Genet. Res., Camb. (2000), **75**, *pp.* 61–73. *With 5 figures.* Printed in the United Kingdom

Distinguishing recombination and intragenic gene conversion by linkage disequilibrium patterns

THOMAS WIEHE[1]*, JOANNA MOUNTAIN[1]**, PETER PARHAM[2] AND MONTGOMERY SLATKIN[1]

[1] *Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA*
[2] *Department of Structural Biology, Stanford University, Stanford, CA 94305, USA*

(*Received 2 April 1999 and in revised form 28 April 1999*)

Summary

Deterministic theory suggests that reciprocal recombination and intragenic, interallelic conversion have different effects on the linkage disequilibrium between a pair of genetic markers. Under a model of reciprocal recombination, the decay rate of linkage disequilibrium depends on the distance between the two markers, while under conversion the decay rate is independent of this distance, provided that conversion tracts are short. A population genetic three-locus model provides a function Q of two-locus linkage disequilibria. Viewed as a random variable, Q is the basis for a test of the relative impact of conversion and recombination. This test requires haplotype frequency data of a sufficiently variable three-locus system. One of the few examples currently available is data from the Human Leukocyte Antigen (HLA) class I genes of three Amerindian populations. We find that conversion may have played a dominant role in shaping haplotype patterns over short stretches of DNA, whereas reciprocal recombination may have played a greater role over longer stretches of DNA. However, in order to draw firm conclusions more independent data are necessary.

1. Introduction

There are different mechanisms which lead to an exchange of alleles at the time of meiosis. Both gene conversion and cross-over are often referred to as recombination and terminology is sometimes not clearly defined. Here we use the term 'recombination' to refer to a single, reciprocal crossing-over event (Fig. 1*a*). It is unlikely that two or more of these events occur within one generation within a short stretch of DNA. We define conversion as a process whereby stretches of DNA are replaced, in a non-reciprocal manner, by stretches transferred from another chromosome or chromosomal region. In this paper we are concerned only with intragenic, interallelic, unbiased conversion (Fig. 1*b*), i.e. we consider events in which homologous alleles are substituted without preference for a particular allele. Such a mechanism has been proposed by Parham *et al.* (1995) as the primary, but possibly not the only, mechanism generating the haplotype[1] patterns observed for the HLA class I loci. Similarly, Geliebter & Nathenson (1988) claimed that 'microrecombinations', i.e. gene conversion events, are responsible for the haplotype diversity observed in the murine MHC-K region.

Because the consequences of reciprocal recombination and conversion are detectable only in highly variable regions, the MHC complex provides a rare opportunity for determining the relative roles of these mechanisms in higher eukaryotes. Analogous mechanisms in bacteria are conjugation and transformation. Clark & Zheng (1997) compared the dynamics of linkage disequilibria for these two mechanisms

* Corresponding author. Department of Molecular Genetics and Evolution, Max Planck Institute for Chemical Ecology, Tatzendpromenade 1, D-07745 Jena, Germany. Tel: +49-3641-64 3631. Fax: +49-3641-64 3668. e-mail: twiehe@ice.mpg.de.
** Current address: Dept. of Anthropological Sciences and Genetics, Stanford University, Stanford, CA 94305-2145, USA.

[1] Here, the term 'haplotype' refers not only to the HLA-A, -B, -C haplotypes, but also, more generally, to the allelic composition of two or more, possibly very short, DNA segments (see Fig. 1) on a chromosome.

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

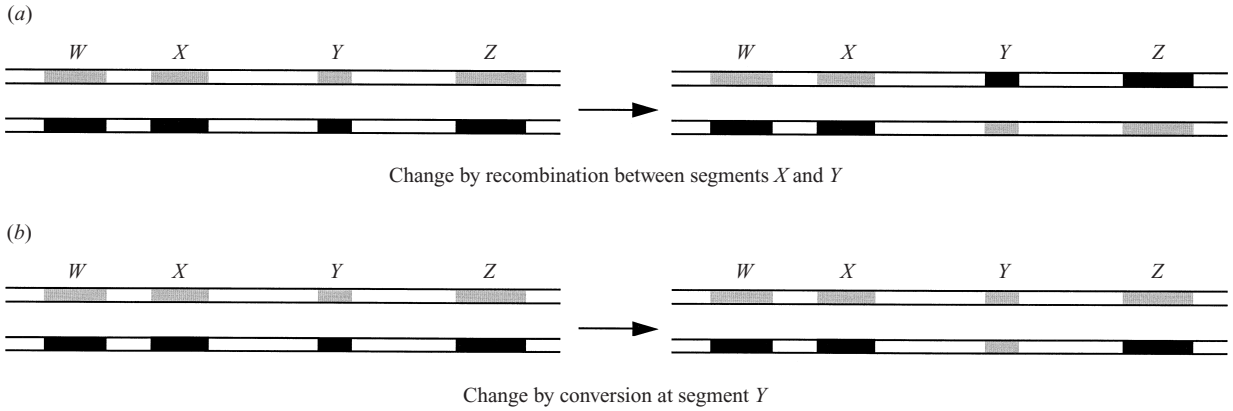

Fig. 1. Diagram of the contrasting effects of a single reciprocal recombination event and a single intragenic, interallelic conversion event. (*a*) Reciprocal recombination exchanges a series of homologous segments in two chromosomes downstream (to the right) of the recombination point (here, between segments $\mathscr{X}$ and $\mathscr{Y}$). (*b*) Conversion, on the other hand, depending on the length of the conversion tract, may alter single segments only (here, segment $\mathscr{Y}$).

with the help of a population genetical three-locus model.

Several statistical approaches to the detection of conversion have been suggested in the past. These generally involve either the examination of the distribution of nucleotide substitutions along the DNA sequence (Sawyer, 1989; Stephens, 1985), or the comparison of phylogenetic trees inferred from different segments within a DNA region (Gyllensten *et al*., 1991). Takahata (1994) found that, while in some cases significant clustering of nucleotide site configurations across sequences could be detected, such clustering does not imply that conversion has taken place. Weiller (1998) describes a graphical method to detect putative recombination sites in sets of homologous sequences. Betran *et al*. (1997) have suggested a method of estimating the number and length distribution of conversion tracts from DNA sequence data. These authors assume that conversion has taken place and then estimate parameters of their conversion model. We are interested, however, in the relative roles of conversion and reciprocal recombination. To this end we develop a maximum likelihood estimator and a statistical test which are based on a function of linkage disequilibria between pairs of short DNA segments. Both statistics require haplotype frequency data, typically obtained from a population sample. We apply the statistics to HLA sequence data for three Amerindian populations of North and South America (Markow *et al*., 1993; Petzl-Erler *et al*., 1993; Parham *et al*., 1997), and to the haplotype frequency data gathered by Tishkoff *et al*. (1998).

When examining DNA sequence data, the general importance of gene conversion as a recombinational mechanism is currently unknown, yet many applications of coalescent theory to sequence data assume only reciprocal recombination. Our results suggest that conversion also must be taken into account.

2. Theoretical background

(i) *Two-locus model*

Consider first a model of two biallelic loci, $\mathscr{X}$ and $\mathscr{Y}$. Here, the terms 'locus' or 'segment' both refer to a short stretch of DNA, typically a few base pairs in length. While the dynamical system can be described in terms of the change over time in haplotype frequencies, $XY(t)$, $Xy(t)$, $xY(t)$ and $xy(t)$, the system can be described equivalently in terms of allele frequencies $X(t)$, $Y(t)$ and linkage disequilibrium $D_{\mathscr{X}\mathscr{Y}}(t) = XY(t)\,xy(t) - Xy(t)\,xY(t)$. The argument t denotes time in generations. Italic letters X, x, etc., denote allelic types; pairs of letters (XY) denote haplotypes. When used in arithmetical expressions, they mean allele frequencies and haplotype frequencies, respectively. For simplicity, we assume that the rate of reciprocal recombination r is a linear function of the distance ν_1 (in base pairs) between loci $\mathscr{X}$ and $\mathscr{Y}$: $r = \rho\nu_1$, where ρ is the recombination rate per base pair. When there is no selection, allele frequencies do not change with time and linkage disequilibrium decays exponentially with rate r. In the case of conversion, we assume that either locus undergoes conversion with rate d. Here, we concentrate on the case of short tract lengths and require the tract length, ν', to be smaller than ν_1 (cf. Fig. 2).

