Genet. Res., Camb. (2005), **86**, *pp.* 149–157. *With* 1 *figure.*
doi:10.1017/S0016672305007743 Printed in the United Kingdom

The detection of shared and ancestral polymorphisms

BRIAN CHARLESWORTH*, CAROLINA BARTOLOMÉ[1] AND VÉRONIQUE NOËL
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, UK

(*Received 18 March 2005 and in revised form 1 August 2005*)

Summary

There is increasing evidence that closely related species contain many polymorphisms that were present in their common ancestral species. Use of a more distant relative as an outgroup increases the ability to detect such ancestral polymorphisms. We describe a method for further improving estimates of the fraction of polymorphisms that are ancestral, and illustrate this with reference to data on *Drosophila pseudoobscura* and *D. miranda*. We also derive formulae for the proportion of fixations arising from ancestral polymorphisms and new mutations, respectively. The results should be useful for tests of selection based on the levels of expected and observed ancestral polymorphisms.

1. Introduction

Inferences concerning the evolutionary histories of closely related species are often complicated by the existence of shared polymorphisms, inherited from their common ancestors, as well as possible ongoing gene flow between them (reviewed in Arbogast *et al.*, 2002). Differences between observed and expected levels of shared polymorphisms may also provide evidence for selection (Clark, 1997; Wiuf *et al.*, 2004; Asthana *et al.*, 2005). The application of population genetic models has greatly enhanced our understanding of this problem (Takahata & Nei, 1985; Clark, 1997; Wakeley & Hey, 1997; Wang *et al.*, 1997; Nielsen & Wakeley, 2001; Wiuf *et al.*, 2004).

It has recently been pointed out that the use of an outgroup to distinguish ancestral from derived variants increases the information on shared and ancestral polymorphisms (Ramos-Onsins *et al.*, 2004). However, there is still a problem in that not all ancestral polymorphisms can be identified, so that estimates of the fraction of all polymorphisms that are ancestral are biased downwards (Ramos-Onsins *et al.*, 2004). In this paper, we describe a method for estimating the true fraction of all polymorphisms in one species that are inherited from an ancestral population, following a speciation event that created complete isolation from a sister species. We also present results on the relative numbers of fixations of mutations within one species that are due to mutations that arose after the split between the two species, and those that come from pre-existing polymorphisms.

This investigation was originally motivated by work on the population genetics of *Drosophila miranda* (Yi *et al.*, 2003; Bartolomé *et al.*, 2005), but the methods can be applied to any suitable set of related species. *D. miranda* is a close relative of *D. pseudoobscura*, a classic subject for studies of evolutionary genetics (Powell, 1997). *D. miranda* provides a model system for the evolution of Y chromosomes, because it has a neo-Y chromosome that was recently formed by the fusion of the homologue of chromosome 3 of *D. pseudoobscura* with the true Y chromosome, allowing detailed studies of the effects of the resulting suppression of crossing over on the neo-Y (Steinemann & Steinemann, 1998; Bachtrog, 2003). It has recently been pointed out that the existence of polymorphisms inherited by *D. miranda* and *D. pseudoobscura* from their common ancestor with *D. miranda* may create biases in estimates of the intensity of selection on codon usage and GC content (Bartolomé *et al.*, 2005). It is therefore important to determine the extent to which contemporary polymorphisms in these two species were present before their split (these constitute *ancestral polymorphisms*). This is also of

* Corresponding author. Tel: 0131-650-5750. Fax: 0131-650-6564. e-mail: Brian.Charlesworth@ed.ac.uk
[1] Present address: Unidade de Xenética Evolutiva, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, 15782- Santiago de Compostela, Spain.

