Copyright © 2004 by the Genetics Society of America
DOI: 10.1534/genetics.104.029488

The Probability and Chromosomal Extent of *trans*-specific Polymorphism

Carsten Wiuf,*,[1] Keyan Zhao,† Hideki Innan†,[2] and Magnus Nordborg†,[3]

**Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom and †Program in Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089-1340*

Manuscript received March 30, 2004
Accepted for publication September 8, 2004

ABSTRACT

Balancing selection may result in *trans*-specific polymorphism: the maintenance of allelic classes that transcend species boundaries by virtue of being more ancient than the species themselves. At the selected site, gene genealogies are expected not to reflect the species tree. Because of linkage, the same will be true for part of the surrounding chromosomal region. Here we obtain various approximations for the distribution of the length of this region and discuss the practical implications of our results. Our main finding is that the *trans*-specific region surrounding a single-locus balanced polymorphism is expected to be quite short, probably too short to be readily detectable. Thus lack of obvious *trans*-specific polymorphism should not be taken as evidence against balancing selection. When *trans*-specific polymorphism is obvious, on the other hand, it may be reasonable to argue that selection must be acting on multiple sites or that recombination is suppressed in the surrounding region.

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

MOST species appear to be monophyletic for most of their genomes. That is, most sites in most genomes have the property that, with respect to that site, all homologous chromosomes in one species are more closely related to each other than they are to any homologous chromosome from another species. This behavior is expected from standard population genetics theory, as long as the species became reproductively isolated sufficiently long ago (see, for example, Hudson and Coyne 2002; Rosenberg 2002). However, exceptions are expected when balancing selection has maintained two or more alleles since the time of speciation. When this occurs, an allele sampled from a particular species may well be more closely related to members of the same functional allelic class in related species than to members of different allelic classes in the same species. This is referred to as *trans-specific polymorphism* (Klein 1980).

A few clear cases of *trans*-specific polymorphism have been found, in particular, in the MHC (*e.g.*, Figueroa *et al.* 1988) and plant self-incompatibility loci (*e.g.*, Ioerger *et al.* 1990). At the same time, studies of sequence variability in several genes that might *a priori* be considered good candidates for *trans*-specific polymorphism have failed to find strong evidence for this hypothesis. Examples include the primate ABO blood group system (Saitou and Yamamoto 1997) and red/green color vision polymorphism in New World monkeys and lemurs (Boissinot *et al.* 1998; Tan and Li 1999). In these cases, are the functionally similar alleles in different species examples of *trans*-specific polymorphism, or are they due to convergent evolution? The purpose of this article is to develop a modeling framework that allows us to address these questions. We focus in particular on our ability to detect *trans*-specific polymorphism when it exists and how this is determined by the length of the chromosomal region that is affected by the presence of a *trans*-specific polymorphism.

Throughout, we discuss relatedness in the genealogical sense, *i.e.*, with reference to "descent" rather than to allelic "state." Thus, when we say that two homologous copies of a site (or locus or nonrecombining sequence) are more closely related to each other than to a third copy, we mean that the most recent common ancestor (MRCA) of these two is more recent than the MRCA of either of them and the third copy. This does not necessarily mean that the two copies are more similar to each other than either is to the third copy (although if they are not, we would typically not be able to infer the true relationship).

BASIC MODEL

We are interested in the following scenario. An ancestral species/population splits into two τ units of time ago. For simplicity, both descendant populations are assumed to be of the same size as the original one. All populations evolve according to a standard neutral model, the coalescent approximation is employed, and time is measured

[1] *Present address:* Bioinformatics Research Center, University of Aarhus, DK-8000 Århus C, Denmark.

[2] *Present address:* Human Genetics Center, School of Public Health, University of Texas, Health Science Center, 1200 Hermann Pressler, Houston, TX 77030.

[3] *Corresponding author:* Molecular and Computational Biology, University of Southern California, SHS 172, 835 W. 37th St., Los Angeles, CA 90089-1340. E-mail: magnus@usc.edu

in units of the effective number of homologous chromosomes in each of the current populations. We consider both selective neutrality and various forms of balancing selection.

We consider the following questions for samples of homologous sequence taken from the two species:

What is the probability that the genealogy of a particular site does not reflect the species tree, *i.e.*, that the samples from the two species are not both monophyletic? *We refer to sites with this property as trans-specific.*

Given that a particular site (or sites) is (or are) *trans*-specific, what is the probability that a linked site is also *trans*-specific?

Given that a particular site (or sites) is (or are) *trans*-specific, what is the distribution of the length of the chromosomal region for which this remains true?

PROBABILITY OF *TRANS*-SPECIFICITY

Consider a sample of n_1 homologous copies of a site from species 1 and n_2 from species 2. The number of ancestral lineages decreases back in time according to a death process. The probability of *trans*-specificity may be calculated by first conditioning on the number of surviving lineages in each species at the time of speciation, τ, and then calculating the probability that these lineages coalesce in the ancestral species in a *trans*-specific manner. Assuming neutrality, the numbers of surviving lineages in each species at τ are independent, identically distributed random variables whose distribution is given by Tavaré (1984), and the conditional probability of *trans*-specificity can be found using the results of Saunders *et al.* (1984). The expression for the total probability can easily be evaluated numerically: see Nordborg (2001), for example. Numerous treatments of the probability of *trans*-specificity exist (*e.g.*, Pamilo and Nei 1988; Takahata 1989; Hey 1994; Hudson and Coyne 2002; Rosenberg 2002); our main purpose here is to introduce the basic concepts and to enable comparisons with later results.

Trans-specificity is impossible unless there are at least three ancestral lineages at the time of speciation. If there are two lineages in one species and one in the other, then the probability of *trans*-specificity is 2/3 (*trans*-specificity is avoided if and only if the two lineages from the same species coalesce with each other; this happens in one out of three equally likely topologies). The probability of *trans*-specificity is higher if there are more than three ancestral lineages. Thus the probability of *trans*-specificity is at least 2/3 given that there are at least three lineages at the time of speciation. The probability that there are at least three lineages at the time of speciation, on the other hand, decreases sharply with τ and can be vanishingly small. The probability of *trans*-specificity is thus mainly determined by this latter probability.

To put this into context, consider a sample of size two. The probability that the MRCA of the sample predates speciation is $e^{-\tau}$. Let T be the time until the MRCA for two genes, and note that under our model, $E[T] = 1$ for two genes sampled from the same species, whereas $E[T] = \tau + 1$ for two genes sampled from different species. Since the average number of pairwise differences between sequences is proportional to pairwise coalescence times under neutrality, an estimate of τ can be obtained as the ratio of the average number of pairwise differences between and within species, -1. For example, if, on average, humans and chimps are at most 99% identical, and humans and humans are at least 99.9% identical, then $\tau \geq 10^{-2}/10^{-3} - 1 = 9$. Let us say $\tau = 8$ to be on the safe side. Then the probability that the MRCA of a sample from humans predates speciation from chimps would be $e^{-8} = 3.3 \times 10^{-4}$ (and probably much smaller).