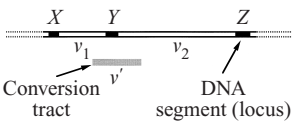

Fig. 2. The theory developed in this paper deals with conversion of short tracts of DNA: the conversion tract does not extend over more than one segment or locus. Distances between segments (ν_1 and ν_2) are larger than the conversion tract length (ν'). For differentiating between reciprocal recombination and conversion using linkage disequilibria between the segments, ν_1 and ν_2 must be different (without loss of generality, we assume $\nu_2 > \nu_1$).

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

Table 1. *Decay rates of linkage disequilibrium*

Type of disequilibrium	Decay rate of disequilibrium				
		Conversion[b]			
	Recombination[a]	$\nu' < \nu_1 < \nu_2$	$\nu_1 < \nu' < \nu_2$	$\nu_1 < \nu_2 < \nu'$ and $\nu' < \nu_1 + \nu_2$	$\nu_1 + \nu_2 < \nu'$
$D_{\mathcal{XY}}$	ν_1	ν'	ν_1	ν_1	ν_1
$D_{\mathcal{YZ}}$	ν_2	ν'	ν'	ν_2	ν_2
$D_{\mathcal{XZ}}$	$\nu_1 + \nu_2$	ν'	ν'	ν'	$\nu_1 + \nu_2$
$D_{\mathcal{XYZ}}$	$\nu_1 + \nu_2$	$\frac{3\nu'}{2}$	$\frac{2\nu' + \nu_1}{2}$	$\frac{\nu_1 + \nu_2 + \nu'}{2}$	$\nu_1 + \nu_2$

[a] Values have to multiplied by the per base pair reciprocal recombination rate ρ.
[b] Values have to be multiplied by the per base pair conversion rate δ; for simplicity, it is here assumed that $l = 1$ and therefore $d = \delta\nu'$ (cf. Appendix A.1).
The case of short conversion tracts, treated in Section 2, corresponds to the first column in the block 'conversion'.

Under unbiased conversion (either allele is converted with the same probability) and in a deterministic model allele frequencies remain constant over time as well. Linkage disequilibrium decays with rate d (Clark & Zheng, 1997; and see Appendix A.1).

Furthermore, we use the normalized linkage disequilibrium D' (Lewontin, 1964) instead of the usual linkage disequilibrium $D = (XY)(xy) - (Xy)(xY)$. The normalized linkage disequilibrium is

$$D' = \begin{cases} \dfrac{D}{\min(X(1-Y), (1-X)Y)}, & \text{if } D > 0 \\ \dfrac{D}{\max(-XY, -(1-X)(1-Y))}, & \text{if } D < 0 \\ 0, & \text{if } D = 0 \\ & \text{and } \min(X(1-Y), (1-X)Y) \neq 0 \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

As is true for D, the normalized linkage disequilibrium D' also decays exponentially:

$$D'_{\mathcal{XY}}(t) = D'_{\mathcal{XY}}(0) \exp(-(.)\, t),$$

where $(.)$ has to be replaced by r or d. Note that $D'(t)$ – but in general not $D(t)$ – is still (to a good approximation) an exponentially decaying function when overdominant selection is incorporated into the model (see Appendix A.2).

The above arguments are for a deterministic model, but carry over to a stochastic model, such as a diffusion model, at least for the first-order moments of the corresponding expressions (results not shown). The diffusion equation has been treated by Ohta & Kimura (1969). These authors have derived the first and second moments of the distribution of $D_{\mathcal{XY}}$. However, the distribution itself, even for the case without selection, is not known analytically.

(ii) *Three-locus model*

We assume that three loci, $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, are arranged in linear order and separated by distances ν_1 and ν_2, where $\nu_1 < \nu_2$ (see Fig. 2). Recombination rates are $r_1 = \rho\nu_1$ and $r_2 = \rho\nu_2$. The conversion rate for either of the three loci is d. There are four kinds of linkage disequilibria: $D_{\mathcal{XY}}$, $D_{\mathcal{XZ}}$, $D_{\mathcal{YZ}}$ and the higher-order linkage disequilibrium $D_{\mathcal{XYZ}}$, which can be written as

$$D_{\mathcal{XYZ}} = (XYZ) - X*Y*Z - X*D_{\mathcal{YZ}} - Y*D_{\mathcal{XZ}} - Z*D_{\mathcal{XY}}.$$

All linkage disequilibria $D_{(.)}$ decay exponentially as a result of reciprocal recombination or conversion. For recombination the decay rates are $r_1, r_1 + r_2, r_2$ and $r_1 + r_2$, respectively. For conversion, the tract length plays a role again. For the case of short tracts ($\nu' < \nu_1 < \nu_2$) the decay rates are d for the three two-locus disequilibria and $(3/2)d$ for the three-locus disequilibrium. Note that, for this case, all three loci are symmetrical in the sense that linkage disequilibrium between any two loci decays at the same rate. Allowing for longer tracts, so that a single conversion event may affect more than one locus, would lead to decay rates which resemble more those under recombination (Table 1). In the following, we deal only with first-order linkage disequilibria. Their normalized versions satisfy

$$D'_{\ldots}(t) = D'_{\ldots}(0) \exp(-(.)\, t), \quad (2)$$

with $(.)$ replaced by $r_1, r_2, r_1 + r_2$ or d and $\ldots$ replaced accordingly by $\mathcal{XY}$, $\mathcal{XZ}$ or $\mathcal{YZ}$. To determine the initial conditions, $D'_{\ldots}(0)$, assume there is a polymorphic locus, say $\mathcal{X}$. At time $t = 0$ a new allele at a second locus, say $\mathcal{Y}$, is introduced. At time $t = 0$ there are at most three of the possible four haplotypes present. Missing haplotypes are eventually created by

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

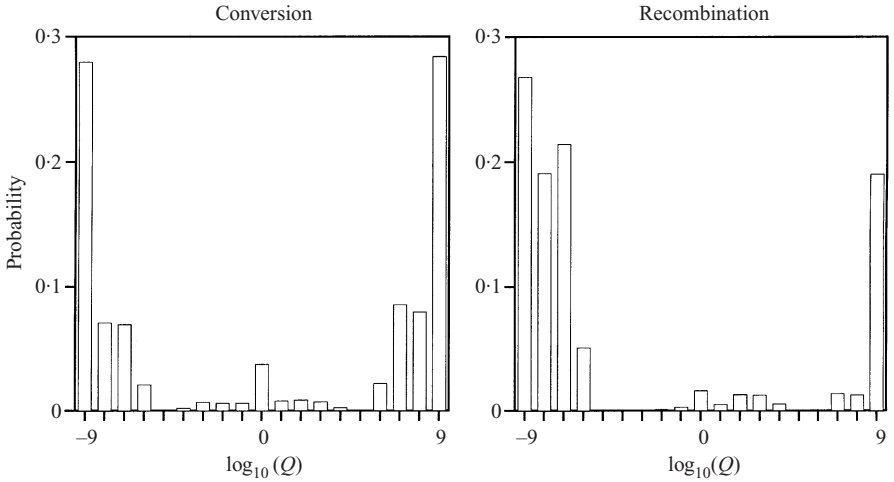

Fig. 3. Distributions of the random variable $Q(t_1)$ under conversion (left-hand panel) and reciprocal recombination (right-hand panel) based on simulations and about 4000 replicates. The log 10-transformed distribution is symmetrical in the first case and more heavily weighted towards negative values in the second case. Q is calculated as described in the text and is plotted here for the time parameter $t_1 = 250$ generations after start. Parameters: $N = 10^4$, $\rho = 5 \times 10^{-8}$, $\nu_1 = 10$, $\nu_2 = 100$, $d = 2 \times 10^{-6}$, $s = 0$ (no selection).