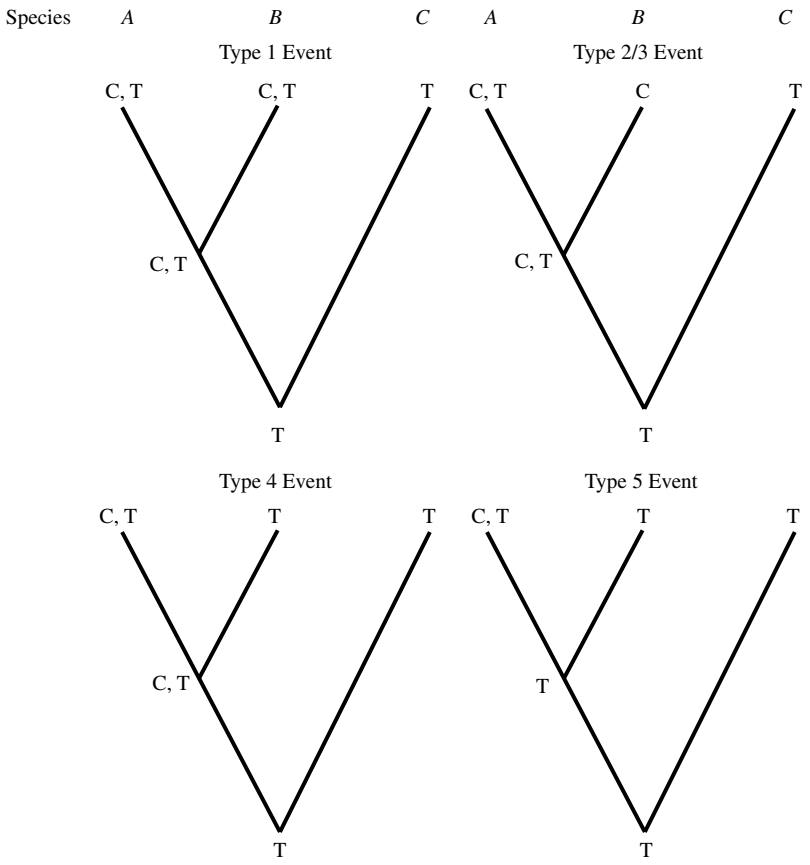

Fig. 1. The most parsimonious interpretations of the three observable patterns of polymorphism and divergence for two species (*A* and *B*) where polymorphism data are available, and an outgroup species (*C*) for which only a single sequence is known. Note that type 4 and 5 events are observationally indistinguishable, but involve different states of the ancestral population. The example assumes T to C transition mutations, but similar principles hold for other types of single nucleotide mutations.

interest in connection with the history of speciation in the group of species comprising *D. pseudoobscura* and its relatives (Wang *et al*., 1997). For this purpose, we assume that *D. miranda* and *D. pseudoobscura* have been completely isolated for some period of time. This assumption is based on the sterility of all viable hybrids formed between them (Dobzhansky & Tan, 1936).

2. Materials and methods

(i) *Nature of the problem*

We assume that we have a pair of related species, *A* and *B*, for which nucleotide site polymorphism data are available for a set of loci, and an outgroup species, *C*, for which a single sequence is known for each locus used in the polymorphism studies. Using parsimony, the status of a variant can be inferred by the following reasoning (Ramos-Onsins *et al*., 2004). Assume that we detect a polymorphism at a given nucleotide site in species *A*, for example C and T. If the outgroup species has a given nucleotide, say T, at the site in question, whereas both *A* and *B* have C and T, we infer that the common ancestor of *A* and *B* was polymorphic for C and T, with T being the ancestral state (in this case, a shared polymorphism has been detected: we designate this as a *type 1* event) (see Fig. 1). If the sequences from species *B* contain only C at this site, either the polymorphism exists in *B* but was missed due to small sample size (a *type 2* event). or C is truly fixed in *B* (a *type 3* event) (see Fig. 1). Clearly, type 2 and 3 events are indistinguishable empirically, but both correspond to a polymorphism that existed in the ancestral population. If species *B* contains only T, we similarly cannot distinguish between two possible events. First, there may be a shared polymorphism, with T being present by chance in all alleles sampled from *B* (a *type 4* event; see Fig. 1). Second, a *de novo* polymorphism may have arisen in species *A* as a result of a mutation from T to C after the split between *A* and *B* (a *type 5* event; see Fig. 1).

The same reasoning can obviously be applied to polymorphisms detected in species *B*. For both species, we can attempt to correct for misclassification of ancestral polymorphisms as *de novo* polymorphisms by estimating the fraction of ancestral polymorphisms

https://doi.org/10.1017/S0016672305007743 Published online by Cambridge University Press

Table 1. *Shared and ancestral synonymous polymorphisms in* D. pseudoobscura and D. miranda

	D. pseudoobscura				*D. miranda*		
Locus	*n*	'Shared' (type 1)	'Ancestral' (2/3)	'*De novo*' (4/5)	*n*	'Ancestral' (2/3)	'*De novo*' (4/5)
bcd	21	1	1	22	11	1	3
rosy	10	3	7	32	12	3	10
Adh	139	0	5	48	13	0	2
Total		4	13	102		4	15

that are identified as such i.e. as events of types 1–3. The observed frequency of these can then be adjusted by dividing by this fraction.

Inferences based on parsimony are subject to errors; corrections for these are discussed in the Appendix.

(ii) *Source of data*

We used data on polymorphisms in the two close relatives *D. miranda* and *D. pseudoobscura*, with the more distant species *D. affinis* as an outgroup. The *D. pseudoobscura* data were obtained from GenBank accessions, with sources of the alleles described by Schaeffer & Miller (1992) and Schaeffer (2002) (*Adh*), Machado *et al.* (2002) (*bcd*), and Riley *et al.* (1992) and Begun & Whitley (2002) (*rosy*). Alleles from the incipient species *D. pseudoobscura bogotana* were excluded from the analysis. The *D. miranda* and *D. affinis* sequences were obtained from the studies by Yi *et al.* (2003) and Bartolomé *et al.* (2005). Sequences were aligned and analysed using SeAl (http://evolve.zoo.ox.ac.uk/) and DnaSP (Rozas *et al.*, 2003), respectively. We restricted our analyses to synonymous variants rather than all silent variants, in order to avoid errors due to misalignment of non-coding sequences. For the purpose of determining whether a synonymous polymorphism observed in a sample from *D. pseudoobscura* is ancestral, we compared the state of the nucleotide site in *D. pseudoobscura* with those in a sample of 11–13 alleles from *D. miranda* and a single allele sampled from *D. affinis*, a more distantly related outgroup (Yi *et al.*, 2003; Bartolomé *et al.*, 2005). A similar procedure was applied to polymorphisms in *D. miranda*.