This is a small number, but the genome is large. If there are G sites in the genome, then we expect $Ge^{-\tau}$ to have MRCAs that predate speciation. If we consider the whole population rather than just two copies of the genome, the expected number of sites with MRCAs that predate speciation increases about threefold: the time until there are two ancestral lineages is ~ 1, so the expected number of sites is $\sim Ge^{-(\tau-1)} \approx 3Ge^{-\tau}$. For the ranges of τ we are interested in, qualitative conclusions are unaffected by sample size. For simplicity, we therefore discuss mainly samples of size two throughout this article.

The probability of *trans*-specificity for a site depends on whether the site is polymorphic or not. The calculations above assumed no knowledge of allelic state. What is the probability that $T > \tau$ for a polymorphic locus A, *i.e.*, for a sample of two different alleles? We consider the process that keeps track of the number of ancestral lineages in each of the two allelic classes. Denote the state of the process at time t by $X_t = (i, j)$, where i is the number of lineages in the first allelic class, and j is the number of lineages in the second allelic class. The probability we seek is

$$P_\tau = \mathbb{P}(X_\tau \notin \{(1, 0), (0, 1)\} | X_0 = (1, 1)),$$

which we write as $P_\tau = Q_\tau / Q_0$, with

$$Q_\tau = \mathbb{P}(X_\tau \notin \{(1, 0), (0, 1)\}, X_0 = (1, 1)).$$

We consider two cases: unidirectional mutation and bidirectional mutation. For the case of unidirectional mutation, assume that allele A_1 mutates into A_2 at rate $\theta/2$ and that further mutation in A_2 does not change the allelic state (we think of A_2 as a loss-of-function allele: this case is motivated by the observation that many examples of balancing selection appear to involve such mutations). Using standard population genetics theory (Hudson 1990), we find

$$Q_\tau = 2\int_\tau^\infty (1 - e^{-\nu\theta/2}) e^{-\nu\theta/2} e^{-\nu} d\nu$$

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

$$= \frac{4}{2+\theta}e^{-(1+\theta/2)\tau} - \frac{2}{1+\theta}e^{-(1+\theta)\tau}$$

and

$$Q_0 = \frac{2\theta}{(1+\theta)(2+\theta)},$$

so that

$$P_\tau = \frac{2(1+\theta)}{\theta}e^{-(1+\theta/2)\tau} - \frac{(2+\theta)}{\theta}e^{-(1+\theta)\tau}.$$

For the case of bidirectional mutation, assume that alleles A_1 and A_2 mutate back and forth at rate $\theta/2$. Here we find

$$Q_\tau = 2\int_\tau^\infty \frac{1}{2}e^{-\theta\nu}\sinh(\theta\nu)e^{-\nu}d\nu$$

$$= \frac{1}{2}e^{-\tau} - \frac{1}{2(1+2\theta)}e^{-(1+2\theta)\tau},$$

$$Q_0 = \frac{\theta}{1+2\theta},$$

and

$$P_\tau = \frac{1+2\theta}{2\theta}e^{-\tau} - \frac{1}{2\theta}e^{-(1+2\theta)\tau}.$$

These results make intuitive sense: for small θ, $P_\tau \approx (1+\tau)e^{-\tau}$ in both cases. The probability of *trans*-specificity conditional on polymorphism is higher than the unconditional probability because the fact that a (rare) mutation must have occurred automatically pushes the time to the MRCA further back in time. For large θ, $P_\tau \approx 0$ with unidirectional mutation and $P_\tau \approx e^{-\tau}$ with bidirectional mutation. In the former case, the MRCA must be recent or all A_1 would have mutated to A_2, whereas in the latter case, mutations occur so frequently that the allelic states tell us nothing about the age of the MRCA.

The main conclusion from the above discussion, however, is that no matter which model is used, the probability of *trans*-specificity under neutrality is always very low for large τ (for recent attempts to estimate it directly, see Chen and Li 2001; O'hUigin *et al.* 2002). In contrast, if some form of balancing selection is acting, *trans*-specificity becomes highly probable. Selected alleles will of course also be lost through genetic drift, but this occurs over entirely different timescales (Takahata 1990; Vekemans and Slatkin 1994), and it is easy to imagine strengths of selection that make loss of polymorphism during speciation unlikely even if one believes that speciation is accompanied by genetic bottlenecks (Vincek *et al.* 1997). *Trans*-specificity may therefore, in and of itself, be viewed as evidence for a history of balancing selection. But how do we detect *trans*-specificity? To consider the traces of *trans*-specificity in sequence data, we need to know something about its chromosomal extent. This is the topic of the following sections.

THE EXTENT OF *TRANS*-SPECIFICITY

What is the probability that a locus is *trans*-specific given that it is linked to a locus that is *trans*-specific? Let the recombination rate between the two loci be $\rho/2$, where $\rho = 4Nr$, N is the effective population size, and r is the recombination fraction, and consider a sample of size two. The site is *trans*-specific if no recombination occurs before coalescence at the other site. The probability of this is

$$\int_\tau^\infty e^{-\nu\rho}e^{-\nu}d\nu \Big/ \int_\tau^\infty e^{-\nu}d\nu = \frac{e^{-\rho\tau}}{1+\rho}. \quad (1)$$

If, as suggested above, $\tau \approx 8$ for humans and chimps, and ρ per site is 5×10^{-4} (Przeworski *et al.* 2000; Pritchard and Przeworski 2001; Innan *et al.* 2003), then the probability is >60% for sites separated by 100 bp, but it decreases rapidly to 1% for 1 kb. Linkage to a *trans*-specific site increases the probability of *trans*-specificity for tighly linked sites dramatically, but we should not expect large chromosomal regions to be *trans*-specific (at least not due to linkage).

The probability just derived is an underestimate: the focal site can of course be *trans*-specific without being identical by descent (with respect to recombination) to the conditional one. Most importantly, whereas two lineages linked to a *trans*-specific site cannot coalesce before τ without at least one recombination, a single recombination does not allow them to coalesce unless it occurs between descendants of different *trans*-specific lineages ("moving" the two lineages into the same *trans*-specific lineage). The probability of this depends on the frequency of descendants of each *trans*-specific lineage in every generation back to τ.