recombination or conversion. According to definition (1), $D' = 1$ if at least one of the four haplotypes is missing. Therefore, $D'_{\mathcal{XY}}(0) = 1$, and, taking the logarithm on both sides of (2), one obtains

$$\log(D'_{\ldots}(t)) = -(.)\,t.$$

Define the ratio

$$Q = \frac{\log(D'_{\mathcal{XY}}(t))}{\log(D'_{\mathcal{XZ}}(t))} = \frac{-(.)_{\mathcal{XY}}\,t}{-(.)_{\mathcal{XZ}}\,t}. \tag{3}$$

Replacing $(.)$ by either the appropriate recombination or conversion rate the ratio becomes

$$Q = \begin{cases} \dfrac{\nu_1}{\nu_1+\nu_2} & \text{(reciprocal recombination)} \\ 1 & \text{(conversion).} \end{cases}$$

Note that Q does not contain d, ρ, the initial conditions or time as explicit parameters. Letting $\nu_2 = h\nu_1$ and assuming $\nu_2 \gg \nu_1$, then, with reciprocal recombination, $Q \approx 1/h$. Thus, the relative distances between the three loci rather than the absolute distances matter in distinguishing recombination from conversion: the (deterministic) Q-values for the two cases differ by as many orders of magnitude as do ν_1 and ν_2. However, in practice, these Q-values will almost never be observed. Finite population and sample sizes introduce random effects. In fact, Q is a compound random variable which depends in a nonlinear manner on other random variables, the haplotype frequencies. Strictly, the validity of (3) is not even established for the expected value $E(Q)$ of Q. Still, some distinguishing features of the distribution of Q for the two cases, conversion and reciprocal recombination, can be characterized: under the model of conversion all involved mechanisms are symmetrical for the three loci. Accordingly, $D'_{\mathcal{XY}}$ and $D'_{\mathcal{XZ}}$ are identically distributed. This implies that Q and $1/Q$ are identically distributed. Therefore, the distribution of $\log Q$ (or $\log_{10} Q$) is symmetrical around 0: $\text{Prob}(\log Q < -k) = \text{Prob}(\log Q > k)$, for all k. In particular, conditioning on $\log Q \neq 0$ one has for the case of conversion

$$q = \text{Prob}(\log Q < 0 \,|\, \log Q \neq 0) = \text{Prob}(\log Q > 0 \,|\, \log Q \neq 0) = 1/2,$$

independently of the numerical value of any involved parameter. On the other hand, for the case of recombination one can only state

$$q = q(N) = \text{Prob}(\log Q < 0 \,|\, \log Q \neq 0) > 1/2 > \text{Prob}(\log Q > 0 \,|\, \log Q \neq 0),$$

where the numerical value of q depends not only on N but also on r_1, r_2 and t. However, for the limit we have

$$\lim_{N\to\infty} q(N) = q = 1.$$

Generally, neither the distribution of Q itself nor its moments are available analytically. We carried out simulations to determine the distribution of Q for a range of parameters. We give an example in Fig. 3.

(iii) *Simulations*

On the basis of a biallelic three-locus model of constant diploid population size N, we advanced the haplotype frequencies in each generation according to a two-step mechanism. We first incorporated the change caused by reciprocal recombination or conversion and selection as a deterministic step. The

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

Table 2. *Simulation results*

		N		ρ		t_1	
k	$\bar{q}$ (SE)	$\bar{q}-$	$\bar{q}+$	$\bar{q}-$	$\bar{q}+$	$\bar{q}-$	$\bar{q}+$
0	0·4900 (0·0048)	0·5179	0·4860	0·4886	0·4862	0·4925	0·4920
50	0·6972 (0·0045)	0·6014	0·7483	0·6728	0·6923	0·7045	0·7112
100	0·8285 (0·0043)	0·6921	0·8648	0·7509	0·8699	0·8273	0·8222
0	0·4952 (0·0049)	0·5168	0·4836	0·4855	0·5040	0·4931	0·4946
50	0·7320 (0·0045)	0·6595	0·7235	0·7222	0·7179	0·7388	0·7327
100	0·9605 (0·0019)	0·8869	0·9883	0·9415	0·9168	0·9726	0·9205

Upper block: short-range case. Lower block: long-range case. Column k: percentage of recombination. Column $\bar{q}$: Values obtained with parameter settings $N = 10^4$, $\rho = 10^{-8}$, $d = 4 \times 10^{-6}$ and $t_1 = 250$. Columns $\bar{q}-$ and $\bar{q}+$ represent $\bar{q}$ values from a series of simulations for which the parameter choices $N = 10^4$, $\rho = 10^{-8}$, and $t_1 = 250$ are individually varied. Columns N: population sizes $N = 10^3$ (column $\bar{q}-$) and $N = 10^5$ (column $\bar{q}+$). Columns ρ: recombination rates $\rho = 2 \times 10^{-9}$ (column $\bar{q}-$) and $\rho = 5 \times 10^{-8}$ (column $\bar{q}+$). Columns t_1: recording time at $t_1 = 125$ (column $\bar{q}-$) and $t_1 = 500$ (column $\bar{q}+$) generations. Each $\bar{q}$ is based on at least 1000 (depending on parameters) simulation runs. For each run the outcome $\tilde{Q}$ is recorded. The fraction of times when $\tilde{Q} = -1$ is q. The table contains $\bar{q}$ (SE), the average (standard error) obtained by resampling (100 samples of size 100 each).

recombination pathway is taken with probability c and the conversion pathway with probability $1-c$. Therefore, the effective recombination and conversion rates for the region comprising segments $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{Z}$ are $c(r_1+r_2)$ and $(1-c)\,3d$, respectively. Apart from the neutral model, we considered separately two selection models: (1) overdominance at one locus and (2) epistasis for the two flanking loci. In the second step, we simulated the change due to random drift by multinomial sampling with replacement (Wright–Fisher model). From the simulated haplotype frequencies we calculated the ratio Q as described in the previous section, and recorded its distribution at a fixed time t_1 generations after start. The initial distribution (at $t = 0$) for the total of eight haplotypes is chosen randomly, subject to the condition (1 a) that all loci are polymorphic and (1 b) that for both pairs of loci, $\mathscr{X}\mathscr{Y}$ and $\mathscr{X}\mathscr{Z}$, at least one recombinant haplotype is missing, so that $D'_{\mathscr{X}\mathscr{Y}}(0) = D'_{\mathscr{X}\mathscr{Z}}(0) = 1$. Furthermore, the recorded distribution at t_1 is conditional on the requirement that (2) all loci remained polymorphic during the simulation interval (since data analysis is performed only on polymorphic loci). The random variables $D'_{\mathscr{X}\mathscr{Y}}$ and $D'_{\mathscr{X}\mathscr{Z}}$ may take values between 0 and 1. To avoid singularities when calculating the logarithm and the ratio we rescaled by

$$D' \to D'(1-2\epsilon)+\epsilon,$$

where ϵ is a small positive number (10^{-10}, say). The simulations suggest that the distribution of Q is independent of t_1 over a wide range of values, but depends more sensitively on N, r_i and d (Table 2). In the remaining sections we concentrate on and make use of the properties of q only.