3. Results

(i) *Nucleotide polymorphisms*

We have only three genes for which there are polymorphism data for both *D. pseudoobscura* and *D. miranda*, as well as a sequence from *D. affinis*. The results of the determination by parsimony of the status of the polymorphisms in the two species are shown in Table 1. Overall, we inferred 4 of 119 *D. pseudoobscura* polymorphisms to be shared with *D. miranda* (type 1), 13 type 2 or 3 polymorphisms, and 102 type 4 or 5 polymorphisms. For *D. miranda*, there are 4 type 1, 4 type 2 or 3, and 15 type 4 or 5 polymorphisms. Taken at face value, this implies that a fraction $17/119 = 0{\cdot}143$ of all *D. pseudoobscura* polymorphisms, and $8/23 = 0{\cdot}348$ of *D. miranda* polymorphisms, are ancestral.

(ii) *Theoretical analysis of ancestral polymorphisms*

The extent of the bias in this estimate because of failure to recognize ancestral polymorphisms can be investigated as follows, using the principle outlined in Section 2. We assume that the polymorphisms whose nature is to be determined are ascertained in a sample of size m from species A; a sample of size n is taken from species B for the purpose of comparison. We also assume that use of the outgroup species C has enabled the mutant and ancestral states of each polymorphism to be determined. In the present case, parsimony may well have produced substantial biases in these determinations (see the Appendix), and the results of correcting for these are presented in the Discussion.

For ease of calculation, we also assume that synonymous mutations are neutral, although this is known not to be true for mutations altering codon usage in *Drosophila* (Akashi *et al.*, 1998; Bartolomé *et al.*, 2005). However, if weak selection acts on synonymous mutations, polymorphisms present in the ancestral population are expected to be enriched in mutations from P to U, where P represents the selectively favoured state and U represents the selectively disadvantageous state (Akashi *et al.*, 1998). This increases the probability that the mutant state will be lost from species B. The action of selection thus decreases our ability to detect ancestral polymorphisms.

We use the formula derived by Kimura (1955), discussed by Crow & Kimura (1970, pp. 383–387), for the probability density of frequency x of an allele at a time t after its frequency was p, in a population with

https://doi.org/10.1017/S0016672305007743 Published online by Cambridge University Press

effective size N_e. For convenience, we scale time to be measured in units of $2N_e$ generations. In our case, p corresponds to the frequency of the mutation at a given site in the ancestral population at the time of the split between the lineages leading to species A and B, and t represents the scaled time since divergence of the two species.

The general version of this formula (equation 8.4.6 of Crow & Kimura, 1970) is a complicated infinite series involving Gegenbauer polynomials (Korn & Korn, 1968); however, when time t is equal to 0·5 or more, the first two terms of the series dominate. (This was confirmed by detailed evaluation of the contributions of the next two higher-order terms, involving exp $(-6t)$ and exp $(-10t)$, to the quantities derived below. The relative errors from these terms are a few per cent at most for $t \geqslant 0{\cdot}5$.)

Ignoring the higher-order terms, we have

$$\phi(x,p,t) \approx 6p(1-p)\exp(-t) + 30p(1-p)(1-2p)(1-2x)\exp(-3t) \quad (1)$$

(Crow and Kimura, 1970, equation 8.4.7 [a misprint has been corrected]).

The restriction to $t \geqslant 0{\cdot}5$ may seem somewhat of a limitation on the method, but species which are less diverged than this are likely to share a good deal of their polymorphism, so that corrections of the kind described below are then largely unnecessary.

The probability that this polymorphism is present in a sample of size m from species A is

$$P_m(p) = \int_{1/(2N)}^{1} \{1-x^m-(1-x)^m\}\,\phi(x,p,t)\mathrm{d}x \quad (2a)$$

where N is the number of breeding adults.

Substituting from (1) into (2) into (3) and carrying out the integration (noting that, for most natural populations, $1/(2N)$ can be replaced by 0 in the integrals to a high level of accuracy), we find that the contribution from the second term in equation (1) vanishes, and so we have

$$P_m(p) \approx 6p(1-p)(m-1)\exp(-t)/(m+1). \quad (2b)$$

If the ancestral population is in drift-mutation equilibrium, we can assume that the probability density of p is given by the standard neutral formula i.e. it is proportional to $1/p$ (Ewens, 1979, p.238). We can then ask: What is the probability density $Q_m(p)$ that an ancestral polymorphism in the sample from A had a frequency p of the mutant variant in the ancestral population? By Bayes' theorem, we have

$$Q_m(p) = \frac{p^{-1}P_m(p)}{\int_{1/(2/N)}^{1} p^{-1}P_m(p)\mathrm{d}p} \quad (3)$$

and we obtain the simple expression

$$Q_m(p) \approx 2(1-p) \quad (4)$$

where the approximation indicates the use of the same number of terms in the series expansion of ϕ as in equation (1).

