Satta & Takahata 1999), this gives $\rho = 4Nr = 0{\cdot}08$ per base pair, or 32 per locus (with 400 sites per locus). We also used ρ values that are half or twice as large.

We also examined the power of the test in a few complex scenarios of gene flow, mimicking gradual build-up of reproductive isolation after speciation or secondary contact. The details will be described later when we present the results. The program MS (Hudson 2002) was used to simulate gene trees under the model of recombination and under complex migration scenarios, and the program SEQ-GEN (Rambaut & Grassly 1997) was used to generate the sequence data under the JC69 model. The data were then analysed using 3s under the JC69 model.

Results

THE FALSE-POSITIVE RATE AND POWER OF THE LRT

The false-positive rate of the test is shown in Table 1 for different data sizes (the number of loci L), different parameter values and different data configurations. The false-positive rate is <5% in most cases. In a few other cases, it is slightly higher than 5%, which may be attributed to random sampling errors. Overall the type-I error rate of the LRT based on the discrete-beta model is acceptable. Note also that the percentage of data sets in which the test statistic $2\Delta\ell = 0$ becomes close to 50% when the number of loci L is large, consistent with the theoretical expectation (Self & Liang 1987).

The power of the test is shown in Table 2. The power is influenced by the parameter values, the data configurations of the loci and the number of loci. For the hominoid set of parameter values, the test has virtually no power when $L = 100$, but relatively high power is achieved when $L = 1000$. For the mangrove set of parameter values, the power is close to 100% even with $L = 100$ when all loci are of configuration 123 and is even higher with $L \geq 1000$. Note that compared with the hominoid set of parameter values, the mangrove set has much larger θs and τs, so that the sequences are more divergent and the data are more informative. Furthermore, loci of configuration 123 are more informative, with the test achieving higher power, than loci of configurations 113 or 223. It is unclear what proportions of the loci of different configurations are optimal and achieve the highest power for the test. Such issues of experimental design are beyond the scope of our small simulation study.

Note that the null hypothesis M0 (no gene flow) is true when the migration rate $M = 0$. However, the power of the test to reject M0 does not always increase with the increase of M. Biologically, when $M = \infty$ (or even $M = 10$), populations 1 and 2 will be in effect one single species, so that the simulation model is one of two species, with different population size parameters for the time period $(0, \tau_1)$ and (τ_1, τ_0), rather than a model of three species with frequent migration. The null model M0 does not match this true model, which explains the relative high power of the test at $M = 10$ (Table 2). Test of migration should be interpreted with caution when the estimates of migration rate are very high.

ROBUSTNESS OF THE TEST TO RECOMBINATION

We examined the robustness of the LRT based on the discrete-beta model to recombination. Note that our model assumes complete linkage within each locus, so that the model assumption is violated when recombination occurs between sites in the locus. The results of this simulation are shown in Table 3.

For the hominoid set of simulation, the false-positive rate is acceptable when $\rho = 4Nr \leq 0{\cdot}148 \times 10^{-3}$ per base pair, which is based on the estimate for humans (Arnheim, Calabrese & Tiemann-Boege 2007). When the recombination rate is much higher, say at $\rho \geq 2 \times 10^{-3}$, excessive false positives are generated. By comparison with previous results of Zhu & Yang

Table 1. False-positive rate of the LRT_1 in simulations

Data configuration	$L = 100$	$L = 1000$	$L = 5000$	$L = 10\,000$
Hominoid set				
Configurations 1 : 0 : 0	0·009 (0·608)	0·039 (0·510)	0·052 (0·506)	0·049 (0·512)
Configurations 4 : 3 : 3	0·031 (0·521)	0·051 (0·527)	0·050 (0·477)	0·057 (0·509)
Configurations 2 : 3 : 5	0·050 (0·535)	0·061 (0·520)	0·051 (0·511)	0·049 (0·509)
Mangrove set				
Configurations 1 : 0 : 0	0·050 (0·541)	0·039 (0·528)	0·048 (0·543)	0·050 (0·538)
Configurations 4 : 3 : 3	0·043 (0·563)	0·044 (0·528)	0·049 (0·518)	0·043 (0·542)
Configurations 2 : 3 : 5	0·048 (0·574)	0·039 (0·543)	0·042 (0·539)	0·056 (0·525)

The test is conducted at the 5% level (with critical value $2\Delta\ell = 2{\cdot}71$). In parentheses is the percentage of replicates in which the test statistic is $2\Delta\ell = 0$. Data of L loci are simulated under M0. We use three experimental designs, in which the numbers of loci of configurations 123, 113 and 223 are in proportions 1 : 0 : 0, 4 : 3 : 3 and 2 : 3 : 5, respectively.