3. Statistical methods

(i) *A sign test to distinguish conversion and reciprocal recombination*

As discussed by Lewontin (1995) it is often hard to establish the significance of linkage disequilibrium in natural data, in particular when allele frequencies are highly asymmetrical. Instead of testing the significance of the numerical value of linkage disequilibrium he suggested applying a sign test of linkage disequilibrium to the absolute numbers of repulsion versus coupling haplotypes. We are in a very similar situation. In the examined data, we found that allele frequencies were often highly asymmetrical. In addition, it is known that the distribution of the standardized linkage disequilibrium is concentrated at the extreme ends of the domain (cf. Hudson, 1985, fig. 4) for a wide range of recombination rates. Consequently, Q is often also at either extreme of its domain. The resulting high sample variance therefore makes a test of recombination versus conversion which is based on numerical Q-values of the sample, such as the sample mean, obsolete. Rather, even at the expense of power, we suggest the following sign test. Let

$$\tilde{Q} = \begin{cases} 1, & \text{if } \log Q > 0 \\ 0, & \text{if } \log Q = 0 \\ -1, & \text{if } \log Q < 0. \end{cases}$$

Then,

$$\text{Prob}\,(\tilde{Q} = -1 \,|\, \tilde{Q} \neq 0) = q.$$

When trying to reject the hypothesis 'only recombination is acting' we choose the null hypothesis H_0:

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

$q = q'$, where $q' > 1/2$ is a value which has been determined by simulations as described above, and the alternative H_1: $q = 1/2$. When trying to reject the hypothesis 'only conversion is acting' we reverse the roles of H_0 and H_1. Let the sample be $\tilde{Q}_1, \ldots, \tilde{Q}_n$ and assume – without loss of generality – $\tilde{Q}_i \neq 0$ for all i. Let $\hat{x}$ be the number of times with $\tilde{Q}_i = -1, i = 1, \ldots, n$ and put $\hat{q} = \hat{x}/n$. We then test both null hypotheses that $\hat{q}$ is sampled from a population in which $q = q'$ and $q = 1/2$. Critical values for such an one-sided sign test, as well as the power, can be derived from the binomial distributions with parameters n, $q = q'$ and $q = 1/2$, respectively (Sokal & Rohlf, 1981).

(ii) *Maximum likelihood estimates of the proportion of recombination*

Reciprocal recombination and conversion may both have contributed to the observed genetic diversity. If so, the relative contribution of each mechanism to haplotype diversity is of interest. The relationship between the percentage of recombination events (k %) and the value of q, $q = f(k)$, is an increasing function of k. A maximum likelihood estimate $\hat{k}$ is obtained from the sample statistic $\hat{q}$ when f is inverted and evaluated at $\hat{q}$. When determining k one has to take the possibly different rates of recombination and conversion into account and to normalize with respect to the combined rate for either recombination or conversion. This rate is $c(r_1 + r_2) + (1 - c)\,3d$. The first summand is the effective recombination rate and the second summand the effective conversion rate in the region comprising segments $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{Z}$. The percentage of events that are reciprocal recombinations is

$$k = \frac{c(r_1 + r_2)}{c(r_1 + r_2) + (1 - c)\,3d} 100\,\%.$$

The maximum likelihood estimate of the parameter q is the fraction

$$\hat{q} = \frac{\hat{x}}{n}.$$

The maximum likelihood estimate $\hat{k}$ of k is then

$$\hat{k} = \begin{cases} f^{-1}(\hat{q}), & \text{if } f(0) \leqslant \hat{q} \leqslant f(100) \\ 0, & \text{if } \hat{q} < f(0) \\ 100, & \text{if } f(100) < \hat{q}, \end{cases}$$

where f^{-1} is the inverse function of f. Since the function $f(k)$ is not known analytically and rests upon simulations, we used linear interpolation to obtain approximate values for f^{-1}. We focused on two scenarios: neutrality and epistatic selection. In the latter case we assumed that haplotypes XZ and xz at the locus pair $\mathscr{X}\mathscr{Z}$ are favoured over haplotypes Xz and xZ with selection coefficient s. Such selection tends to preserve linkage disequilibrium between $\mathscr{X}$ and $\mathscr{Z}$ through the elimination of recombinants. Thus, the value of q will be lower than is expected under a neutral model (see Fig. 4); in particular q may be smaller than 1/2.

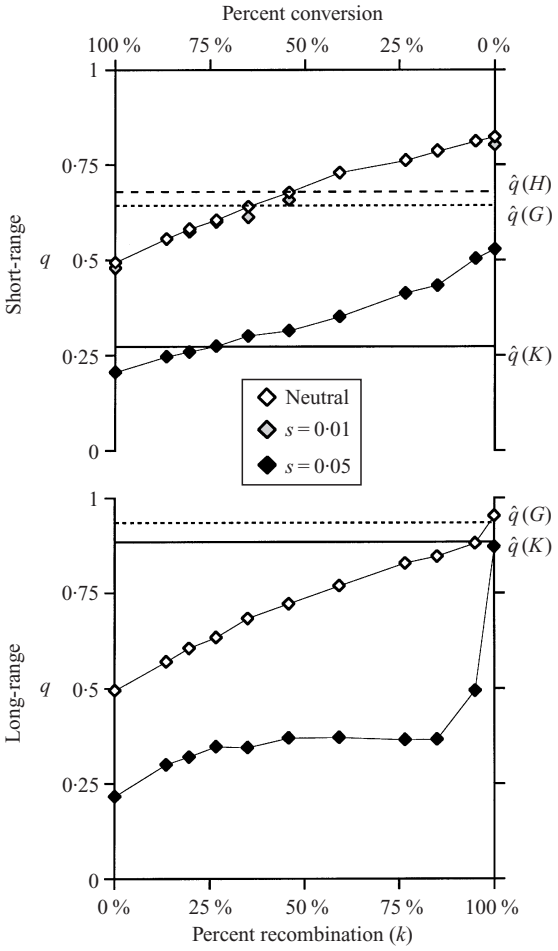

Fig. 4. Simulation results of $q = f(k)$ for a combination of reciprocal recombination (k %) and conversion ($(100-k)$ %). Parameters: $N = 10^4$, $t_1 = 250$, $d = 4 \times 10^{-6}$; $r_1 = 10^{-6}$, $r_2 = 10^{-5}$ (short-range) and $r_1 = 10^{-3}$, $r_2 = 1{\cdot}3 \times 10^{-2}$ (long-range). Only strong epistasis ($s = 0{\cdot}05$, filled diamonds) appreciably changes q with respect to its neutral values (open diamonds). Observed levels of q are printed as horizontal black lines. Continuous line, Kaingang; dotted line, Guarani; dashed line, Havasupai.

4. Applications

(i) *Haplotype data from HLA class I loci*

It had been suggested that interallelic conversion plays a major role in generating haplotype variability in the HLA complex (Parham *et al.*, 1995). As pointed out by Hughes (1991), however, the MHC literature is

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

Table 3. *HLA-B allele and HLA B-C-A haplotype frequencies*

Guarani		Havasupai		Kaingang	
Type	Frequency	Type	Frequency	Type	Frequency
Allele frequencies					
B1504	0·281	B27052	0·037	B1520	0·038
B3505	0·065[a]	B3501	0·164	B3501	0·136[a]
B3511	0·144[a]	B39011	0·061	B3505	0·179[a]
B4003	0·079[a]	B40012	0·102	B3506	0·052[a]
B4004	0·303[a]	B4002	0·008	B3903	0·102[a]
B5104	0·066	B4801	0·422	B3905	0·090[a]
—	—	B5101	0·086	B4801	0·081
—	—	B5102	0·119	B5101	0·248
Haplotype frequencies					
A0201C0303B1504	0·277			A31012C0401B3505	0·191
A68012C0304B4004	0·220			A31012C0304B5101	0·152
A31012C1503B5104	0·085			A31012C0401B3501	0·115
A68012C0304B3511	0·078			A31012C0702B3905	0·083
A0211C0304B4003	0·057			A0212C0102B5101	0·076
A0211C0401B3505	0·035			A0212C0702B3903	0·059
A31012C0303B1504	—			A2402C0401B3506	0·056
—	—			A31012C0702B3903	0·051
—	—			A0212C0401B3501	—
—	—			A2402C0304B1520	—
—	—			A31012C0102B5101	—

[a] Allele frequencies have been inferred from haplotype frequencies; cf. fig. 1 and table 2 in Parham *et al.* (1995).

replete with claims of gene conversion, and yet the null hypothesis of no gene conversion is rarely excluded. It may be that other mechanisms, including reciprocal recombination possibly combined with some form of selection, have generated existing diversity. We examined haplotype data of the HLA-A, -B and -C loci from three Amerindian populations: Kaingang (K), Guarani (G) and Havasupai (H). Haplotypes for the entire HLA-B-C-A region as well as for the individual loci have been characterized and frequency data have been compiled (Markow *et al.*, 1993; Petzl-Erler *et al.*, 1993; Parham *et al.*, 1997; see Table 3). For data analysis we consider two scenarios. (1) The three segments $\mathscr{X}$, $\mathscr{Y}$, $\mathscr{Z}$ are chosen one each from the three HLA loci (long-range comparisons between loci) and (2) the three segments are chosen from within a single locus (short-range comparisons within a locus). We searched the available coding sequences of haplotypes for DNA segments which are variable among the characterized and annotated haplotypes. In our effort to be conservative in choosing variable sites, we required that the candidate segments should:

(a) comprise at least two nucleotides (to exclude point mutations as a potential source of the variation),
(b) be polymorphic, but contain not more than two alleles,
(c) not extend over exon/exon boundaries (with respect to the cDNA),
(d) extend at most as far as the polymorphism pattern across the aligned haplotypes does not change.