We can then determine the probability P_1 that a sample of n alleles from species B segregates for a polymorphism present in the ancestral population, given that this polymorphism was detected in the sample from species A (a *type 1* event, in the above terminology). This is given by

$$P_1 = \int_{1/(2N)}^{1}\int_{1/(2N)}^{1} \{1-x^n-(1-x)^n\}\phi(x,p,t)Q_m(p)\mathrm{d}x\,\mathrm{d}p \approx \frac{(n-1)}{(n+1)}\exp(-t) \quad (5)$$

(in this case, t is scaled by the effective population size for species B).

Similarly, the probability P_2 that the sample from species B is fixed for a mutation that was polymorphic in the ancestral population, and which is still polymorphic in species B (a *type 2* event), is

$$P_2 = \int_{1/(2N)}^{1}\int_{1/(2N)}^{1} x^n\phi(x,p,t)Q_m(p)\mathrm{d}x\,\mathrm{d}p \approx \left\{1-\frac{n}{(n+2)}\exp(-2t)\right\}\frac{\exp(-t)}{(n+1)}. \quad (6)$$

The probability P_3 that species B is fixed for a mutation that was polymorphic in the ancestral population (a *type 3* event) is

$$P_3 = \int_{1/(2N)}^{1} u(p,t)Q_m(p)\mathrm{d}p \quad (7)$$

where $u(p, t)$ is the probability that a variant present at frequency p is fixed in the population by time t.

From equation (8.4.12) of Crow & Kimura (1970), $u(p, t)$ is given to the same order of approximation as equation (1) by

$$u(p,t) \approx p - 3p(1-p)\exp(-t) + 5p(1-p) \times (1-2p)\exp(-3t) \quad (8)$$

so that

$$P_3 \approx \frac{1}{3} - \frac{1}{2}\exp(-t) + \frac{1}{6}\exp(-3t). \quad (9)$$

The probability that an ancestral polymorphism present in species A is classified as an ancestral

https://doi.org/10.1017/S0016672305007743 Published online by Cambridge University Press

polymorphism by the criteria defined above is simply the sum of the P_i from 1 to 3, P_d.

If $t \gg 1$, then we have

$$P_1 \approx 0, \quad P_2 \approx 0, \quad P_3 \approx \frac{1}{3}. \tag{10}$$

The asymptotic value of P_d is thus one-third. The more general expression is

$$P_d \approx \frac{1}{3} + \frac{(n-1)}{2(n+1)} \exp(-t) + \frac{\{(n+1)(n+2)-6n\}}{6(n+1)(n+2)} \exp(-3t). \tag{11a}$$

For moderate to large sample sizes, this can be approximated further by

$$P_d \approx \frac{1}{3} + \frac{1}{2} \exp(-t) + \frac{1}{6} \exp(-3t). \tag{11b}$$

(iii) *Estimation of the frequency of ancestral polymorphisms*

These formulae can be used to estimate the frequency of ancestral polymorphisms from the net probability of detecting type 1, 2 and 3 events (P_d), together with the observed proportion of polymorphisms that are in this category, as explained in section 2 (i). To do this, we need to have an estimate of t. One way to obtain this is to use the ratio of the frequency of events of type 1 to that of events of types 2 and 3. For $t > 1$, the dominant term controlling the frequencies of type 2 and 3 events is given by the sum of the first terms on the right-hand sides of equations (6) and (9), so we have

$$\frac{P_1}{(P_2+P_3)} \approx \frac{(n-1)}{\left\{1+\frac{(n+1)}{3}\exp(t)\right\}}. \tag{12}$$

The value of t can therefore be estimated by equating the left-hand side of equation (12) and the ratio of the observed number of type 1 events to the observed number of (indistinguishable) type 2 or 3 events. If this ratio is denoted by r_s, we have

$$\hat{t} \approx \mathrm{In}\left\{\frac{3([n-1]-r_s)}{(n+1)r_s}\right\}. \tag{13}$$

In the case of *D. pseudoobscura* (species *A*) and *D. miranda* (species *B*), we have $r_s = 0{\cdot}308 \pm 0{\cdot}219$ for the pooled data, and $n = 12$ on average, so that the estimate of t is 2·08. Given the rather small total number of events, the confidence interval on this (derived from the exact binomial distribution confidence interval for the proportion of type 1 events) is wide: 1·52–3·34. This method of obtaining confidence intervals assumes independence among sites, which is a reasonable approximation given the low levels of linkage disequilibrium found in these species (Schaeffer & Miller, 1992; Yi *et al.*, 2003).

