Estimation of the Neutrality Index

Nina Stoletzki[1] and Adam Eyre-Walker*,[1]

[1]Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton, United Kingdom

***Corresponding author:** E-mail: a.c.eyre-walker@sussex.ac.uk.

Associate editor: Dan Graur

Abstract

The McDonald–Kreitman (MK) test is a simple and widely used test of selection in which the numbers of nonsilent and silent substitutions (D_n and D_s) are compared with the numbers of nonsilent and silent polymorphisms (P_n and P_s). The neutrality index ($\mathrm{NI} = D_sP_n/D_nP_s$), the odds ratio (OR) of the MK table, measures the direction and degree of departure from neutral evolution. The mean of NI values across genes is often taken to summarize patterns of selection in a species. Here, we show that this leads to statistical bias in both simulated and real data to the extent that species, which show a pattern of adaptive evolution, can apparently be subject to weak purifying selection and vice versa. We show that this bias can be removed by using a variant of the Cochran—Mantel–Haenszel procedure for estimating a weighted average OR. We also show that several point estimators of NI are statistically biased even when cutoff values are employed. We therefore suggest that a new statistic be used to study patterns of selection when data are sparse, the direction of selection: $\mathrm{DoS} = D_n/(D_n + D_s) - P_n/(P_n + P_s)$.

Key words: adaptive evolution, McDonald–Kreitman test, bias of odds ratios, Mantel–Haenszel test, heterogeneity.

Introduction

Understanding the nature of natural selection on DNA sequences is one of the central goals of molecular evolution. The McDonald–Kreitman (MK) test of selection (McDonald and Kreitman 1991) compares the numbers of nonsilent (P_n) and silent (P_s) polymorphisms to the numbers of nonsilent (D_n) and silent (D_s) substitutions per locus. When silent and nonsilent substitutions are interspersed, and therefore share a common genealogy or genealogies and sampling scheme, one may perform a simple test of independence on the 2 × 2 contingency table to test for a deviation from neutrality. Under strict neutrality, where mutations are either strongly deleterious or neutral, we expect the two ratios, P_n/P_s and D_n/D_s to be the same. The direction and degree of departure from neutrality can then be quantified using the neutrality index (NI; Rand and Kann 1996), the odds ratio (OR) of the MK contingency table. NI is defined as $(P_n/P_s)/(D_n/D_s)$, however, the inverse of the NI has also been used (e.g., Tachida 2000; Presgraves 2005). Under the assumption that silent mutations are neutral, $\mathrm{NI} > 1$ indicates an excess of amino acid polymorphisms (as expected when there are slightly deleterious mutations), and $\mathrm{NI} < 1$ indicates an excess of nonsilent divergence (as expected under positive selection). Typically, the MK test is applied to protein-coding data, but it can be applied to any two categories of sites that are interspersed, such as protein and nonprotein-binding sites in a regulatory element (e.g., Jenkins et al. 1995).

There are, however, problems with using ORs such as NI: First, being a ratio of two ratios, NI tends to be biased and to have a large variance, especially when numbers of observations are small. Additionally, NI is undefined when either D_n or P_s is 0. A mean NI estimate is often obtained by averaging NI after the exclusion of genes for which NI is undefined or after removing genes that do not have sufficient numbers of substitutions or polymorphisms (see, e.g., Bazin et al. 2006; Meiklejohn et al. 2007; Hughes et al. 2008). Such an average will be biased, and the loss of genes can be severe; for example, in a recent analysis of NI values in bacteria, Hughes et al. (2008) had to exclude one-third of all genes due to undefined NI values. Here, we demonstrate the biases that can arise in the estimation of NI in both real and simulated data.

Materials and Methods

Data

We analyzed D_n, D_s, P_n, and P_s values from a number of studies: 1) 115 genes for which we have polymorphism data from *Drosophila simulans* and divergence between *D. simulans* and *D. yakuba* (Welch 2006), 2) 98 genes for which we have polymorphism data from *D. melanogaster* and divergence from the ancestor of *D. melanogaster*–*D. simulans* to *D. melanogaster* (Presgraves 2005), 3) 410 genes for which we have polymorphism data from both *Escherichia coli* and *Salmonella enterica* and divergence between the two species (Charlesworth and Eyre-Walker 2006), and 4) 11,624 genes for which we have polymorphism data from humans and divergence between humans and chimpanzees (Bustamante et al. 2005).