Some examples of such segments which passed these criteria are shown in Fig. 5 (framed boxes). To apply the model we also required that the triplets of segments $(\mathscr{X}, \mathscr{Y}, \mathscr{Z})$ be unevenly spaced. This requirement is fulfilled *a fortiori* for the long-range comparisons ($\nu_1 = 100$ kb, $\nu_2 = 1300$ kb, $h \approx 13$); for the short-range comparisons we required

(e) the distance between $\mathscr{Y}$ and $\mathscr{Z}$ to be at least 10 times ($h \geqslant 10$, ν_1 is typically on the order of 100 bp or less) as large as the distance between $\mathscr{X}$ and $\mathscr{Y}$.

One example of a triplet which passes all criteria (a) to (e) are the three shaded boxes in Fig. 5. Table 4 lists the final counts of accepted triplets of segments for both scenarios – long-range and short-range comparisons. For the short-range comparisons, only HLA-B yielded admissible triplets; neither HLA-A nor HLA-C contained a triplet of segments which passed all criteria (a)–(e). We used the frequency data (Parham *et al.*, 1997) to obtain linkage disequilibria for the selected triplets of DNA segments. With missing frequency data, we let values range from 0 to 10% (test results depended only slightly on the missing frequency data; we report those results which were most conservative). For each admissible triplet i of DNA segments we determined $\tilde{Q}_i$. Finally, for all

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

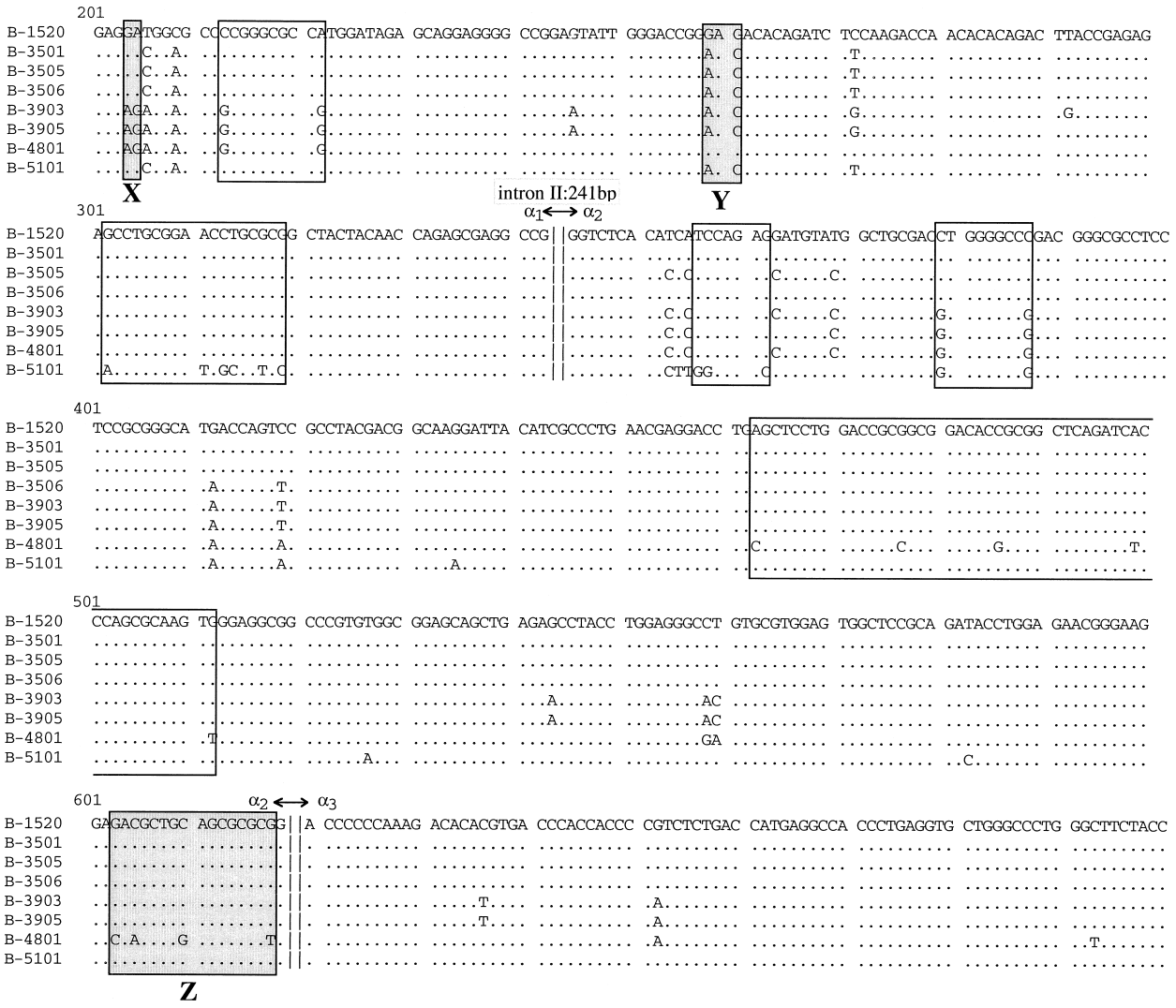

Fig. 5. Part of the HLA-B locus (no introns shown) and the set of haplotypes found in the Kaingang population. Boxes show the candidate segments for analysis. Only those segments which fulfil requirements (a) to (e) (see the text) may ultimately be used in the test (grey boxes).

samples (long- and short-range data for three populations) we calculated the fractions $\hat{q}$ (see Table 4). Generally, we observed $\hat{q}$ to be smaller in short-range cases than in long-range cases. We tested the hypotheses H_0: $q = 1/2$ and H_0: $q = q'$ as described above (see Table 4). Recombination-only ($q = q'$) as null hypothesis was rejected for the short-range data in populations K and H. On the other hand, the model of conversion-only ($q = 1/2$) was also rejected for population H; for the other populations the conversion-only model was not rejected. For the long-range data, the conversion-only model was clearly rejected in all cases. The recombination-only model was also rejected for population K. The power of the test is moderately high for the short-range data and almost 1·0 for the long-range data where a large number of observations is available. The test is based on the assumption of independence of the values $\tilde{Q}_1, \ldots, \tilde{Q}_n$ within each sample. This assumption can be violated, in particular with data from a single chromosomal region or a single population. With enough sequence data available it would be obvious to additionally require that any pair of admissible triplets must be non-intersecting (i.e. segment $\mathscr{X}$ of triplet 2 should be downstream of segment $\mathscr{Z}$ of triplet 1). Another approach to cope with the problem of non-independence is to pool data from isolated populations. As an example we performed the test for the pooled observations in populations G and K.

Finally, we determined the maximum likelihood estimates $\hat{k}$ of the percentage of recombination for the three populations. Table 4 lists the derived maximum likelihood estimates $\hat{k}$. Generally, $\hat{k}$ is lower for the short-range cases. The results also suggest (see Fig. 4) that q, and therefore the derived estimates, change appreciably with respect to their neutral values only if epistasis is very strong. Overdominant selection has almost no influence on q.