Table 2. Power of the LRT_1 in simulations

Simulation model	$L = 100$	$L = 1000$	$L = 5000$	$L = 10\,000$
Hominoid set				
$M = 0{\cdot}1$ (1 : 0 : 0)	0·194	0·991	1·000	1·000
$M = 0{\cdot}1$ (4 : 3 : 3)	0·159	0·768	1·000	1·000
$M = 0{\cdot}1$ (2 : 3 : 5)	0·105	0·519	0·988	1·000
$M = 1$ (1 : 0 : 0)	0·183	0·906	1·000	1·000
$M = 1$ (4 : 3 : 3)	0·141	0·462	0·884	0·977
$M = 1$ (2 : 3 : 5)	0·105	0·345	0·681	0·841
$M = 10$ (1 : 0 : 0)	0·109	0·841	0·999	1·000
$M = 10$ (4 : 3 : 3)	0·128	0·798	1·000	1·000
$M = 10$ (2 : 3 : 5)	0·116	0·758	0·999	1·000
Mangrove set				
$M = 0{\cdot}1$ (1 : 0 : 0)	0·957	1·000	1·000	1·000
$M = 0{\cdot}1$ (4 : 3 : 3)	0·618	1·000	1·000	1·000
$M = 0{\cdot}1$ (2 : 3 : 5)	0·344	0·998	1·000	1·000
$M = 1$ (1 : 0 : 0)	0·703	1·000	1·000	1·000
$M = 1$ (4 : 3 : 3)	0·402	0·965	1·000	1·000
$M = 1$ (2 : 3 : 5)	0·316	0·849	0·997	1·000
$M = 10$ (1 : 0 : 0)	0·469	0·988	1·000	1·000
$M = 10$ (4 : 3 : 3)	0·444	0·960	0·999	1·000
$M = 10$ (2 : 3 : 5)	0·414	0·951	0·998	1·000

The test is conducted at the 5% level (with critical value $2\Delta\ell = 2{\cdot}71$). Data of L loci are simulated under M2 (SIM3s), with the migration rate M to be the expected number of migrants per generation. See notes for Table 2.

Table 3. False-positive rate of the LRT_1 in presence of recombination

Recombination rate (ρ)	$L = 100$	$L = 1000$	$L = 5000$	$L = 10\,000$
(a) Hominoid set, $N = 500$ bp				
$0{\cdot}037 \times 10^{-3}$	0·007	0·035	0·051	0·059
$0{\cdot}074 \times 10^{-3}$	0·011	0·044	0·052	0·049
$0{\cdot}148 \times 10^{-3}$	**0·008**	**0·038**	**0·073**	**0·077**
$0{\cdot}3 \times 10^{-3}$	0·006	0·052	0·069	0·086
2×10^{-3}	0·018	0·083	0·193	0·306
8×10^{-3}	0·041	0·195	0·537	0·813
(b) Drosophila set, $N = 400$ bp				
0·04	0·154	0·803	0·998	1·000
0·08	**0·034**	**0·399**	**0·900**	**0·981**
0·16	0·010	0·078	0·408	0·634

Recombination rate $\rho = 4Nr$ is per base pair, with the value $0{\cdot}148 \times 10^{-3}$ to be the estimate for the human, and 0·08 to be the estimate for *Drosophila melanogaster*. The values of parameters θs and τs for the hominoid set are $\theta = 0{\cdot}005$ for all populations, $\tau_0 = 0{\cdot}006$ and $\tau_1 = 0{\cdot}004$. For the *Drosophila* set, θs are 0·01 for modern species and 0·02 for ancestors, and $\tau_0 = 0{\cdot}025$ and $\tau_1 = 0{\cdot}013$. The sequence length is 500 bp for the hominoid set and is 400 bp for the flies set. All loci have the configuration 123. Results highlighted in bold are for recombination rates estimated for the human and Drosophila, respectively.