An alternative method of estimating t is to use the ratio of the mean divergence at silent sites between the two species (K_S) to the within-species silent nucleotide site diversity (π_S) for *D. miranda* (Hudson *et al.*, 1987). Bartolomé *et al.* (2005) provide estimates of mean K_S and π_S of 0·032 and 0·0041, respectively. This gives a considerably higher t value of 7·80. The reason for this discrepancy is not clear, but may either reflect errors in the parsimony assignments (see Appendix), or a recent reduction in population size in *D. miranda*. This would reduce π_S but would have little effect on the P_i values, which are controlled by the long-term evolutionary process. In support of this interpretation, the mean silent site diversity in *D. pseudoobscura* is approximately 0·020, much higher than in *D. miranda* (Yi *et al.*, 2003; Bartolomé *et al.*, 2005). but it should be noted that *D. pseudoobscura* shows signs of a population expansion (Machado *et al.*, 2002), so that it is likely that its high diversity is in part of recent origin.

With the lowest of these t values (1·52), the frequency of ancestral polymorphisms detected in *D. pseudoobscura* (the total fraction of type 1, 2 and three polymorphisms) is about 0·44 times the true value, using equation (11b). With $t = 2{\cdot}08$, it is 0·392, for $t = 3{\cdot}34$ it is 0·368, and for $t = 7{\cdot}8$, it is 0·333. It thus seems likely that the frequency of ancestral polymorphisms in *D. pseudoobscura* is at least twice, and more probably around three times, the apparent frequency of 0·143 (see above). Using the latter value, the estimate of the fraction of all polymorphisms in *D. pseudoobscura* that are ancestral becomes 0·43, with approximate 95% confidence limits of $\pm 0{\cdot}19$. With $t = 2{\cdot}08$, the adjusted estimate for *D. miranda* is about 2·53 times the observed value, yielding an estimate of $0{\cdot}88 \pm 0{\cdot}49$. These estimates need, however, to be corrected for possible errors in assigning the status of variants by parsimony, as outlined in the Discussion.

(iv) *Fixations of ancestral polymorphisms*

Similar questions can also be asked about fixations that are detected on a given branch of the phylogeny connecting the two species A and B. A proportion F_1 of these fixations will be due to mutations that arose after the split, and $F_2 = 1 - F_1$ to polymorphisms present at the time of the split between the two species. The theoretical values of these proportions can be evaluated as follows. The expected number E_1 of fixations due to mutations arising after the split is approximated by the sum of $u(p, \tau)$ from $\tau = 0$ to $\tau = t$ at $p = 1/(2N)$, multiplied by the number of new mutations entering the population per unit time. The latter is equal to $(2N_e)(2Nkv)$ on the time-scale of $2N_e$

https://doi.org/10.1017/S0016672305007743 Published online by Cambridge University Press

generations, where k is the number of sites and v is the mutation rate per site.

Using equations (8.4.6) and (8.4.12) of Crow & Kimura (1970), neglecting second and higher order terms in the expansion of u with respect to p, and using the recursion relation for generating Gegenbauer polynomials (Korn & Korn, 1968), it follows by induction that E_1 is given by the following infinite series

$$E_1 \approx 2N_e kv\left\{t+\sum_{i=1}^{\infty}(-1)^i\frac{2(2i+1)}{i(i+1)} \times\left(1-\exp\left(-\frac{i(i+1)}{2}t\right)\right)\right\}$$

which simplifies to

$$E_1 \approx 2N_e kv\left\{t-2-\sum_{i=1}^{\infty}(-1)^i\frac{2(2i+1)}{i(i+1)} \times\exp\left(-\frac{i(i+1)}{2}t\right)\right\}. \quad (14)$$

From standard neutral theory (Wright, 1938; Kimura, 1968), the expected total number of fixations occurring over a time period t is equal to $2N_e kvt$, so that the expected number of fixations arising from ancestral polymorphisms, E_2, is simply $2N_e kvt - E_1$, i.e.

$$E_2 \approx 4N_e kv\left\{1+\sum_{i=1}^{\infty}(-1)^i\frac{(2i+1)}{i(i+1)} \times\exp\left(-\frac{i(i+1)}{2}t\right)\right\}. \quad (15)$$

This can be verified by the more laborious procedure of evaluating the integral of the fixation probabilities of polymorphic variants over their stationary frequency distribution, using equations (8.4.6) and (8.4.12) of Crow & Kimura (1970). Exact matrix calculations for population sizes of the order of 100 show that equations (14) and (15) provide excellent approximations.

We have $F_2 = E_2/(E_1+E_2)$. As expected, as t increases, F_2 approaches zero, and tends to one as t tends to zero. Even for $t>2$, it is surprisingly high, reflecting the fact that E_1 is asymptotically $2N_e kv$ $(t-2)$ and E_2 tends to $4N_e kv$. With $t=2$, we have $F_2 = 0{\cdot}799/(0{\cdot}799+0{\cdot}209) = 0{\cdot}796$; for $t=5$ it is 0·200.