To take this into account, we consider the model of balancing selection first described by Hudson and Kaplan (1988) and extended by Nordborg and Innan (2003). Imagine that some form of strong balancing selection maintains two alleles, A_1 and A_2, at a locus. Selection is strong enough to maintain the alleles at frequencies x and $1 - x$, respectively. The recombination rate between the locus under selection and the locus of interest is $\rho/2$, as before. Depending on the allelic state at the former locus, each haplotype belongs to one of two allelic classes. The state of a sample of size two from the focal locus can be described by (z_1, z_2), where z_i denotes the number of lineages belonging to the A_i allelic class. The ancestry of the sample can be described by the Markov process $z = (z_1, z_2)$ with states (1, 1), (2, 0), (0, 2), (1, 0), (0, 1). Let i = 1, 2, 3, 4, 5 refer to these states in the order given. The rate matrix $Q = \{q_{ij}\}_{i,j}$ of z is

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

$$Q = \begin{bmatrix} - & \rho x/2 & \rho(1-x)/2 & 0 & 0 \\ \rho(1-x) & - & 0 & 1/x & 0 \\ \rho x & 0 & - & 0 & 1/(1-x) \\ 0 & 0 & 0 & - & 0 \\ 0 & 0 & 0 & 0 & - \end{bmatrix}$$

with diagonal elements given by $q_{ii} = -\Sigma_j q_{ij}$. The states (1, 0) and (0, 1) are absorbing, and the process starts in (1, 1).

Probability of *trans*-specificity: We are interested in $P_\tau(\rho, x)$, the probability that two lineages that start in (1, 1) are still distinct at the time of speciation. An exact solution can be found using standard methods. For $x = ½$ we find

$$P_\tau(\rho, ½) = \frac{1}{2\sqrt{1+\rho^2}}(c_2 e^{-c_1\tau} - c_1 e^{-c_2\tau}), \tag{2}$$

where $c_1 = 1 + \rho - \sqrt{1+\rho^2}$ and $c_2 = 1 + \rho + \sqrt{1+\rho^2}$. The solution for general x is highly intractable, but it can be shown that $\lim_{\rho\to\infty} P_\tau(\rho, x) = e^{-\tau}$, in agreement with Equation 2 and with our intuition for unlinked loci.

Several approximations are possible for $\rho \approx 0$. We consider two: the first is the best one we found; the second, the simplest.

Approximation 1: The first approximation is obtained by modifying the rate matrix Q so that the recombination rate is set to zero once the process has left (1, 1). This prevents the process from reentering (1, 1), which simplifies calculations considerably. It is readily verified that this modified process corresponds to the original one in the limit $x \to 0$ or $x \to 1$, so the approximation is exact for these cases. Using the modified matrix, we find

$$P_\tau(\rho, x) \approx e^{-\rho\tau} + (e^{-\rho\tau} - e^{-\tau/x})\frac{\rho x^2}{1-\rho x} + (e^{-\rho\tau} - e^{-\tau/(1-x)})\frac{\rho(1-x)^2}{1-\rho(1-x)}. \tag{3}$$

Approximation 2: Assume that the lineages stay distinct if and only if no recombination occurs. This yields

$$P_\tau(\rho, x) \approx e^{-\rho\tau}, \tag{4}$$

which should be compared to Equation 1.

Equations 2–4 give the probability that two lineages linked to different alleles in a balanced polymorphism stay distinct until speciation. If this happens, *trans*-specificity is highly probable for the range of parameters in which we are interested ($\rho \approx 0$, $\tau \gg 1$): one of the two lineages is likely to coalesce with lineages within the same allelic class from the other species long before a recombination event occurs. Equations 2–4 can thus be seen as approximations of the same probability as Equation 1.

Returning to our human-chimp example, and assuming $x = ½$, we find that Equations 2–4 give probabilities of *trans*-specificity of 69, 69, and 67%, respectively, for sites separated by 100 bp; and 6, 2, and 2%, respectively, for sites separated by 1 kb. As predicted, Equation 1 underestimates the probability of *trans*-specificity; however, the results are qualitatively similar. It can be shown that the probability is increased further when $x \neq ½$: intuitively, this is because the probability of recombination between the allelic classes is maximized when allele frequencies are even. With $x = 0.01$, Equation 3 gives 70 and 4%, respectively, in the above two cases.

Length of *trans*-specificity: Let $L_\tau(\rho, x)$ be the length of the region on one side of the site under selection where two haplotypes from different allelic classes still have distinct lineages at time τ. $L_\tau(\rho, x)$ is possibly the total length of a number of disjoint intervals. We have $\mathbb{E}[L_\tau(\rho, x)] = \int_0^\rho P_\tau(u, x)\,du$. From this, and by considering the properties of P_τ, it follows that for arbitrary $0 < x < 1$,

$$\lim_{\rho\to\infty}\frac{\mathbb{E}[L_\tau(\rho, x)]}{\rho} = e^{-\tau}, \tag{5}$$

$$\lim_{\rho\to 0}\frac{\mathbb{E}[L_\tau(\rho, x)]}{\rho} = 1, \tag{6}$$

$$\lim_{\tau\to\infty}\mathbb{E}[L_\tau(\rho, x)] = 0, \tag{7}$$

and

$$\lim_{\tau\to 0}\mathbb{E}[L_\tau(\rho, x)] = \rho. \tag{8}$$

These equations are useful for evaluating approximations to $\mathbb{E}[L_\tau(\rho, x)]$, the exact value of which is not known for any x. The two approximations introduced above can be applied, however.

Approximation 1: The density of $L_\tau(\rho, x)$ can be approximated by

$$-\frac{1}{1 - xe^{-\tau/x} - (1-x)e^{-\tau/(1-x)}}\frac{d}{du}P_\tau(u, x),$$

where $P_\tau(u, x)$ is given by Equation 3, but the expectation cannot be obtained analytically. We refer to this expectation as $\mathbb{E}_1[L_\tau(\rho, x)]$. Equations 6–8 hold for $\mathbb{E}_1[L_\tau(\rho, x)]$, but Equation 5 does not. Instead we have

$$\lim_{\rho\to\infty}\frac{\mathbb{E}_1[L_\tau(\rho, x)]}{\rho} = xe^{-\tau/x} + (1-x)e^{-\tau/(1-x)}.$$

Note that Equation 5 holds if $x \approx 0$ or $x \approx 1$.