Simulations

To investigate the effect of complete linkage on the bias in estimates of the NI, that is, when there is no recombination, we randomly generated the total length of a genealogy by sampling from a set of exponential distributions; that is, the time for m branches to coalesce to $m - 1$ branches is

Table 1. Example MK Tables. Tables Contain the Expected Numbers of D_n, D_s, P_n, and P_s. The Mean of $\mathrm{NI_{TG}}$ Is Given with the Standard Error; These Were Obtained by Generating 100 Data Sets of 100 Contingency Tables Using the Expected Numbers. The $\mathrm{NI_{TG}}$ Value in Parentheses Is for Simulations in Which the NI Is Lognormally Distributed with a Variance Parameter of One.

$\mathrm{NI_{True}} = 1$							
(a)	Polymorphism	Divergence	$\mathrm{NI_{simple}} = 0.83$	(b)	Polymorphism	Divergence	$\mathrm{NI_{simple}} = 1.53$
Silent	2	1	$\mathrm{NI_R} = 0.55$	Silent	4	4	$\mathrm{NI_R} = 0.85$
Nonsilent	1	1	$\mathrm{NI_{TG}} = 1.00 \pm 0.01$	Nonsilent	2	2	$\mathrm{NI_{TG}} = 0.98 \pm 0.02$
			$(\mathrm{NI_{TG}} = 1.06 \pm 0.03)$				$(\mathrm{NI_{TG}} = 1.03 \pm 0.02)$
$\mathrm{NI_{True}} = 1.5$							
(c)	Polymorphism	Divergence	$\mathrm{NI_{simple}} = 0.88$	(d)	Polymorphism	Divergence	$\mathrm{NI_{simple}} = 2.49$
Silent	1	1.5	$\mathrm{NI_R} = 0.60$	Silent	5	7.5	$\mathrm{NI_R} = 1.48$
Nonsilent	1	1	$\mathrm{NI_{TG}} = 1.53 \pm 0.04$	Nonsilent	5	5	$\mathrm{NI_{TG}} = 1.51 \pm 0.01$
			$(\mathrm{NI_{TG}} = 1.48 \pm 0.04)$				$(\mathrm{NI_{TG}} = 1.51 \pm 0.02)$
$\mathrm{NI_{True}} = 0.75$							
(e)	Polymorphism	Divergence	$\mathrm{NI_{simple}} = 0.44$	(f)	Polymorphism	Divergence	$\mathrm{NI_{simple}} = 1.30$
Silent	1	0.75	$\mathrm{NI_R} = 0.30$	Silent	4	3	$\mathrm{NI_R} = 0.72$
Nonsilent	1	1	$\mathrm{NI_{TG}} = 0.77 \pm 0.02$	Nonsilent	4	4	$\mathrm{NI_{TG}} = 0.73 \pm 0.01$
			$(\mathrm{NI_{TG}} = 0.79 \pm 0.03)$				$(\mathrm{NI_{TG}} = 0.78 \pm 0.01)$

exponentially distributed with a mean of $\theta/(m(m-1)$, where θ is a constant. If we scale our trees so the average total length is 1, then if the total length of a randomly generated tree is λ, the number of nonsynonymous and synonymous polymorphisms are Poisson distribution with means of $\lambda E(P_\mathrm{n})$ and $\lambda E(P_\mathrm{s})$, respectively.