These results ignores the possibility that a polymorphic mutation is misclassified as a fixation on the branch in question, due to its presence in all sampled alleles. The expected number of such events in a sequence of length k is $E_3 = 4N_e kv/n$ for sample size n. This follows from the fact that the expected number of polymorphisms with a mutation present at frequency x is $4N_e kvx^{-1}$ (Ewens, 1979, p.276), and the chance that a sample of size n is fixed for such a variant is x^n. Given that diversity per site is $4N_e v$ (whose estimated value is at most 1·75% for the *D. miranda* genes used here), and n is 11–13, this term can be neglected in the present case. In general, however, the fraction of apparent fixations that represent new mutations that have truly gone to fixation after the population split is $E_2/(E_1+E_2+E_3)$.

4. Discussion

Our methods indicate that the frequency of polymorphisms in a species that were present in its common ancestor with a close relative may be much higher than is indicated by the fraction of polymorphisms that are directly inferred to be ancestral, using information on an outgroup (section 3(iii)). We can also use data on polymorphism and divergence for two related species, without considering an outgroup. In this case, we employ equation (2) to calculate the expected number of ancestral polymorphisms at k sites in a sample of size m from one of these species. This uses the fact that the expected number of such polymorphisms at frequency p in the population is $4N_e kvp^{-1}$ (see section 3(iii) above); integrating the equivalent of P_m over p, as in the denominator of equation (3), yields the expected number of ancestral polymorphisms as $12N_e kv(m-1)\exp(-t)/(m+1)$. The expected total number of polymorphisms in the sample is $4N_e kvS_m$, where S_m $1+1/2+1/3+\ldots$ $1/(m-1)$ (Ewens, 1979, p.276). The ratio of these is $3(m-1)\exp(-t)/\{(m+1)\ S_m\}$, which is the *a priori* probability that a polymorphism is ancestral, given t.

A more accurate expression, which is useful when t is smaller than 0·5, can be obtained by evaluating two higher order exponential terms in the expression for the probability density ϕ, additional to those displayed in equation (1) (the net contribution from terms in $\exp(-10t)$ is found to be zero):

$$\left\{\frac{3(m-1)}{(m+1)}\exp(-t)+\frac{7}{6}\left(1-\frac{12(m^2+1)}{(m+1)(m+2)(m+3)}\right) \times\exp(-6t)\right\}/S_m. \quad (16)$$

For *D. pseudoobscura* polymorphisms, t for divergence from *D. miranda* is about 1·6, as estimated from the ratio of divergence to diversity in *D. pseudoobscura* (see 3 (iii) above). In the case of *rosy*, the expected proportion of ancestral polymorphisms is 0·175, slightly but not significantly lower than the fraction actually observed (0·200) and much smaller than the proportion estimated after the corrections described in section 3(iii) (more than 0·400). This may well reflect biases in the identification of ancestral

https://doi.org/10.1017/S0016672305007743 Published online by Cambridge University Press

polymorphisms by the use of parsimony, as discussed in the Appendix. The likelihood of such error is quite large when the divergence of *A* and *B* from the outgroup species *C* is as high as it is here (a mean value of K_s of 0·22; Bartolomé *et al*., 2005). If the inferred number of types 1 and type 2/3 polymorphisms is only 8 after correcting for parsimony errors, as suggested by the analysis in the Appendix, the final estimate of the net frequency of ancestral polymorphisms after using the corrections proposed in section 3(iii) becomes approximately 0·186, which is not significantly different from the *a priori* value of 0·147 for the three loci pooled (weighting the value for each gene by the number of polymorphic sites). This suggests that our estimates for *D. pseudoobscura* are moderately reliable. Too few data are available for *D. miranda* polymorphisms to make this procedure worthwhile.

Examination of discrepancies between *a priori* and estimated frequencies of ancestral polymorphisms in larger datasets, especially where the use of a closer outgroup allows more certain identification of the status of polymorphisms, would provide a means for testing for the existence of larger amounts of ancestral polymorphism than expected under neutrality, as expected with long-terms balanced polymorphism, or for smaller amounts, as expected with directional selection. This approach should be more powerful than the comparison of observed and expected levels of shared polymorphisms, as has been used previously (Clark, 1997; Wiuf *et al*., 2004; Asthana *et al*., 2005).

We have also shown that a high proportion of fixations subsequent to the divergence of two related species may be contributed by ancestral polymorphisms rather than new mutations, even as many as $10N_e$ generations since the split (section 3(iv)). This reflects the long tail in the probability distribution of the time that variants remain segregating in a population (Clark, 1997). Identification of fixed differences by using polymorphism data thus does not guarantee that the fixed variants represent mutations that arose since the divergence of the species. This has implications for estimates of selection on codon usage bias, or the intensity of biased gene conversion on noncoding sequences, as noted by Bartolomé *et al*. (2005). The standard test for equilibrium with respect to base composition and/or codon usage is to compare the numbers of fixations in each direction (e.g. GC to AT mutations versus AT to GC) over a large number of sites. These are expected to be the same if base composition is in equilibrium. However, this expectation applies only to mutations arising *de novo* and not to fixations derived from polymorphisms present at the time of the split. From the standard formula for fixation probability under additive selection in a finite population (Crow & Kimura, 1970, p.426), it is easily seen that the fixation probability of a deleterious mutation relative to that for an advantageous mutation with the same selection coefficient increases with the initial frequency of the mutation. It follows that there will be a relative enrichment of selectively disfavoured mutations among fixations of ancestral polymorphisms, so that an excess of such events among closely related species does not necessarily imply a non-equilibrium situation. The extent of this bias depends on the strength of selection, and requires numerical investigation, which we plan to carry out in a future study.