Table 4. Power of LRT_1 and LRT_2 when the rate of gene flow varies over time

Model	$L = 100$	$L = 1000$	$L = 5000$	$L = 10\,000$
LRT1 (M0–M1)				
Constant migration	0·189	0·930	1·000	1·000
Gradual isolation	0·146	0·797	1·000	1·000
Secondary contact	0·112	0·450	0·968	0·997
LRT2 (M0–M2)				
Constant migration	0·060	0·687	0·938	0·968
Gradual isolation	0·050	0·506	1·000	1·000
Secondary contact	0·026	0·171	0·677	0·944

The test is conducted at the 5% level, with critical value $2\Delta\ell = 2{\cdot}71$ for test 1 and 5·99 for test 2. With constant migration, the scaled migration rate M is 1. With gradual isolation, M is 0·1, 1 and 5 during the three time periods $\left(0, \frac{1}{3}\tau_1\right)$, $\left(\frac{1}{3}\tau_1, \frac{2}{3}\tau_1\right)$ and $\left(\frac{2}{3}\tau_1, \tau_1\right)$. With secondary contact, M is 0, 10 and 0, during the three periods. Parameters θs and τs are from the Hominoid set. L is the number of loci, with the sequence length to be 500. All loci have the configuration 123.

(2012) (Table 2), we note that test 1, based on the discrete-beta model, is more sensitive to recombination than test 2. For example, when $L = 1000$ and $\rho = 8 \times 10^{-3}$ per base pair, the false positive is 0·195 for test 1 but only 0·047 for test 2.

For the Drosophila set of simulation, the sequences are much more divergent and the recombination rate is much higher. The false-positive rate is unacceptably high, especially in large data sets with $L > 1000$ loci. Interestingly, the false-positive rate decreases rather than increases with the increase of the recombination rate (Table 3). We suggestion caution should be exercised when applying the test to species with very high recombination rate.

THE POWER OF THE TEST IN COMPLEX SCENARIOS OF GENE FLOW

We considered a few complex migration scenarios in which the migration rate M changes over time to examine the power of the LRT to detect gene flow. The results are summarized in Tables 4 and 5. One scenario, called gradual isolation (Table 4), mimics the gradual build-up of reproductive isolation after speciation with the migration rate decreasing over time. We break the time interval $(0, \tau_1)$ after speciation into three equal-width periods: $\left(\frac{2}{3}\tau_1, \tau_1\right)$, $\left(\frac{1}{3}\tau_1, \frac{2}{3}\tau_1\right)$ and $\left(0, \frac{1}{3}\tau_1\right)$ and let M be 5, 1 and 0·1 during the three periods. Another scenario, called secondary contact (Table 4), has M to be 0, 10 and 0 during the three time periods. Other parameter values (θs and τs) are from the hominoid set, and all loci have the configuration 123. We also include the case of constant migration (with $M = 1$) for comparison. In addition to test 1 based on M1 (discrete-beta), we also apply the LRT of Zhu & Yang (2012) (test 2) to those data, which compares models M0 (no gene flow) and M2 (SIM3s). Similar to results of Table 2 (hominoid set), $L = 1000$ loci appear to be necessary for the tests to have any substantial power. It is interesting to note that test 1 has more power than test 2 in all three scenarios: constant migration, gradual isolation and secondary contact.

Next, we consider a scenario in which there is considerable gene flow following speciation but then gene flow stops completely a certain time later. The time interval $(0, \tau_1)$ is broken into two segments, with $M = 0$ in the interval $(0, c\tau_1)$ and $M = 5$ in the interval $(c\tau_1, \tau_1)$. We consider five different values for the fraction c: 0, 1/5, 2/5, 3/5 and 4/5, referred to as cases A, B, C, D and E, respectively. Note that case A ($c = 0$) is the case of constant migration, with ongoing gene flow for the whole period, as in the simulation of Table 2, while in case E ($c = 4/5$), gene flow ceased a long time ago. The results are summarized in Table 5. The longer the time since gene flow had ceased (e.g. case E), the more difficult it is to detect gene flow. Indeed, for case E, the power is only 0·112 for the hominoid set even with 10 000 loci. A short period of gene flow followed by complete isolation is very difficult to distinguish from complete isolation with more recent species divergence. Again we note that test 1 is more powerful than test 2 in all those cases (A–E).