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

Table 4. *Results of statistical analysis*

				Sign test						
				H_0: $p = p'$ (recombination)			H_0: $p = 0{\cdot}5$ (conversion)			
Data set[a]	n[b]	$\hat{x}$	$\hat{q}$	Critical region	Power	Reject H_0?	Critical region	Power	Reject H_0?	MLE[c]
G (SR)	14	9	0·6429	$\leqslant 8$	0·7880	No	$\geqslant 11$	0·7916	No	36·2
H (SR)	28	19	0·6786	$\leqslant 19$	0·9822	Yes	$\geqslant 19$	0·9856	Yes	46·4
K (SR)	11	3	0·2727	$\leqslant 6$	0·7256	Yes	$\geqslant 9$	0·7112	No	0·0
G+K (SR)	25	12	0·4800	$\leqslant 16$	0·9461	Yes	$\geqslant 18$	0·9482	No	0·0
G (LR)	370	346	0·9351	$\leqslant 342$	1	No	$\geqslant 216$	1	Yes	98·7
K (LR)	1455	1287	0·8845	$\leqslant 1372$	1	Yes	$\geqslant 787$	1	Yes	95·2
G+K (LR)	1825	1633	0·8948	$\leqslant 1725$	1	Yes	$\geqslant 979$	1	Yes	95·9
DM locus	7	5	0·7143	$\leqslant 5$	0·9375	Yes	$\geqslant 7$	0·8412	No	76·0

[a] The first two blocks correspond to the short-range (SR) and long-range (LR) HLA data (Parham *et al.*, 1997) for the Guarani (G), Kaingang (K) and Havasupai (H) populations. Counts for the Kaingang and Guarani were also pooled (G+K). The last row corresponds to the data from Tishkoff *et al.* (1998) (DM locus).
[b] n = total counts of segment triplets which fulfilled the criteria (a)–(f) given in the text (cf. Fig. 5). $\hat{x}$ = number of counts with $\tilde{Q} = -1$. $\hat{q} = \hat{x}/n$.
[c] Maximum likelihood estimate of the percentage (k) of (reciprocal) recombination events among recombination and conversion events, assuming neutrality ($s = 0$).
Significance levels for the signtest were $\alpha = 5\,\%$ (short range data and DM locus) and $\alpha = 0{\cdot}1\,\%$ (long range data).

To perform the test and to obtain values for the MLE and function f we needed to specify parameters for the recombination and conversion rates, the population size and time t_1. Most critical is the assumption for the population size N. We assumed $N = 10^4$ (Grimes, 1992). A population size smaller by one order of magnitude ($N = 10^3$) considerably lowers the values of q and potentially leads to overestimates in the percentage of conversion events (cf. Table 2). This problem is due to the strong influence of random drift upon statistic Q in this case. Furthermore, we assumed $d = 4 \times 10^{-6}$ (cf. Zangenberg *et al.*, 1995) and $r_1 = 10^{-6}$ (short-range case) and $r_1 = 10^{-3}$ (long-range case), assuming a recombination rate of 1 % per 1 Mb. There is still some uncertainty about the recombination rate in the HLA class I gene region, with estimates ranging between the value above (Thomsen *et al.*, 1994) and the lower estimate of 0·31 % recombination per meiotic event for the 1·4 Mb spanning HLA class I (Martin *et al.*, 1995). Again, q is somewhat dependent on the assumed recombination rate. Finally, we assumed $t_1 = 250$ generations, corresponding to a lower bound for the time since arrival in the Americas. Simulations revealed that q depended only slightly on changes of the parameter t_1. We checked a range from $t_1 = 125$ to $t_1 = 500$ (see Table 2). Note for the long-range case that q may decrease with increasing recombination rate or time t_1: in this case both linkage disequilibria, D'_{xy} and D'_{xz}, decay very fast to values close to zero and then either of them may become zero by chance with equal likelihood. This leads to a deficit of outcomes $\tilde{Q} = -1$.

(ii) *Haplotype data from the DM region*

Tishkoff *et al.* (1998) analysed haplotypes of the $(CTG)_n$ repeat and three flanking markers at the myotonic dystrophy (DM) locus for 25 human populations. The three markers are each biallelic (one Alu deletion polymorphism, two restriction site polymorphisms). All three polymorphisms are presumably more recent than the divergence of humans and great apes. Furthermore, these markers are unevenly spaced ($\nu_1 = 2{\cdot}5$ kb, $\nu_2 = 17{\cdot}5$ kb, $h = 7$). In this case, each population provides one single observation. Conditioning on $\tilde{Q} \neq 0$, the sample size reduces to seven for the pooled data (cf. table 3 in Tishkoff *et al.*, 1998), and $\hat{q} = 5/7 = 0{\cdot}7143$. To determine q' we chose parameters $t_1 = 5000$, $r_1 = 2{\cdot}5 \times 10^{-5}$, $r_2 = 1{\cdot}75 \times 10^{-4}$, $d = 4 \times 10^{-6}$ and $N = 10^4$. For instance, we obtain $q' = 0{\cdot}5349$ ($k = 50\,\%$) and $q' = 0{\cdot}9756$ ($k = 100\,\%$). The null hypothesis 'recombination only' (H_0: $q = 0{\cdot}9756$) is rejected, the null hypothesis 'conversion only' (H_0: $q = 0{\cdot}5$) is not rejected. The MLE for the proportion of reciprocal recombination events for these data is $\hat{k} = 76\,\%$.

5. Discussion

Reciprocal recombination and conversion with short conversion tracts both lead to a decay of linkage disequilibrium. However, the decay rate depends on the distance between loci only in the first case. The same observation was made earlier by Clark & Zheng

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

(1997) for a model of reciprocal (conjugation) versus non-reciprocal (transformation) recombination in bacteria. These authors also studied a deterministic three-locus model (in a haploid context, though) and derived recursions for the pairwise and the three-locus linkage disequilibria under neutrality and directional selection. Here, we describe a measure, Q, which is derived from pairwise linkage disequilibria. In contrast to the distribution of plain linkage disequilibria, the distribution of Q is qualitatively different for the two modes of recombination and provides the basis for a statistical test which is applicable to haplotype data from natural populations. We have applied the method developed in the theory section to the HLA class I haplotype frequencies observed in three Amerindian populations. Results for the short-range (within the HLA-B locus) cases are consistent with previous suggestions: haplotype frequencies of South American populations fit a conversion-only model, while those of the North American population are consistent with neither a model of recombination-only nor one of conversion-only. For the long-range (across HLA-A, -B and -C loci) cases a model of conversion-only can be rejected for both the populations for which data are available (Guarani and Kaingang). A model of recombination-only is rejected only in the Kaingang. We have taken an additional step and estimated the fraction of the total (all reciprocal recombination and conversion events) that are reciprocal recombination events. For the short-range cases these estimates are all below 50%, with the estimate for the Havasupai highest at 46%. For the long-range cases, on the other hand, these estimates are close to 100%. Our interpretation is that conversion may have played a major role in generating the diversity within loci, at least for the HLA-B locus. This is most apparent for the South American populations. Reciprocal recombination has apparently played a much larger role in breaking up haplotypes at the multiple loci level.

Both the model of recombination-only and that of conversion-only are rejected for the Havasupai HLA-B locus data. Both mechanisms have had a similar impact, as suggested by the estimate of k near 50%. One possible explanation is that neither of these mechanisms is responsible for generating the observed data. By chance the Havasupai may have retained haplotypes generated prior to the founding of the population. If so, the assumptions underlying the application of this test may not hold.

Unfortunately, suitable data from other genes or species are still scarce. One other data set is provided by a haplotype study of the DM region (Tishkoff *et al.*, 1998). The involved interlocus distances are intermediate between the short- and long-range cases considered for the HLA analysis. One expects that the role of reciprocal recombination should be intermediate compared with the short- and long-range cases as well. This expectation is confirmed by the derived maximum likelihood estimate of $k = 76\%$.