Approximation 2: $L_\tau(\rho, x)$ is approximately exponential with intensity τ, Exp(τ), truncated at ρ, and the expectation is

$$\mathbb{E}[L_\tau(\rho, x)] \approx \mathbb{E}_2[L_\tau(\rho, x)] = \int_0^\rho e^{-u\tau}du = \frac{1 - e^{-\rho\tau}}{\tau},$$

where $P_\tau(u, x)$ is given by Equation 4. Equations 6–8

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

TABLE 1

The performance of approximations for $\mathbb{E}[L_\tau(\rho, x)]$

	$\rho =$				
	0.1	1	10	100	1000
		$\tau = 5$			
$\mathbb{E}[L_\tau(\rho, x)]$	0.081	0.271	0.410	1.07	7.18
$\mathbb{E}_1[L_\tau(\rho, x)]$	0.081	0.223	0.227	0.231	0.272
$\mathbb{E}_2[L_\tau(\rho, x)]$	0.079	0.199	0.200	0.200	0.200
$\mathbb{E}_3[L_\tau(\rho, x)]$	0.100	0.250	0.250	0.250	0.250
		$\tau = 10$			
$\mathbb{E}[L_\tau(\rho, x)]$	0.065	0.120	0.123	0.128	0.169
$\mathbb{E}_1[L_\tau(\rho, x)]$	0.065	0.106	0.106	0.106	0.106
$\mathbb{E}_2[L_\tau(\rho, x)]$	0.063	0.100	0.100	0.100	0.100
$\mathbb{E}_3[L_\tau(\rho, x)]$	0.100	0.111	0.111	0.111	0.111
		$\tau = 25$			
$\mathbb{E}[L_\tau(\rho, x)]$	0.038	0.043	0.043	0.043	0.043
$\mathbb{E}_1[L_\tau(\rho, x)]$	0.037	0.041	0.041	0.041	0.041
$\mathbb{E}_2[L_\tau(\rho, x)]$	0.037	0.040	0.040	0.040	0.040
$\mathbb{E}_3[L_\tau(\rho, x)]$	0.042	0.042	0.042	0.042	0.042

All calculations were made at $x = \frac{1}{2}$, where $\mathbb{E}[L_\tau(\rho, x)]$ can be calculated numerically. Note that ρ can be thought of either in terms of the (scaled) genetic length of the fragment or as a physical length (given assumptions about the recombination rate per base pair).

hold for $\mathbb{E}_2[L_\tau(\rho, x)]$, but instead of Equation 5 we have $\lim_{\rho\to\infty}\mathbb{E}_2[L_\tau(\rho, x)]/\rho = 1/\tau$.

Approximation 3: A third approximation comes from the expected coalescence time for a linked locus. As is discussed further below, this is $\sim 1 + 1/\rho$ if τ is large. Solving $\tau = 1 + 1/\mathbb{E}[L_\tau(\rho, x)]$ gives the estimate $\mathbb{E}_3[L_\tau(\rho, x)] = 1/(\tau - 1)$, which should be truncated at ρ if greater than ρ.

Table 1 shows how these approximations perform for a range of parameters. It can be seen that $\mathbb{E}[L_\tau(\rho, x)] > \mathbb{E}_1[L_\tau(\rho, x)] > \mathbb{E}_2[L_\tau(\rho, x)]$ (this can be proved for all r and x). $\mathbb{E}_3[L_\tau(\rho, x)]$ works surprisingly well as long as $\rho > 1/(\tau - 1)$. The approximations can be extended to handle both sides of the balanced polymorphism simply by assuming independence of recombination on each side and multiplying by two.

How should we interpret these results? Note that the expected length decreases with τ in Table 1. This is in agreement with Equation 7 and is perfectly intuitive: larger τ means more time for recombination to decrease the size of the *trans*-specific segment. However, we also see that $\mathbb{E}[L_\tau(\rho, x)]$ increases with ρ, in particular for $\tau = 5$. This may seem paradoxical: if we imagine that the recombination rate per base pair is constant, then increasing ρ simply corresponds to looking at a larger section of the genome. The length of the *trans*-specific segment surrounding a balanced polymorphism should not depend on how large a section of the genome we look at (unless, of course, we are looking at a region that is too small to contain the entire *trans*-specific segment, but this hardly explains the difference between $\rho = 100$ and $\rho = 1000$). The reason for the increase is that $L_\tau(\rho, x)$ includes *trans*-specific regions that have nothing to do with the *trans*-specific polymorphism. As noted earlier, there is a small but positive probability that any site is *trans*-specific. The more of the genome we look at, the more of these we will encounter. The intuitive interpretation of Equation 5 is that, for sufficiently large regions, the fraction of the genome that has not coalesced by τ is simply $e^{-\tau}$, which is the probability that a particular site has not coalesced by τ. The case $\tau = 5$, $\rho = 1000$ is approaching this limit: $7.18 \times 10^{-3} \approx e^{-5} = 6.74 \times 10^{-3}$. Thus, in this case, most of the fragments that have not coalesced by τ are not associated with the balanced polymorphism. These fragments may or may not be *trans*-specific (the probability for each fragment is $\sim 2/3$), whereas the fragments that are linked to the balanced polymorphism are almost certain to be *trans*-specific. For $\tau = 5$, the various approximations give a much better idea of the length of the *trans*-specific fragment that is associated with the balanced polymorphism than does the exact calculation. For approximations 1 and 2 this is not surprising, given that they are defined in terms of a region contiguous with the selected site. For larger values of τ, all expectations agree because the probability of noncoalescence that is not due to linkage to the balanced polymorphism is negligible. A slight increase is seen between $\rho = 0.1$ and $\rho = 1$: this is due to the former region being too small to contain the *trans*-specific region with sufficiently high probability.

Simulation results: The process described here can be simulated, for example, using the algorithm described by Nordborg and Innan (2003). One simply simulates two independent realizations of balancing selection for time τ and then merges the states of the two processes and continues the simulation until all fragments have reached their MRCA.

We used simulations to investigate how well our analytical results concerning $\mathbb{E}[L_\tau(\rho, x)]$ predict the actual length of *trans*-specificity. Recall that $L_\tau(\rho, x)$ is the length of the region on one side of the site of selection where two haplotypes belonging to different allelic classes in a single species still have distinct lineages at the time of speciation, τ. To obtain the length that is *trans*-specific in samples, we have to consider $L_\tau(\rho, x)$ on both sides of the polymorphism, $L_\tau(\rho, x)$ in each species, the probability that lineages distinct at speciation actually coalesce in a *trans*-specific manner, and samples >2.

By assuming that the lengths of either side are independent, noting that the lengths in different species *are* independent, and ignoring the final two issues (*i.e.*, we assume that lineages belonging to different allelic classes at speciation will almost always coalesce in a *trans*-specific manner and that samples >2 will have coalesced to 2 long before speciation for the parameters of interest

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

TABLE 2

The performance of $\mathbb{E}[L_\tau(\rho, x)]$ as an approximation for the expected length of *trans*-specificity