Results and Discussion

Summary Estimators

The bias that can arise by simply averaging NI values, after excluding undefined values, is illustrated in table 1. Here, each pair of MK tables shares the same NI value, $\mathrm{NI_{True}} = E(D_\mathrm{s})\ E(P_\mathrm{n})/(E(D_\mathrm{n})\ E(P_\mathrm{s}))$, where $E(x)$ refers to the expected value of x (i.e., the mean value for a sample of infinite size). If we assume free recombination and no epistasis, then P_s, D_s, P_n, and D_n are independently and Poisson distributed and we can calculate the expected value of NI excluding cases in which NI is undefined, $\mathrm{NI_{simple}}$ as follows:

$$E[\mathrm{NI_{simple}}] = E[D_\mathrm{s}]E[P_\mathrm{n}]\frac{\sum_{x=1}^{\infty} Z(x, E[D_\mathrm{n}])1/x}{\sum_{x=1}^{\infty} Z(x, E[D_\mathrm{n}])}\frac{\sum_{x=1}^{\infty} Z(x, E[P_\mathrm{s}])1/x}{\sum_{x=1}^{\infty} Z(x, E[P_\mathrm{s}])}, \tag{1}$$

where

$$Z(x,\mu) = \frac{e^{-\mu}\mu^x}{x!}$$

is the Poisson distribution. Note that the denominators in equation (1) can be simplified to $1 - e^{-E[D_\mathrm{n}]}$ and $1 - e^{-E[P_\mathrm{s}]}$, respectively.

We find that $\mathrm{NI_{simple}}$ is substantially above or below $\mathrm{NI_{True}}$ due to the exclusion of genes and small sample bias. There is sufficient bias in these examples for the left-hand tables to indicate positive selection, whereas the right-hand tables indicate negative selection. This means that two species subject to the same levels of positive and negative selection can have different mean NI values simply because one species has more or less polymorphism or substitution data per gene than the other. Such a bias is clearly undesirable.

The bias in $\mathrm{NI_{simple}}$ arises from two sources: first, the exclusion of genes where NI is undefined and second, from a tendency for the average value of a ratio to overestimate the true value (even when the denominator is relatively large). The two sources of bias can also be illustrated as follows (table 1). We can estimate the expected value of NI for the restricted set of genes that have a defined NI, $\mathrm{NI_R}$, estimated as $E[D_\mathrm{s}]\ E[P_\mathrm{n}]/(E[D_\mathrm{n}]\ E[P_\mathrm{s}])$, when $D_\mathrm{n} > 0$ and $P_\mathrm{s} > 0$. $\mathrm{NI_R}$ is considerably below the true NI for all genes, illustrating that excluding genes with undefined NI tends to reduce the estimate of NI. This is because excluding genes with either undefined $D_\mathrm{s}/D_\mathrm{n}$ or with undefined $P_\mathrm{n}/P_\mathrm{s}$ excludes potentially large NI values. However, $\mathrm{NI_{simple}}$ is greater than $\mathrm{NI_R}$ in each case, illustrating that even for genes for which NI is defined, $\mathrm{NI_{simple}}$ tends to be an overestimate due to the skew in the distribution of a ratio.

The bias can be more formally quantified as follows. NI is the product of two independent ratios, and if we assume free recombination, the numerator and denominator are independent of each other. We therefore need only consider the bias inherent in the estimation of $1/x$ to understand the biases in NI. This can be calculated as follows. The expected value of a statistic subject to a cutoff value z, below which values are excluded, is

$$E[f(x)] = \frac{\sum_{x=z}^{\infty} Z(x, E[D_\mathrm{n}])f(x)}{\sum_{x=z}^{\infty} Z(x, E[D_\mathrm{n}])} \tag{2}$$

from which the bias can be quantified as follows:

$$B = \frac{E[f(x)]}{f(E[x])}.$$

In the case of $1/x$ this means setting $z = 1$ and dividing the expected value of $1/x$ by $1/E[x]$. This is plotted against $E[x]$ in figure 1. When $E[x]$ is small, $E[1/x]$ underestimates $1/E[x]$ due to the loss of cases when $x = 0$ and $1/x$ is undefined. However, $E[1/x]$ increases and becomes greater than $1/E[x]$ for large $E[x]$ due to the skew in the distribution of