It is useful to note, however, that the reasoning leading to equations (14) and (15) implies that the number of fixed neutral differences between two species can be used to estimate their divergence time, even in the presence of a substantial fraction of fixations arising from ancestral polymorphisms, providing that there have not been radical changes in population size since divergence. This is because the result that the number of neutral fixations over a fixed time interval depends only on the product of time and mutation rate (Wright, 1938; Kimura, 1968) is independent of the time of origination of the mutations in question. Estimates of mutation rates for truly neutral mutations from divergence between isolated populations should therefore be independent of elapsed time, provided that within-population variability has been removed. This is relevant to the interpretation of recent evidence for apparent dependence of estimated mutation rates on divergence time (Ho *et al*., 2005), suggesting that these must reflect the effects of selection, population bottlenecks, or failure to correct for within-population variability.

This work was supported by grants from the Biotechnology and Biological Sciences Research Council and the Royal Society to BC. We thank an anonymous reviewer for pointing out the simplification of the power series leading to equation (14), and Jody Hey and both reviewers for suggestions for improving the manuscript.

Appendix

Here we examine the likely errors resulting from the use of parsimony in inferring the status of polymorphisms, as represented in Fig. 1.

(i) *Errors in inferring type 1 polymorphisms*

The alternative to an observed type 1 polymorphism being a true shared polymorphism is that independent mutations causing the same polymorphisms arose in species *A* and *B*. Since we are conditioning on having observed a polymorphism at a given site in one species (e.g. *A*), we need to consider the chance that this polymorphism arose *de novo*, and that an independent *de novo* mutation of the same type also is observed at that site in the other species. Unless the site in

question is exceptionally mutable, the chance that a *de novo* polymorphism exists at this site in species *B* is the product of the complement of the *a priori* probability that a polymorphism in species *B* is ancestral, and $4N_e vS_n$ (see sections 3(iii) and 4). In the case of *D. miranda* as species *B*, the maximum nucleotide site diversity is for *rosy* (0·0175), which can be equated to $4N_e v$, and $S_n = 3{\cdot}02$ An upper bound estimate of the chance that an independent *de novo* polymorphism was established in *B* at *rosy* is provided by $3{\cdot}02 \times 0{\cdot}0175 = 0{\cdot}053$, so that, out of a total of 42 polymorphisms at *rosy* in *D. pseudoobscura*, we expect less than 2·06 independent *de novo* polymorphisms at the same sites in *D. miranda*. This should be weighted by the *a priori* probability that the *A* polymorphism is *de novo*, which was estimated to be approximately 0·825 in section 4 above, reducing the number to 1·70.

In addition, we need to include the probability that the two independent mutations are identical. Examination of the pattern of polymorphism in these genes in *D. miranda* and *D. pseudoobscura* shows that 98/122 independent polymorphisms are transitions, so that the chances that two independent mutations are transitions or transversions are 0·64 and 0·04, respectively. The overall probability that two mutations derived from the same ancestral state are identical is thus approximately $0{\cdot}64 + 0{\cdot}02 = 0{\cdot}66$, so that the final estimate of the expected number of shared polymorphisms is $1{\cdot}70 \times 0{\cdot}66 = 1{\cdot}12$. Similar calculations for *bcd* and *Adh* suggest that the total expected number of spurious type 1 polymorphisms in the pooled set for *D. pseudoobscura* may be as high as 1·76, compared with the 4 observed.

(ii) *Errors in inferring type 2/3 and 4/5 polymorphisms*

Here, the most likely alternative to the parsimonious interpretation shown in Fig. 1 is that the common ancestor was C, that the lineage leading to species *C* experienced a C to T mutation, and that C mutated in species *A* to give the C, T polymorphism. (We will ignore the much less likely case of dual events on the lineages leading to species *A* and *B* from their common ancestor.) The above estimate of the chance of identical mutations at the same site (0·66) will be used. If the site in question is not unusually mutable, the chance of the substitution along the *C* lineage being the same as a mutation generating a *de novo* polymorphism in *A* is thus approximately $0{\cdot}33K_s$, where K_s is the divergence between species *A* or *B* and *C* (see Table 3 of Bartolomé *et al.*, 2005 for the values of K_s for each locus). Again, this should be weighted by the *a priori* probability that the polymorphism in *A* is *de novo*. For *rosy*, the expected number of false type 2/3 polymorphisms among the total of types 2/3 and 4/5 is thus $39 \times 0{\cdot}099 \times 0{\cdot}825 = 3{\cdot}18$, compared to the 7 inferred. For the pooled set of 3 loci, the value is 7·67, compared with an inferred total of 13.