ANALYSIS OF GENOMIC DATA FROM DROSOPHILA AND THE HOMINOIDS

We apply the discrete-beta model to the genomic sequences of the human (H), chimpanzee (C) and gorilla (G) (Burgess & Yang 2008). The data set included 9861 neutral autosomal loci and 510 X-linked loci. The average sequence length per locus is about 508 bp. Each locus has three sequences, with one sequence from each of the three species, so that all loci have the configuration 123. These data were analysed by Yang (2010) using the (continuous) beta model. Here we re-analyse the same data using the new discrete-beta model for comparison. We also apply test 2, which compares models M0 (no gene flow) and M2 (SIM3s) (Zhu & Yang 2012).

The results are summarized in Table 6. For the autosomal data set, the results under M1 (discrete-beta) when $K = 16$ and 32 points are used in the Gaussian quadrature are very similar. The test statistic for test 1 (which compares M0 with M1) is

Table 5. Power of LRT_1 and LRT_2 when gene flow ceased a certain time after speciation

	LRT_1			LRT_2		
	$L = 100$	$L = 1000$	$L = 10\,000$	$L = 100$	$L = 1000$	$L = 10\,000$
Hominoid set						
A (0%)	0·120	0·852	1·000	0·041	0·401	0·683
B (20%)	0·141	0·628	1·000	0·029	0·290	0·989
C (40%)	0·080	0·360	0·995	0·020	0·083	0·791
D (60%)	0·052	0·160	0·617	0·008	0·035	0·203
E (80%)	0·018	0·061	0·112	0·009	0·004	0·025
Mangrove set						
A (0%)	0·530	0·994	1·000	0·233	0·717	0·907
B (20%)	0·509	0·987	1·000	0·190	0·706	1·000
C (40%)	0·384	0·983	1·000	0·082	0·415	0·999
D (60%)	0·158	0·736	1·000	0·031	0·129	0·879
E (80%)	0·074	0·142	0·626	0·009	0·016	0·064

The test is conducted at the 5% level, with critical value $2\Delta\ell = 2{\cdot}71$ for test 1 and 5·99 for test 2. All loci have the configuration 123. The migration rate is $M = 0$ for $0 \le t \le c\tau_1$ and $M = 5$ for $c\tau_1 \le t \le \tau_1$, with the fraction $c = 0\%$, 20%, 40%, 60% and 80% for cases A, B, C, D and E, respectively. Note that case A corresponds to constant migration, with $M = 5$.

Table 6. Maximum likelihood estimates of parameters and LRT statistic for human, chimpanzee and gorilla (9861 loci of configuration 123)

Model	θ_4	θ_5	$\theta_H = \theta_C$	τ_0	τ_1	q	M	ℓ	$2\Delta\ell$
(a) 9861 nuclear autosomal loci									
$K = 16$									
M0	0·00358	0·00431		0·00661	0·00432			−649 936·62	
M1	0·00368	0·00044		0·00656	0·00564	0·491		−649 923·16	26·92
M2	0·00362	0·00369	0·0260	0·00659	0·00459		0·125	−649 930·90	11·44
$K = 32$									
M0	0·00359	0·00430		0·00660	0·00432			−649 936·86	
M1	0·00371	0·00026		0·00655	0·00571	0·394		−649 923·27	26·98
M2	0·00363	0·00368	0·0260	0·00659	0·00460		0·125	−649 932·12	9·48
(b) 510 X loci									
$K = 16$									
M0	0·00305	0·00143		0·00521	0·00362			−25 600·75	
M1	0·00309	0·00061		0·00519	0·00401	2·04		−25 600·41	0·68
M2	0·00308	0·00120	0·0147	0·00520	0·00381		0·117	−25 600·37	0·76
$K = 32$									
M0	0·00304	0·00143		0·00522	0·00362			−25 600·77	
M1	0·00309	0·00061		0·00519	0·00401	2·04		−25 600·44	0·66
M2	0·00307	0·00119	0·0147	0·00520	0·00381		0·117	−25 600·39	0·76

θ_4 is for the HCG ancestor while θ_5 is for the HC ancestor. K is the number of points in the Gaussian quadrature.