	$\rho =$		
	0.1	1	10
	$\tau = 5$		
Approximation	0.099	0.633	1.178
$n = 4$	0.096 (0.009)	0.563 (0.219)	0.983 (0.516)
$n = 8$	0.096 (0.010)	0.595 (0.214)	1.190 (0.603)
$n = 12$	0.096 (0.010)	0.603 (0.211)	1.276 (0.642)
$n = 16$	0.096 (0.010)	0.610 (0.215)	1.336 (0.668)
$n = 20$	0.096 (0.010)	0.614 (0.215)	1.361 (0.679)
	$\tau = 10$		
Approximation	0.095	0.349	0.378
$n = 4$	0.092 (0.014)	0.306 (0.162)	0.330 (0.188)
$n = 8$	0.091 (0.015)	0.322 (0.170)	0.358 (0.206)
$n = 12$	0.091 (0.014)	0.321 (0.167)	0.365 (0.204)
$n = 16$	0.091 (0.014)	0.327 (0.165)	0.373 (0.211)
$n = 20$	0.091 (0.015)	0.331 (0.169)	0.370 (0.205)
	$\tau = 25$		
Approximation	0.079	0.129	0.129
$n = 4$	0.075 (0.021)	0.103 (0.070)	0.106 (0.068)
$n = 8$	0.073 (0.021)	0.102 (0.070)	0.109 (0.070)
$n = 12$	0.073 (0.021)	0.103 (0.069)	0.108 (0.070)
$n = 16$	0.074 (0.021)	0.104 (0.071)	0.110 (0.071)
$n = 20$	0.073 (0.021)	0.104 (0.071)	0.110 (0.071)

All calculations were made at $x = \frac{1}{2}$, where approximation (9) can be calculated numerically. Mutation between the two allelic classes was symmetric at rate 0.01 (except for $\rho = 0.1$, when a rate of 0.001 was used). Samples were evenly distributed among species and allelic classes; *i.e.*, $n = 16$ means four in each allelic class in each of two species. Estimates are based on 5000 simulations; numbers in parentheses are standard deviations.

here), we obtain the following approximation for the expected length of *trans*-specificity in a region of length ρ surrounding a balanced polymorphism in a pair of species:

$$\rho - 2\int_0^{\rho/2} (1 - P_\tau(u, x))^2 du. \qquad (9)$$

Table 2 illustrates the performance of this approximation for various parameter values and sample sizes. In agreement with the argument just given, the expected length of *trans*-specificity increases only weakly with sample size. In general, the approximation is quite good, although it overestimates the length slightly. Whether this is due to nonindependence between the two sides or due to some distinct lineages not coalescing in a *trans*-specific manner is not clear.

The pattern around a particular site: The results in Table 2 are averages over thousands of realizations. While these results are helpful in understanding the behavior of the process, data are likely to come from a single locus or a small number of loci. Expected values are not sufficient to interpret such data. By studying individual realizations of the process, we can get some idea of how variable it is and what real data might look like. Figure 1 summarizes the results of a single realization that used the human-chimp parameters, by plotting the time to the MRCA along a 10-kb region. The different plots show the coalescence time for different samples. Note that all members of the same allelic class that were sampled within the same species typically coalesce much more recently than speciation. This behavior should be contrasted with samples that include members of different allelic classes: regions closely linked to the balanced polymorphism typically coalesce much further back in time, leading to regional *trans*-specificity. Members of the same allelic class sampled from different species can of course coalesce only in the ancestral species, but they do so much faster than do members of different allelic classes. Note that the pattern is highly variable and that it is sometimes possible for members of the same allelic class to have a MRCA that is older than speciation (Figure 1, center). As discussed above, lineages that are older than speciation need not be *trans*-specific.

Figure 2 shows the time to MRCA for within-species samples in three more realizations. The expected length of the *trans*-specific region is on the order of 0.5–1 kb for these parameters. Note that *trans*-specific regions may often be disjoint from the region surrounding the balanced polymorphism. These additional regions are nonetheless caused by linkage to the balanced polymorphism: as we have discussed, *trans*-specificity in the absence of balancing selection is highly unlikely. The behavior in the absence of balancing selection is completely different, as is illustrated in Figure 3.

In summary, the four realizations shown in Figures 1 and 2 illustrate the enormous variability of the process and thus the danger of relying on expected values when analyzing data. While the genealogy surrounding a *trans*-specific polymorphism is in general expected to be quite different from what is expected in the absence of balancing selection (Figure 3), the variability between different *trans*-specific cases is striking. Not only does the length and genealogical depth of the *trans*-specific region vary between realizations, but also it is the case that *trans*-specific regions may not be centered on, or even contain, the site under selection. Furthermore, peaks of polymorphism may sometimes occur within allelic classes in a single species.

Two selected loci: Our model can easily be extended to two or more selected sites, using the approach described in Nordborg and Innan (2003). This is of relevance because balancing selection may well act to maintain complex alleles that are distinguished by more than a single functionally important mutation (the MHC is a case in point). While it is perfectly possible under this model to obtain analytical results analogous to those presented for the single-locus model, they are too com-