However, a similar argument also applies to type 4/5 polymorphisms; the most likely alternative to the parsimonious interpretation of the example in Fig. 1 is that a C to T mutation occurred in the lineage leading to species *C*, and in the common ancestor of *A* and *B*. In this case, the probability that the sample of alleles from *B* is fixed for an ancestral polymorphism present in *A* is given by $P_2 + P_3$ in equations (6) and (9). For large *t*, this is approximately 0·33. This is discounted by the product of $0{\cdot}33K_s$ and the *a priori* probability that the C to T polymorphism in *A* is ancestral. The expected number of false 4/5 polymorphisms for the pooled set of loci is the product of this and the total number of apparent 2/3 and 4/5 polymorphisms; the sum of this over each locus is 0·50, which should be deducted from the above number of false 2/3 polymorphisms, leading to an overall estimate of an expected number of 7·17 false 2/3 polymorphisms.

The use of parsimony is thus likely to produce a substantial bias in favour of overestimation of the frequency of type 2/3 polymorphisms, working in the opposite direction to the correction proposed in section 3(iii). The extent of this bias would clearly be greatly reduced by the use of a closer outgroup species, which unfortunately is hard to obtain in the present case.

References

Akashi, H., Kliman, R. M. & Eyre-Walker, A. (1998). Mutation pressure, natural selection and the evolution of base composition in *Drosophila*. *Genetica* **102/103**, 49–60.

Arbogast, B. S., Edwards, S. V., Wakeley, J., Beerli, P. & Slowinski, J. B. (2002). Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics* **33**, 707–740.

Asthana, S., Schmidt, S. & Sunyaev, S. (2005). A limited role for balancing selection. *Trends in Genetics* **21**, 30–32.

Bachtrog, D. (2003). Protein evolution and codon usage bias on the neo-sex chromosomes of *Drosophila miranda*. *Genetics* **165**, 1221–1232.

Bartolomé, C., Maside, X., Yi, S., Grant, A. L. & Charlesworth, B. (2005). Patterns of selection on synonymous and non-synonymous variants in *Drosophila miranda*. *Genetics* **169**, 1495–1507.

Begun, D. J. & Whitley, P. (2002). Molecular population genetics of Xdh and the evolution of base composition in Drosophila. *Genetics* **162**, 1725–1735.

Clark, A. G. (1997). Neutral behavior of shared polymorphism. *Proceedings of the National Academy of Sciences of the USA* **94**, 7730–7734.

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper and Row.

Dobzhansky, T. & Tan, C. C. (1936). Studies on hybrid sterility. III. A comparison of the gene order in two species, *Drosophila pseudoobscura* and *Drosophila miranda*. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* **72**, 88–114.

Ewens, W. J. (1979). *Mathematical Population Genetics*. Berlin: Springer-Verlag.

Ho, S. Y. W., Phillips, M. J., Cooper, A. & Drummond, A. J. (2005). Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution* **22**, 1561–1568.

Hudson, R. R., Kreitman, M. & Aguadé, M. (1987). A test of molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Kimura, M. (1955). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology* **20**, 33–53.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Korn, G. A. & Korn, T. M. (1968). *Mathematical Handbook for Scientists and Engineers*, *2nd ed*. NewYork, NY: McGraw-Hill.

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

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

Powell, J. R. (1997). *Progress and Prospects in Evolutionary Biology*. *The Drosophila Model*. New York: Oxford University Press.

Ramos-Onsins, S. E., Stranger, B. E., Mitchell-Olds, T. & Aguadé, M. (2004). Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. *Genetics* **166**, 373–388.

Riley, M. A., Kaplan, S. R. & Veuille, M. (1992). Nucleotide polymorphism at the xanthine dehydrogenase locus in *Drosophila pseudoobscura*. *Molecular Biology and Evolution* **9**, 56–69.

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497.

Schaeffer, S. W. (2002). Molecular population genetics of sequence length diversity in the Adh region of *Drosophila pseudoobscura*. *Genetical Research* **80**, 163–175.

Schaeffer, S. W. & Miller, E. L. (1992). Molecular population genetics of an electrophoretically monomorphic protein in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* **132**, 163–178.

Steinemann, M. & Steinemann, S. (1998). Enigma of Y chromosome degeneration: neo-Y and neo-X chromosomes of *Drosophila miranda* a model for sex chromosome evolution. **102/103**, 409–420.

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

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

Wang, R. L., Wakeley, J. & Hey, J. (1997). Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* **147**, 1091–1106.

Wiuf, C., Zhao, K., Innan, H. & Nordborg, M. (2004). The probability and chromosomal extent of transpecific polymorphism. *Genetics* **168**, 2363–2372.

Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the USA* **24**, 253–259.

Yi, S., Bachtrog, D. & Charlesworth, B. (2003). Genetic and karyotypic variation in *Drosophila miranda*. *Genetics* **164**, 1369–1381.