Table 7. Maximum likelihood estimates of parameters and LRT statistic for the Drosophila data

Model	θ_4	θ_5	θ_M	θ_S	τ_0	τ_1	q	M	ℓ	$2\Delta\ell$
(a) Assuming different θ_M and θ_S										
$K = 16$										
M0	0·0782	0·0228	0·0060	0·0129	0·0258	0·0109			−6 174 221·95	
M1	0·0809	0·0022	0·0062	0·0147	0·0251	0·0192	0·454		−6 173 207·77	2028·36
$K = 32$										
M0	0·0803	0·0226	0·0060	0·0130	0·0256	0·0109			−6 173 854·16	
M1	0·0841	0·0018	0·0062	0·0147	0·0247	0·0193	0·394		−6 172 742·35	2222·63
(b) Assuming $\theta_M = \theta_S$										
$K = 16$										
M0	0·0806	0·0227	0·0126		0·0256	0·0109			−6 174 155·00	
M1	0·0854	0·0017	0·0143		0·0246	0·0193	0·391		−6 173 055·45	2199·10
M2	0·0814	0·0195	0·0121		0·0254	0·0131		0·0146	−6 173 972·72	364·56
$K = 32$										
M0	0·0803	0·0227	0·0126		0·0256	0·0109			−6 173 925·14	
M1	0·0841	0·0018	0·0143		0·0247	0·0193	0·394		−6 172 827·76	2194·74
M2	0·0810	0·0195	0·0121		0·0254	0·0131		0·0147	−6 173 742·79	364·69
(c) Wang & Hey (2010)										
M0		0·0085	0·0059	0·0135		0·0158				
M2		0·0069	0·0055	0·0135		0·0172		$M_{21} = 0·0067$ $M_{12} = 0·0$		

Wang & Hey (2010) analysed loci with two sequences only and allowed different migration rates: M_{21} for the migration rate from *Drosophila simulans* to *Drosophila melanogaster* and M_{12} in the opposite direction.

$2\Delta\ell = 26·92$, as opposed to $2\Delta\ell = 36·92$ from Yang (2010), using $K = 16$. The difference is consistent with the expectation that the inaccuracies in the numerical integration under the (continuous) beta model of Yang (2010) tend to inflate the test statistic. However, the result of the test stays the same with the use of the discrete-beta model. The null hypothesis is rejected, and the data suggest gene flow between the human and the chimpanzee.

The estimate of q is 0·39–0·49 under the discrete-beta model, compared with 1·189 under the continuous-beta model (Yang 2010). It is obvious that at the same parameter value q, the discrete-beta model involves less variation in τ_1 among loci than the continuous beta. Thus, we expect the estimate of q under the discrete-beta model to be smaller than the estimate under the continuous beta. We note that parameter θ_5 is very poorly estimated under M1. The model (incorrectly) attributes much of the variation in sequence divergence between species 1 and 2 to variation in τ_1, rather than to the natural fluctuation in the coalescent process in the ancestor of species 1 and 2. As a result, the ancestral population size (θ_5) is seriously underesti-

mated. This is in contrast to models M0 and M2, which produced similar and more reliable estimates of θ_5. The strong correlation between estimates of θ_5 and τ_1 means that M1 underestimate θ_5 and overestimate τ_1. Parameters θ_4 and τ_0 are much more stable.

For the data set of X-linked loci, the test statistic for the M0–M1 comparison is $2\Delta\ell = 0{\cdot}68$, as opposed to $\Delta\ell = 0{\cdot}70$ from Yang (2010) using $K = 16$. In neither case is the LRT significant.

We then apply the model to the genomic sequences from three fruit fly species in the *Drosophila melanogaster* subgroup: *D. melanogaster* (M), *D. simulans* (S) and *D. yakuba* (Y). The data are from Wang & Hey (2010) and include 19 889 MSY loci, 10 056 SSY loci and 378 MMY loci. The average locus length is 426 bp. The data are analysed using the program 3s. We also apply LRT 2, which compares M0 (no gene flow) and M2 (SIM3s) (Zhu & Yang 2012). The current implementation of M2 assumes that the two θ parameters for species 1 and 2 are equal: $\theta_1 = \theta_2$, while this restriction is not enforced in M1.