To draw firm conclusions larger data sets would be required. For the HLA data we have considered multiple triplets of DNA segments for each population. The most severe shortcoming is that the chosen triplets, originating from not only a single population but also a single gene locus, are generally not independent. Correction is not straightforward. Ideally one would consider a large number of triplets, each spanning a different stretch of DNA. This is possible if the stretch of DNA under consideration is long, or if multiple DNA regions are considered. Alternatively one can consider jointly the results for different populations. As an illustration we combined the data for the two South American (Guarani and Kaingang) samples. Similarly, in the DM sample any one triplet is from a different population. As a further alternative one may consider data of homologous regions of a set of closely related species.

It remains to be investigated whether the error incurred by neglecting the requirement of independence is substantial. Simulations based on coalescent theory rather than a Wright–Fisher model may yield some insight. This will be the topic of future work.

Our test for the roles of conversion and recombination does not explicitly incorporate natural selection and might therefore seem inapplicable to HLA data. We have examined the robustness of our test to two models of selection. We found that overdominant selection at any one locus has little effect on the ratio, Q. Epistatic selection, even in the absence of conversion, biases the values of q downward, resembling those which result from conversion. Selection, however, must be very strong to have such an effect. Furthermore, one can distinguish between the effects of conversion and of a combination of epistasis and recombination. If epistasis were playing a dominant role we would expect haplotypes to appear to be old. That is, we would expect them to be found in multiple populations, at relatively high frequencies. For the HLA data considered here, many of the haplotypes found in the South American populations appear to be relatively new (Parham *et al.*, 1997). Epistasis may, however, have played a role in generating the frequencies of the North American population, where neither conversion alone nor reciprocal recombination alone would explain the observed data.

On the other hand, most of the variability at the CTG_n repeat locus in the DM region appears to be neutral (Tishkoff *et al.*, 1998). Therefore, the observed variability for the three adjacent neutral markers is most likely not the result of natural selection.

We have considered a model wherein the recombination rate increases linearly with the distance between two DNA segments. There may exist,

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

however, hotspots of reciprocal recombination along the chromosome. If such a hotspot exists between the two DNA segments that are closer together ($\mathscr{X}$ and $\mathscr{Y}$ in Fig. 2), the fraction q will be small, as it is under the conversion model. Thus, for the entire HLA-B, HLA-C, HLA-A region we would expect to observe a much lower q value if there were a hotspot between HLA-B and HLA-C. For the within-locus samples (HLA-B) multiple triplets are considered. If a subset of these triplets were to include the hypothetical hotspot, we would still expect the effect of such a hotspot to be minimal.

An essential assumption for the model to be valid is that conversion tracts are short. More precisely, this means that the tract length may vary but the admissible length is limited by the distance between two adjacent markers as described in Section 4. This condition appears to be satisfied in the HLA system (Kuhner & Peterson, 1992; Parham *et al*., 1995). Other regions and other species (Hilliker *et al*., 1994), however, may exhibit much longer average conversion tracts, limiting the applicability of the above model.

In the DM region two markers are restriction site polymorphisms characterized by a few base pairs. One is an Alu deletion polymorphism of about 1 kb. However, the markers are spaced by 2·5 kb and 17·5 kb. Thus, the requirement that the conversion tract covers at most one locus is satisfied.

Several strategies to document conversion events have been applied in the past. So far, it has rarely been successful in identifying single conversion events by genealogical analysis of sequence data (Warner *et al*., 1996; van der Steege *et al*., 1996). Usually, nucleotide sequence patterns of a number of different haplotypes in a population are used to characterize candidate regions for conversion events which may have taken place in the past. Phylogenetic (Gyllensten *et al*., 1991), graphical or statistical (Sawyer, 1989; Kuhner *et al*., 1991; Betran *et al*., 1997; Weiller, 1998) methods are employed. However, it is generally very difficult to reconstruct ancestral haplotypes. Even further complication arises if the observed haplotype variation may be due to a combination of mechanisms, such as crossing-over and conversion. There is generally not a unique pathway for reconstructing the ancestral haplotypes and as a consequence the relative numbers of recombination and conversion events cannot be determined. In our model we do not attempt to single out individual sequence stretches which may have undergone conversion but we use linkage disequilibria between triplets of admissible loci within the region of interest. Since three-locus linkage disequilibria behave differently under the two mechanisms, conversion and reciprocal recombination, we were able to derive an estimate for their relative percentages. A drawback of this approach is that it requires reliable measurements of haplotype frequencies. This problem, however, should be eased in the near future when re-sequencing strategies, such as chip-based sequencing technologies, are routinely available.

Evaluating the relative roles of recombination and conversion, we find that conversion may have had a greater impact than reciprocal recombination in generating local haplotype patterns. The highly variable HLA loci provide us with a rare opportunity for such evaluation. The challenge now is to confirm these conclusions with additional data from the HLA system and other loci. A recent study of Zhao *et al*. (1998) reveals that gene conversion may play a major role in the evolution of human red and green opsin genes, questioning the view that opsin variability was primarily generated by unequal cross-over (Vollrath *et al*., 1988). Unfortunately, haplotype frequency data applicable in our test are not yet available. For the moment, it has to remain a hypothesis that conversion might be a much more ubiquitous mechanism for generating novelty in a species' genome than previously thought.

Appendix

A.1. *Derivation of the decay rate of* D *due to conversion*

Consider a region with two loci (DNA segments), $\mathscr{X}$ and $\mathscr{Y}$. We assume that these DNA segments have length $l > 1$ and comprise a few base pairs – some 30, say. These segments may be overlapped by a conversion tract of size ν'. At each base a conversion event originates at rate δ and its conversion tract extends ν' bases downstream. There are $(\nu' - l + 1)$ possible positions where conversion may originate and overlap a segment – $\mathscr{X}$ say. Therefore, the rate at which $\mathscr{X}$ is converted is $\delta(\nu' - l + 1)$. To simplify the notation we substitute $\delta(\nu' - l + 1)$ by d. Note that δ is not necessarily an overall per base pair conversion rate. Conversion events with long tracts may originate at a different rate from those with short tracts (Hilliker *et al*., 1994). Here, we are concerned only with conversion events with short tracts.

Let the frequencies of the four haplotypes XY, Xy, xY and xy be ξ_1 to ξ_4, respectively. After one generation the haplotype frequencies will change as a result of conversion. Recalling that alleles are converted with equal probabilities, for instance ξ_1 becomes

$$(1-2d)\,\xi_1 + 2d(\xi_1^2 + \xi_1\xi_2 + \xi_1\xi_3 + (1/2)\xi_1\xi_4 + (1/2)\xi_2\xi_3)$$

which simplifies to

$$\xi_1 - dD,$$

where $D = \xi_1\xi_4 - \xi_2\xi_3$. Generally, frequency ξ_i becomes $\xi_i + \eta_i\, dD$, where $\eta_1 = \eta_4 = -1$ and $\eta_2 = \eta_3 = 1$. Thus, after one generation, D changes to $D(1-d)$. A

https://doi.org/10.1017/S0016672399004036 Published online by Cambridge University Press

similar calculation leads to the decay rate $(3/2)d$ of the higher order linkage disequilibrium $D_{\mathscr{X}\mathscr{Y}\mathscr{X}}$.