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

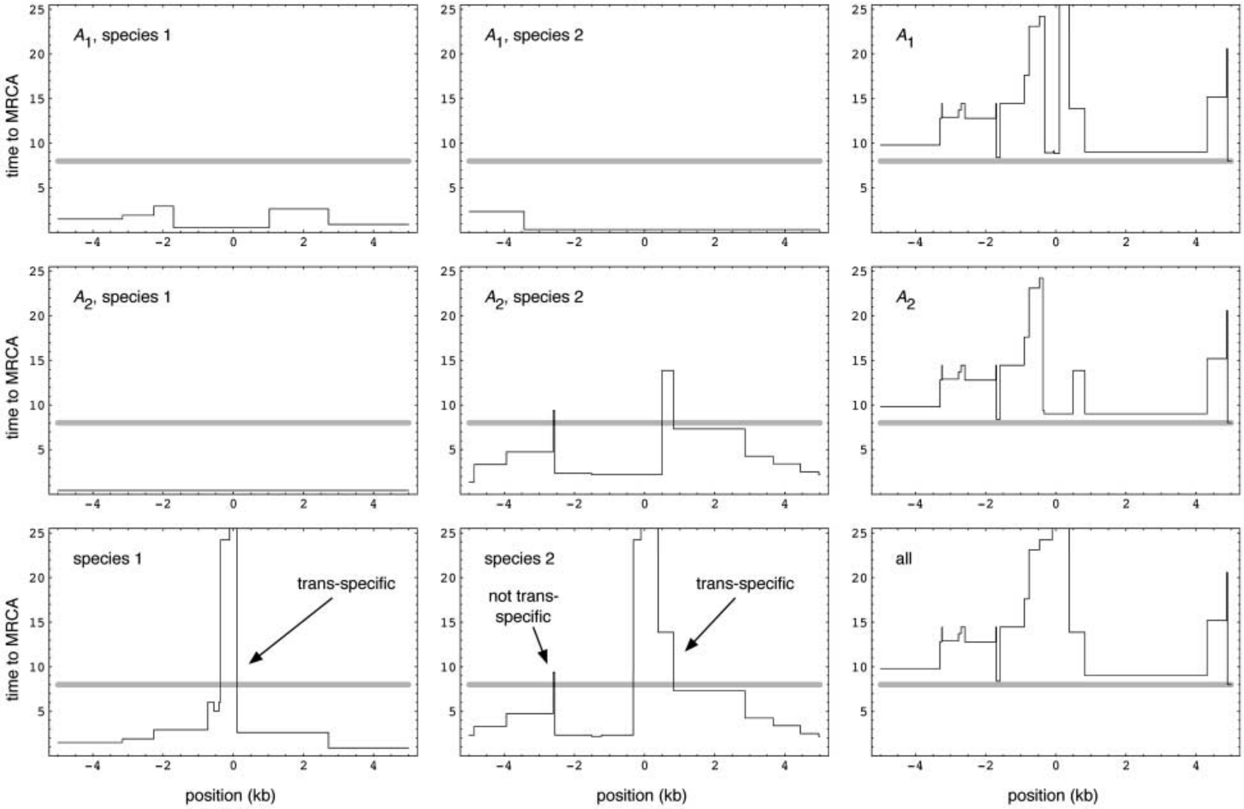

FIGURE 1.—Summary of a single realization of the single-locus *trans*-specific balanced polymorphism. Five lineages were sampled in each allelic class in each species, for a total of 20. Each plot shows the time to MRCA for a subset of the sample. The selected locus was located in the center of the region. Mutation at this locus was symmetric at rate 0.01. Other parameters used were $\rho = 5$ (note that in this and subsequent figures, ρ refers to the total length of the region containing the selected site or sites), $x = 0.5$, and $\tau = 8$ (marked by a shaded line).

plicated to be useful except in very special cases. In particular, because coalescence between allelic classes in the two-locus model must often involve more than a single recombination event (*e.g.*, for a site located between the selected loci sampled in A_1B_1 and A_2B_2), the simple approximations used above do not apply.

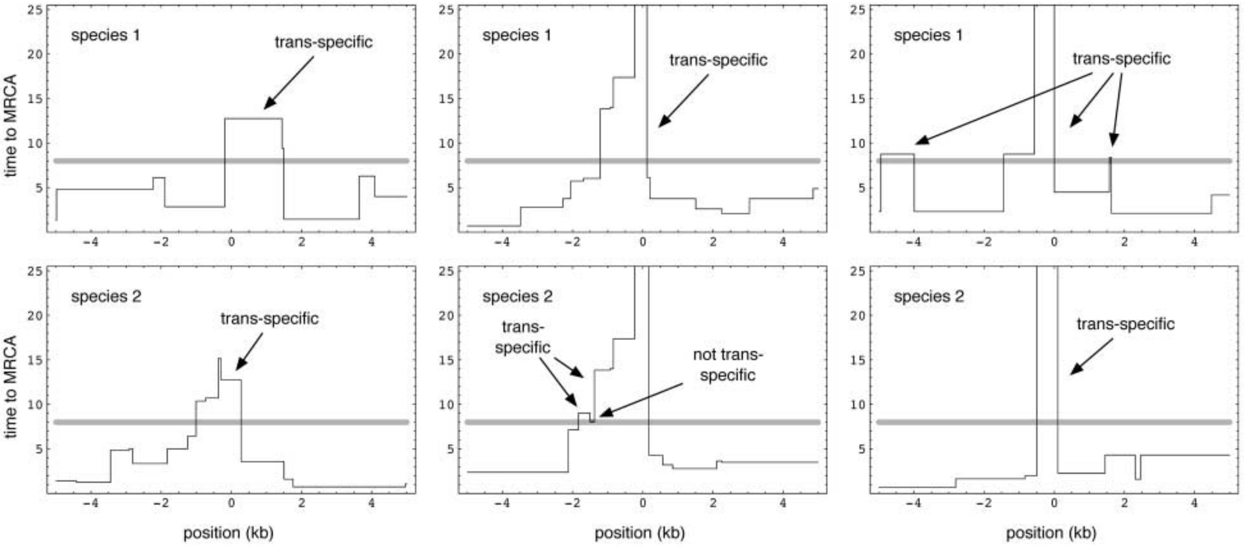

FIGURE 2.—Three further examples of the process used in Figure 1. The parameters are the same, and each column represents one realization.

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

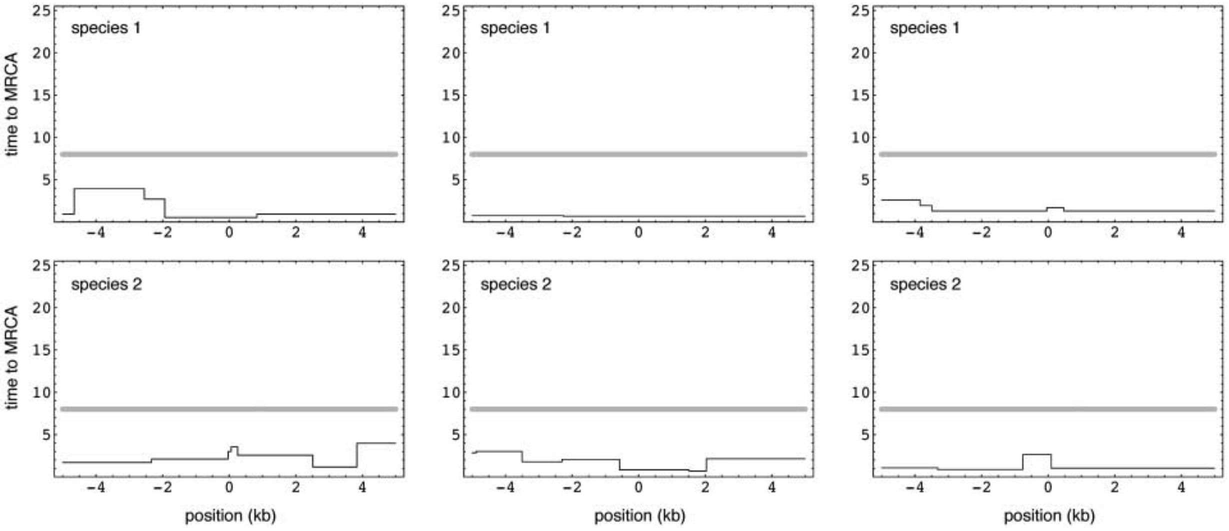

FIGURE 3.—Three examples of the process used in Figures 1 and 2, but without balancing selection. Instead, 10 neutral lineages were sampled from each species.

Because of this, and also due to space limitations, we content ourselves with showing simulation results that illustrate the main points. Figure 4 shows a straightforward extension of the other examples to two loci. As we would expect, there are now regions of *trans*-specificity around each selected site. In addition, the variance in time to MRCA in the general region has clearly increased due to the very complex history of recombination among the four haplotypic classes.