The results are summarized in Table 7. The Drosophila data are far more divergent and informative than the hominoid data. As a result, the numerical integration by quadrature does not work as well as for the hominoid data, and the log likelihood values calculated using $K = 16$ or 32 points show much larger differences. Nevertheless, the parameter estimates and the test results are virtually identical between $K = 16$ and 32. The estimates of θs for modern species are very stable among models and also very similar to those of Wang & Hey (2010): $\theta_M = 0{\cdot}006$ and $\theta_S = 0{\cdot}014$. *Drosophila melanogaster* apparently has a smaller effective population size than *D. simulans* (e.g. Aquadro, Lado & Noon 1988; Langley *et al.* 2012). Under the assumption of $\theta_M = \theta_S$, the estimate is 0·012, very close to θ_S, at 0·014. This is apparently because the data include many more 223 loci than 113 loci; in other words, the population data are mostly from *D. simulans* rather than from *D. melanogaster*. As discussed earlier, model M1 does not provide reliable estimates of ancestral parameters: it underestimates θ_5 and overestimates τ_1. Models M0 and M2 produced quote stable estimates, with $\theta_5 = 0{\cdot}020$ and $\tau_1 = 0{\cdot}013$. In comparison, the estimates from Wang & Hey (2010) are $\theta_5 = 0{\cdot}007$ and $\tau_1 = 0{\cdot}017$. The differences may be partly due to the fact that Wang and Hey analysed doublet data (with two sequences per locus) while we analysed triplet data. Note that parameters θ_5 and τ_1 tend to be strongly negatively correlated (see, e.g. Zhu & Yang 2012; Table 4), so that underestimation of one means overestimation of the other in the same analysis.

Both tests 1 and 2 suggest gene flow between *D. melanogaster* and *D. simulans*, whether $K = 16$ or 32 is used in the calculation. Our simulation results described early suggest that the tests (in particular test 1) may generate excessive false positives at high recombination rates typical of *Drosophila*. Thus, caution should be exercised in the interpretation of the results of Table 7. Nevertheless, we note that previous analyses by Wang & Hey (2010) and Lohse & Barton (2011) both suggested gene flow. The estimate of the migration rate (the expected number of migrants per generation) from model M2 is $M = 0{\cdot}015$, which is very small. Wang & Hey (2010) assumed different migration rates in the two directions, with estimates $M_{12} = 0$ from *D. melanogaster* to *D. simulans* and $M_{21} = 0{\cdot}0067$ in the opposite direction. Those estimates are even smaller than our estimates. It seems that migration between the two species if present occurs only at very low rates.

Discussion

The beta model of variable species divergence times (τ_1) among loci is an empirical model that attempts to describe the main features of the sequence data without modelling the biological mechanisms that have generated those features. In this case, the main feature is the variation in sequence divergence between species across the genome. Neither the continuous-beta nor the discrete-beta is a mechanistic biological model. In this regard, they are both wrong models. The fact that the discrete-beta model has the false-positive rate under control while the continuous-beta model does not is due to the mathematical construction of the model, and has nothing to do with the biological realism of either model. With the same number of points (K) used in each dimension, numerical quadrature is far more reliable for calculating 2D integrals than for 3D integrals. Thus, the discrete-beta model does not suffer from the problem of numerical inaccuracies of the continuous-beta model, which involve calculation of 3D integrals.

In our simulation where the data are generated under M2 (SIM3s), test 1 (which compares model M0 with M1) has more power than test 2 (which compares model M0 with M2). The analysis of real data shows similar patterns with larger values of test statistic for test 1 than for test 2, even though both tests are significant in both data sets. This may appear counter-intuitive. In particular, in the case of comparing two sharp hypotheses, the LRT is known to have the highest power among tests that have the same false-positive rate, a result known as the Neyman–Pearson lemma (Neyman & Pearson 1933). Here, both hypotheses are composite hypotheses with unknown parameters. In this regard, test 1 may be viewed a test based on summary statistics. The higher power of test 1 than test 2 may be explained as follows. First, the additional variation in sequence divergence among loci may be the most important feature in the sequence data that distinguishes models M0 and M2, which is captured in the empirical model M1. Secondly, in test 2, the χ^2_2 distribution (chi-square with 2 degrees of freedom) is not the correct null distribution (Zhu & Yang 2012), because $M = 0$ is at the boundary of the parameter space under model M2, and because parameters θ_1 and θ_2 are not identifiable and become redundant when $M = 0$. As the correct null distribution is unknown, Zhu & Yang (2012) recommended the use of χ^2_2 to conduct the test, but this apparently has made test 2 extremely conservative. At any rate, test 1 based on model M1 (discrete-beta) appears to be a powerful test of gene flow, whether migration is constant as in a simple IM model or it is decreasing over time.

A major difficulty with model M1 is that parameter q does not have an easy biological interpretation, and estimates of certain parameters such as θ_5 and τ_1 are seriously biased and unre-

liable. Both a large θ_5 and a small q will cause the sequence distance between species 1 and 2 to vary among loci, so that the use of the beta model with parameter q to describe the effect of migration makes it impossible to estimate θ_5 reliably under the model. The strong correlation between θ_5 and τ_1 means that τ_1 will be affected as well. We suggest that caution be exercised in interpretation of parameter estimates under M1. In this regard, the mechanistic model M2 (SIM3s) may be very useful for estimating biologically important parameters such as the migration rate M.