A.2. *Overdominant selection*

Although the theory above assumes neutrality for all loci, it remains valid when overdominance is incorporated. Consider two loci. Interpreting w_i, $i = 1, \ldots, 4$, as Malthusian fitness parameters, and γ as a dummy for r_1 or d, the change of the haplotype frequencies (as defined in Section A 1) in time is given by the differential equations

$$\dot{\xi}_i(t) = \xi_i(w_i - \bar{w}) + \eta_i \gamma(\xi_1 \xi_4 - \xi_2 \xi_3), \ (i = 1, \ldots, 4).$$

Expressed in terms of allele frequencies X, Y and linkage disequilibrium $D_{\mathscr{X}\mathscr{Y}}$ and with overdominant selection (selective advantage of the heterozygotes $= s$) at $\mathscr{Y}$, the transformed differential equations are

$$\begin{aligned}
\dot{X}(t) &= sD_{\mathscr{X}\mathscr{Y}}(t)\,(1-2Y(t)),\\
\dot{Y}(t) &= sY(t)\,(1-Y(t))\,(1-2Y(t)),\\
\dot{D}_{\mathscr{X}\mathscr{Y}}(t) &= D_{\mathscr{X}\mathscr{Y}}(t)\,(-\gamma + s(1-2Y(t))^2). \qquad \text{(A 1)}
\end{aligned}$$

Replacing D by the normalized linkage disequilibrium D', the last equation becomes

$$\dot{D}'_{\mathscr{X}\mathscr{Y}}(t) = D'_{\mathscr{X}\mathscr{Y}}(t)\,[-\gamma \pm sY(t)\,(1-2Y(t))\,(1-D'_{\mathscr{X}\mathscr{Y}})], \qquad \text{(A 2)}$$

where the plus sign is valid if $D > 0$ and $X < Y$ or if $D < 0$ and $X < (1-Y)$ and the minus sign in the opposite cases. Obviously, the decay of D', which is determined by γ under neutrality, may accelerate or slow as a result of selection. If selection is very strong, however, then allele Y will quickly reach its equilibrium $Y = 1/2$, thereby making (A 2) identical to the neutral case. For weak to intermediate selection, one may consider a perturbation argument. Let

$$D' \approx D'^{(0)} + sD'^{(1)} + O(s^2).$$

The zeroth-order equation is readily solved:

$$D'^{(0)}(t) = D'^{(0)}(0)\exp(-\gamma t).$$

If t is small enough and since $D'^{(0)}(0) = 1$, we may approximate $(1 - D'^{(0)}(t))$ by γt and the right side in (A 2) by

$$D'_{\mathscr{X}\mathscr{Y}}(t)\,(-\gamma \pm \gamma st\,Y(t)\,(1-2Y(t))).$$

Thus, the contribution to the decay rate of D' from selection is governed by the parameter γs, which may be neglected with respect to γ. Numerical results show that this approximation is reasonably accurate.

We would like to thank Kelly Arnett, William Klitz and Yannis Michalakis for stimulating discussions and two anonymous reviewers for valuable comments on an earlier draft of the manuscript. This work was supported by US National Institute of Health grant no. GM28248 to M.S. and in part supported by a research grant (to T.W.) from the German Academic Exchange Service. T.W. would like to thank A. Hatzigeorgiou for providing office space and computation facilities while on a stay in Heraklion, Greece.

References

Betran, E., Rozas, J., Navarro, A. & Barbadilla, A. (1997). The estimation of the number and length distribution of gene conversion tracts from population DNA sequence data. *Genetics* **146**, 89–99.

Clark, A. & Zheng, Y. (1997). Dynamics of linkage disequilibrium in bacterial genomes undergoing transformation and/or conjugation. *Journal of Evolutionary Biology* **10**, 663–676.

Geliebter, J. & Nathenson, S. G. (1988). Microrecombinations generate sequence diversity in the murine major histocompatibility complex: analysis of the K^{bm3}, K^{bm4}, K^{bm10} and K^{bm11} mutants. *Molecular and Cellular Biology* **8**, 4342–4352.

Grimes, B. F. (ed.) (1992). *Ethnologue: Languages of the World*, 12th edn. Dallas, Texas: Summer Institute of Linguistics.

Gyllensten, U. B., Sundvall, M. & Erlich, H. A. (1991). Allelic diversity is generated by intraexon sequence exchange at the DRB1 locus of primates. *Proceedings of the National Academy of Sciences of the USA* **88**, 3686–3690.

Hilliker, A. J., Harauz, G., Reaume, A. G., Gray, M., Clark, S. H. & Chovnick, A. (1994). Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. *Genetics* **137**, 1019–1026.

Hudson, R. R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631.

Hughes, A. L. (1991). Testing for interlocus genetic exchange in the MHC: a reply to Andersson and coworkers. *Immunogenetics* **33**, 243–246.

Kuhner, M. K. & Peterson, M. J. (1992). Genetic exchange in the evolution of the human MHC class II loci. *Tissue Antigens* **39**, 209–215.

Kuhner, M. K., Lawlor, D. A., Ennis, P. & Parham, P. (1991). Gene conversion in the evolution of the human and chimpanzee MHC class I loci. *Tissue Antigens* **38**, 152–164.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* **49**, 49–67.

Lewontin, R. C. (1995). The detection of linkage disequilibrium in molecular sequence data. *Genetics* **140**, 377–388.

Markow, T., Hedrick, P. W., Zuerlein, K., *et al.* (1993). HLA polymorphism in the Havasupai: evidence for balancing selection. *American Journal of Human Genetics* **53**, 943–952.

Martin, M., Mann, D. & Carrington, M. (1995). Recombination rates across the HLA complex: use of microsatellites as a rapid screen for recombinant chromosomes. *Human Molecular Genetics* **4**, 423–428.

Ohta, T. & Kimura, M. (1969). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13**, 47–55.

Parham, P., Adams, E. & Arnett, K. L. (1995). The origins of HLA-A,B,C polymorphism. In *Immunological Reviews: Origin of Major Histocompatibility Complex Diversity* (ed. G. Möller), vol. 143, pp. 141–180. Copenhagen: Munksgaard.

Parham, P., Arnett, K., Adams, E. J., Little, A. M., Tees, K., Barber, L. D., Marsh, S. G. E., Ohta, T., Markow, T.

& Petzl-Erler, M. L. (1997). Episodic evolution and turnover of HLA-B in the indigenous human populations of the Americas. *Tissue Antigens* **50**, 219–232.

Petzl-Erler, M. L., Luz, R. & Sotomaior, V. S. (1993). The HLA polymorphism of two distinctive South American Indian tribes: The Kaingang and the Guarani. *Tissue Antigens* **41**, 227–237.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution* **6**, 526–538.

Sokal, R. R. & Rohlf, F. J. (1981). *Biometry*. San Francisco: W. H. Freeman.

Stephens, J. C. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Molecular Biology and Evolution* **2**, 539–556.

Takahata, N. (1994). Comments on the detection of reciprocal recombination or gene conversion. *Immunogenetics* **39**, 146–149.

Thomsen, M., Neugebauer, M., Arnaud, J., Borot, N., Sevin, A., Baur, M. & Cambon-Thomsen, A. (1994). Recombination fractions in the HLA system based on the data set 'Provinces Françaises': indications of haplotype-specific recombination rates. *European Journal of Immunogenetics* **21**, 33–43.

Tishkoff, S., Goldman, A., Calafell, F., Speed, W., Deinard, A., Bonne-Tamir, B., Kidd, J., Pakstis, A., Jenkins, T. & Kidd, K. (1998). A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and the origin of myotonic dystrophy mutations. *American Journal of Human Genetics* **62**, 1389–1402.

van der Steege, G., Grootscholten, P., Cobben, J., Zappata, S., Scheffer, H., den Dunnen, J., van Ommen, G. J., Brahe, C. & Buys, C. H. (1996). Apparent gene conversions involving the SMN gene in the region of the spinal muscular atrophy locus on chromosome 5. *American Journal of Human Genetics* **59**, 834–838.

Vollrath, D., Nathans, J. & Davis, R. W. (1988). Tandem array of human visual pigment genes at Xq28. *Science* **240**, 1669–1671.

Warner, J., Barron, L., Fitzpatrick, D. & Brock, D. (1996). A gene conversion event at the Huntington's CAG repeat. *American Journal of Human Genetics* **59**, A292.

Weiller, G. F. (1998). Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Molecular Biology and Evolution* **15**, 326–335.

Zangenberg, G., Huang, M. M., Arnheim, N. & Erlich, H. (1995). New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. *Nature Genetics* **10**, 407–414.

Zhao, Z., Hewett-Emmett, D. & Li, W. (1998). Frequent gene conversion between human red and green opsin genes. *Journal of Molecular Evolution* **46**, 494–496.