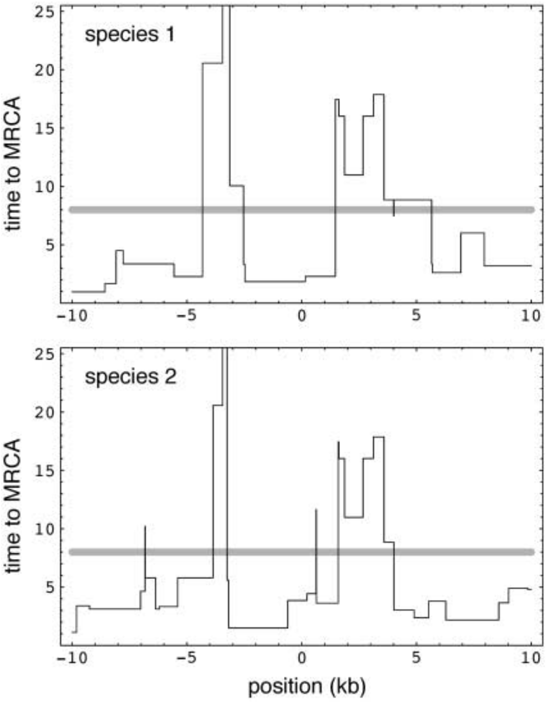

FIGURE 4.—A realization of a model of two-locus balanced polymorphism. The selected sites were located one-third and two-thirds of the distance from the left side of the plot, respectively (*i.e.*, at ±3.33 kb). The frequencies of all four haplotypic classes were taken to be the same, 0.25, and $\rho = 10$ for the region. Two lineages were sampled in each of the four classes, for a total of eight lineages per species. In this plot, all regions with a MRCA older than speciation were *trans*-specific.

Figure 5 illustrates what happens when the sites are closer to each other. In this case, there are two sites in a 10-kb region, rather than one (as in the other cases). Note that there is a tendency for much of the region between the two loci to be *trans*-specific. Clearly, exten-

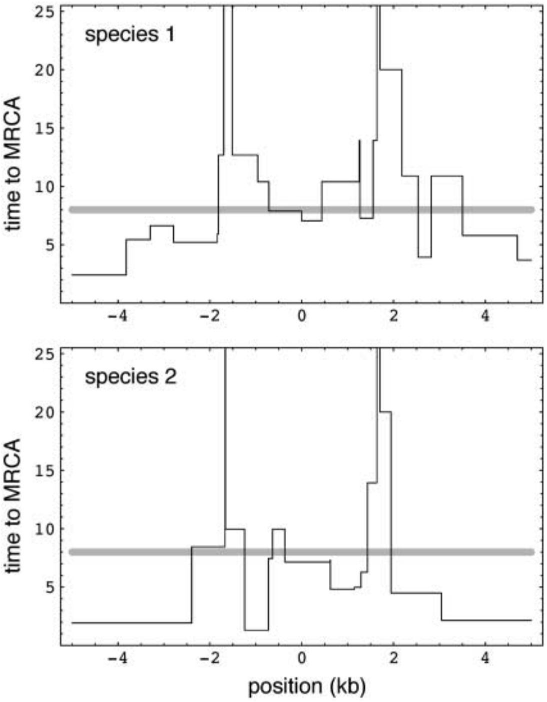

FIGURE 5.—Another realization of the model used in Figure 4, but with $\rho = 5$. Again, all regions with a MRCA older than speciation were *trans*-specific.

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

TABLE 3

The probability that *trans*-specificity is apparent using phylogenetic methods

	ρ =				
	0.1	1	2	5	10
	Finite sites model				
θ = 0.1	0.796	0.369	0.224	NA	NA
θ = 1	0.832	0.414	0.164	0.017	0.002
θ = 10	0.833	0.397	0.081	0.000	0.000
	Finite synonymous sites model				
θ = 0.1	0.801	0.400	0.194	NA	NA
θ = 1	0.839	0.365	0.148	0.008	0.001
θ = 10	0.840	0.264	0.026	0.000	0.000

Power was estimated using 1000 replicates for each parameter combination. The selected polymorphism was located in the center of the region, and mutation between allelic classes occurred with symmetric rate 0.01. Other parameters used were $x = 0.5$, $\tau = 8$, and $n = 4$ (*cf.* Table 2). Neutral mutations were added according to a simple Jukes-Cantor substitution model (Jukes and Cantor 1969) using a method similar to the one used by Schierup and Hein (2000). The simulated region was assumed to correspond to 1 kb, but in the "finite synonymous site" model, only one-third of all sites were allowed to vary. Mutations were added according to the rates in the table, but data sets with less than three segregating sites were not used as they provide too little information for phylogenetic methods. We used PHYLIP for phylogenetic reconstruction. Each data set was run through seqboot (to generate 1000 bootstrap data sets), dnadist, neighbor, and finally consense to generate a consensus tree. If the final tree showed a *trans*-specific topology with a bootstrap support of >70% for the terminal branches, the data set was said to support *trans*-specificity.

sive *trans*-specificity is expected in a region where multiple closely linked sites are subject to balancing selection. It should be noted that this in no way relies on epistatic interaction between the selected sites.

DETECTING *TRANS*-SPECIFICITY

We have shown the extent of *trans*-specificity around a *trans*-specific polymorphism maintained by balancing selection is likely to be quite small and therefore probably difficult to detect. To explore this further, we considered the power of simple phylogenetic methods to detect *trans*-specific polymorphism. We simulated large numbers of data sets using the model above and then used a simple phylogenetic reconstruction method to determine what fraction of these data sets supported a *trans*-specific relationship.

Table 3 shows the results of this study. It is clear that power decreases rapidly with recombination, as would be expected. It also decreases with increased mutation rate. This may seem counterintuitive given that more polymorphism should provide more information about the underlying genealogy. However, more mutation also means increased probability of repeat mutation, *i.e.*, more noise from the point of view of phylogenetic reconstruction.

The region used in our study was too small to explore the effects of using different window sizes when searching for *trans*-specificity. This will clearly influence power: the window size used must be large enough to obtain statistical significance, yet not so large as to drown out any unusual pattern in the surrounding neutral "noise." The problem is analogous to detecting balancing selection within species (Nordborg and Innan 2003).

DISCUSSION

We have described how the structured ancestral recombination graph (Nordborg and Innan 2003) can be used to model *trans*-specific polymorphism. We show that *trans*-specific balancing selection will lead to a distinctive (and highly complex) local distortion in the genealogical graph, but that the extent of the region affected is expected to be quite small. The main implication of these results is that we should not necessarily expect to be able to detect *trans*-specific polymorphism by simply applying phylogenetic tree-building algorithms to genes or parts of genes: the *trans*-specific region may be too small. To the problem of the size of the region should be added that any region that is in fact *trans*-specific is likely to have a very distant MRCA indeed. The time axis in the figures used in this article was cut off at 25 to show detail: in most cases, the time to the MRCA for the *trans*-specific region was several hundred. With such distances, repeated mutation events may start to interfere with the phylogenetic signal by causing homoplasy. It should also be noted that gene conversion may well make the affected region even smaller (Andolfatto and Nordborg 1998; Wiuf and Hein 2000). Finally, it cannot be emphasized enough that the model of recombination used is simplistic and unlikely to be accurate for the short chromosomal distances that are relevant here (Nordborg 2000). In terms of robustness of our predictions, uncertainty about the local recombination process is likely to be far more important than uncertainty about speciation and selection.