Acknowledgements

We thank Tianqi Zhu, Daniel Dalquen and an anonymous reviewer for many constructive comments. This study is supported by a grant from Biotechnological and Biological Sciences Research Council (BBSRC, UK) to Z.Y., and a Ministry of Science and Technology of China Grant (No. 2011CB808001) to D.-X. Z.

Data accessibility

No new data are generated in this study.

References

Aquadro, C.F., Lado, K.M. & Noon, W.A. (1988) The rosy region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. *Genetics*, **119**, 875–888.

Arnheim, N., Calabrese, P. & Tiemann-Boege, I. (2007) Mammalian meiotic recombination hot spots. *Annual Review of Genetics*, **41**, 369–399.

Burgess, R. & Yang, Z. (2008) Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular Biology and Evolution*, **25**, 1979–1994.

Fiston-Lavier, A.S., Singh, N.D., Lipatov, M. & Petrov, D.A. (2010) *Drosophila melanogaster* recombination rate calculator. *Gene*, **463**, 18–20.

Futuyma, D.J. & Mayer, G.C. (1980) Non-allopatric speciation in animals. *Systematic Zoology*, **29**, 254–271.

Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Jukes, T.H. & Cantor, C.R. (1969) Evolution of protein molecules. *Mammalian Protein Metabolism* (ed. H.N. Munro), pp. 21–123. Academic Press, New York.

Langley, C.H. & Fitch, W.M. (1974) An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution*, **3**, 161–177.

Langley, C.H., Stevens, K., Cardeno, C., Lee, Y.C., Schrider, D.R., Pool, J.E. *et al*. (2012) Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics*, **192**, 533–598.

Li, Y.J., Satta, Y. & Takahata, N. (1999) Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. *Genes & Genetic Systems*, **74**, 117–127.

Lohse, K. & Barton, N.H. (2011) A general method for calculating likelihoods under the coalescent process. *Genetics*, **189**, 977–987.

Mallet, J., Meyer, A., Nosil, P. & Feder, J.L. (2009) Space, sympatry and speciation. *Journal of Evolutionary Biology*, **22**, 2332–2341.

Neyman, J. & Pearson, E.S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, **231**, 289–337.

Osada, N. & Wu, C.I. (2005) Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics*, **169**, 259–264.

Rambaut, A. & Grassly, N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, **13**, 235–238.

Rannala, B. & Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**, 1645–1656.

Saitou, N. (1988) Property and efficiency of the maximum likelihood method for molecular phylogeny. *Journal of Molecular Evolution*, **27**, 261–273.

Self, S.G. & Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of American Statistical Association*, **82**, 605–610.

Smadja, C.M. & Butlin, R.K. (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, **20**, 5123–5140.

Steinbrecher, G. & Shaw, W.T. (2008) Quantile mechanics. *European Journal of Applied Mathematics*, **19**, 87–112.

Wang, Y. & Hey, J. (2010) Estimating divergence parameters with small samples from a large number of loci. *Genetics*, **184**, 363–379.

Wu, C.I. & Ting, C.T. (2004) Genes and speciation. *Nature Reviews Genetics*, **5**, 114–122.

Yang, Z. (1994) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology*, **43**, 329–342.

Yang, Z. (2002) Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. *Genetics*, **162**, 1811–1823.

Yang, Z. (2010) A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biology and Evolution*, **2**, 200–211.

Yang, Z. & Rannala, B. (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 9264–9269.

Zhang, C., Zhang, D.-X., Zhu, T. & Yang, Z. (2011) Evaluation of a Bayesian coalescent method of species delimitation. *Systematic Biology*, **60**, 747–761.

Zhou, R., Zeng, K., Wu, W., Chen, X., Yang, Z., Shi, S. & Wu, C.-I. (2007) Population genetics of speciation in nonmodel organisms: I. ancestral polymorphism in mangroves. *Molecular Biology and Evolution*, **24**, 2746–2754.

Zhu, T. & Yang, Z. (2012) Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Molecular Biology and Evolution*, **29**, 3131–3142.

Received 10 December 2014; accepted 16 February 2015
Handling Editor: Louise Johnson