Nonetheless, the basic conclusion that regions of *trans*-specificity are likely to be quite short seems hard to avoid. Suitable data for testing our predictions are available in primates, for the ABO system (Saitou and Yamamoto 1997) and for color vision genes (Shyue *et al.* 1995, 1998; Boissinot *et al.* 1998). Table 4 shows our rough estimates of the extent of *trans*-specificity surrounding putatively selected sites in these data. The extent of *trans*-specificity, if that is what it is, seems to be at most a few hundred base pairs.

In summary, we should not expect *trans*-specific balanced polymorphism to be easy to detect, at least not by looking simply for *trans*-specific regions. As a consequence, failure to detect such regions should not be

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024

TABLE 4

The extent of *trans*-specificity in data

Selected codon	Direction	L_{min}	L_{max}
ABO: human/chimp			
156 (010)	3′	0	46
Opsin: saki/howler			
180 (010)	5′	0	31
229 (100)–233 (100)	5′	0	12
	3′	0	54
277 (010)–285 (100)	5′	0	93
	3′	0	199
Opsin: capuchin/squirrel			
180 (010)	5′	0	97
	3′	>20	?
229 (100)–233 (100)	5′	0	12
	3′	54	130
275 (100)–276 (101)–277 (010)–285 (100)	5′	22	181
Opsin: tamarin/marmoset			
180 (010)	5′	22	181
229 (100)–233 (100)	5′	2	67
	3′	0	62
275 (100)–285 (010)	5′	0	181

The position of the *trans*-specific site in each selected codon is given by the three numbers in parentheses: 1 represents the *trans*-specific site. L_{min} is the distance, in base pairs, to the last putatively nonselected *trans*-specific polymorphism from the selected sites in the given direction. L_{max} is similarly the distance to the first fixed difference between the species. Data are from Saitou and Yamamoto (1997), Shyue *et al.* (1995, 1998), and Boissinot *et al.* (1998).

taken as evidence against *trans*-specific balancing selection. On the other hand, when *trans*-specific regions are detected, there is good reason to suspect balancing selection and possibly also a local reduction in the rate of recombination or multiple selected sites. Both will tend to increase the length of *trans*-specific regions.

We thank N. Rosenberg and two anonymous reviewers for comments on an earlier version of this article. Much of this work was done when C.W. was visiting the University of Southern California Center for Computational and Experimental Genomics.

LITERATURE CITED

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397–1399.

Boissinot, S., Y. Tan, S.-K. Shyue, H. Schneider, I. Sampaio *et al.*, 1998 Origins and antiquity of X-linked triallelic color vision systems in New World monkeys. Proc. Natl. Acad. Sci. USA **95:** 13749–13754.

Chen, F.-C., and W.-H. Li, 2001 Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68:** 444–456.

Figueroa, F., E. Günter and J. Klein, 1988 MHC polymorphism pre-dating speciation. Nature **335:** 265–267.

Hey, J., 1994 Bridging phylogenetics and population genetics with gene tree models, pp. 435–449 in *Molecular Ecology and Evolution: Approaches and Applications*, edited by B. Schierwater, B. Streit, G. P. Wagner and R. DeSalle. Birkhäuser Verlag, Basel, Switzerland.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–43 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics.. Oxford University Press, Oxford.

Hudson, R. R., and J. A. Coyne, 2002 Mathematical consequences of the genealogical species concept. Evolution **56:** 1557–1565.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Innan, H., B. Padhukasahasram and M. Nordborg, 2003 The pattern of polymorphism on human chromosome 21. Genome Res. **13:** 1158–1168.

Ioerger, T. R., A. G. Clark and T.-H. Kao, 1990 Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc. Natl. Acad. Sci. USA **87:** 9732–9735.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Klein, J., 1980 Generation of diversity at MHC loci: implications for T-cell receptor repertoires, pp. 239–253 in *Immunology 80*, edited by M. Fougereau and J. Dausset. Academic Press, London.

Nordborg, M., 2000 Linkage disequilibrium, gene trees, and selfing: an ancestral recombination graph with partial self-fertilization. Genetics **154:** 923–929.

Nordborg, M., 2001 On detecting ancient admixture, pp. 123–136 in *Genes, Fossils, and Behaviour: An Integrated Approach to Human Evolution*, edited by P. Donnelly and R. A. Foley. NATO Science Series, IOS Press, Amsterdam.

Nordborg, M., and H. Innan, 2003 The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics **163:** 1201–1213.

O'hUigin, C., Y. Satta, N. Takahata and J. Klein, 2002 Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. Mol. Biol. Evol. **19:** 1501–1513.

Pamilo, P., and M. Nei, 1988 Relationships between gene trees and species trees. Mol. Biol. Evol. **5:** 568–583.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.

Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.

Rosenberg, N. A., 2002 The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. **61:** 225–247.

Saitou, N., and F.-I. Yamamoto, 1997 Evolution of primate ABO blood group genes and their homologous genes. Mol. Biol. Evol. **14:** 399–411.

Saunders, I. W., S. Tavaré and G. A. Watterson, 1984 On the genealogy of nested subsamples from a haploid population. Adv. Appl. Probab. **16:** 471–491.

Schierup, M. H., and J. Hein, 2000 Consequences of recombination on traditional phylogenetic analysis. Genetics **156:** 879–891.

Shyue, S.-K., D. Hewett-Emmett, H. G. Sperling, D. M. Hunt, J. K. Bowmaker *et al.*, 1995 Adaptive evolution of color vision genes in higher primates. Science **269:** 1265–1267.

Shyue, S.-K., S. Boissinot, H. Schneider, I. Sampaio, M. P. Schneider *et al.*, 1998 Molecular genetics of special tuning in New World monkey color vision. J. Mol. Evol. **46:** 697–702.

Takahata, N., 1989 Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics **122:** 957–966.

Takahata, N., 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species polymorphism. Proc. Natl. Acad. Sci. USA **87:** 2419–2423.

Tan, Y., and W.-H. Li, 1999 Trichromatic vision in prosimians. Nature **402:** 36.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetic models. Theor. Popul. Biol. **26:** 119–164.

Vekemans, X., and M. Slatkin, 1994 Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics **137:** 1157–1165.

Vincek, V., C. O'hUigin, Y. Satta, N. Takahata, P. T. Boag *et al.*, 1997 How large was the founding population of Darwin's finches? Proc. R. Soc. Lond. Ser. B **264:** 111–118.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Communicating editor: M. K. Uyenoyama

Downloaded from https://academic.oup.com/genetics/article/168/4/2363/6059413 by guest on 23 April 2024