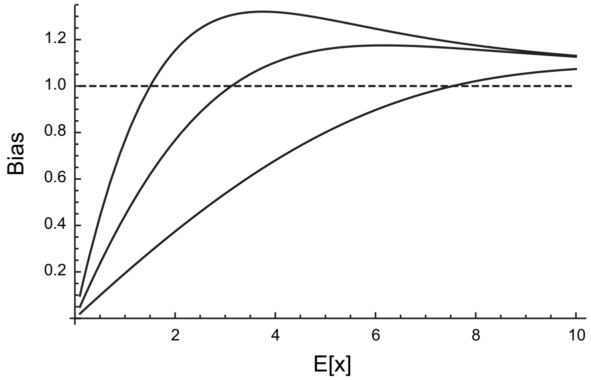

FIG. 1. The bias associated with NI_{simple}. Figure shows $E[1/x]/(1/E[x])$ as a function of $E[x]$ for different cutoff values, z, below which cases are excluded; from left to right $z = 1$, 2, and 5. The straight line indicates the situation of no bias.

$1/x$; so $1/x$ is on average an overestimate of $1/E[x]$ even with very large sample sizes.

The above results were obtained assuming free recombination but remain qualitatively unaffected if there is complete linkage. With no recombination, the distributions of P_n and P_s become complex because they depend both upon the sampling variance, due to the fact we have sampled of sequences of finite length, and the variance due to the coalescence; however, this latter source of variation covaries between P_n and P_s (i.e., they share the same genealogy) with the effect that this becomes largely irrelevant (results not shown).

The problem of estimating an OR and potential approaches to deal with them has received much attention in the statistical literature (see, e.g., Cochran 1954; Haldane 1956; Mantel and Haenszel 1959; Jewell 1984, 1986). It is not appropriate to simply sum contingency tables because of Simpson's paradox (Simpson 1951); summing two contingency tables with the same OR can yield a contingency table with a different OR (see, e.g., Shapiro et al. 2007). A popular method of obtaining the mean OR is to use the Mantel and Haenszel (1959) procedure, which is a particular instance of a general method suggested by Cochran (1954; henceforth, the Cochran-Mantel-Haenszel [CMH] procedure) in which a weighted average OR is computed. The classic CMH procedure has previously been applied to the summary of MK tables (Bartolomé et al. 2005; Maside and Charlesworth 2007). However, Greenland (1982) has pointed out that the CMH method may yield a biased estimate of the OR when there is heterogeneity in the OR between the tables being combined. Because we might reasonably expect NI to vary between genes—either because they have different proportions of slightly deleterious mutations or advantageous substitutions—we need an estimator that takes the heterogeneity into account. It is important to note that tests of heterogeneity in the OR are generally weak (Jones et al. 1989; O'Gorman et al. 1990). Consequently, ORs may still be heterogeneous, even if no significant heterogeneity is detected, and the CMH estimate of the OR may therefore be biased. Fortunately, a very simple variant of the CMH method yields an unbiased estimate of the mean NI, when there is heterogeneity, under most conditions (Tarone 1981; Greenland 1982):

$$NI_{TG} = \frac{\sum D_{si}P_{ni}/(P_{si} + D_{si})}{\sum P_{si}D_{ni}/(P_{si} + D_{si})}, \tag{3}$$

where the index refers to ith gene. To illustrate the power of this estimator to overcome the biases inherent in simply averaging NI values, we generated 100 data sets of 100 contingency tables based on the expected values for P_s, D_s, P_n, and D_n for each of the cases in table 1. NI_{TG} is almost exactly equal to and not significantly different to NI_{True}. This is even the case if we simulate data in which NI varies substantially between genes; for the examples in table 1, we allowed NI to be lognormally distributed with a variance parameter of one; this means that the 5% of genes with the lowest NI have at least a 30-fold lower NI than the 5% of genes with the highest NI. The results from these simulations are given in table 1 and show that NI_{TG} is almost completely unbiased even with very substantial variation in NI. Any residual bias can be removed by increasing the sample size of genes (results not shown).

In common with other CMH type estimators, NI_{TG} is expected to give an unbiased estimate of NI even if there is very little data for each gene, so long as the overall sample size is substantial (the sum of D_n, D_s, P_n, and P_s is of the order of 100s), and there are no systematic (i.e., not due to sampling error) correlations between NI and any of the cells within the MK table. However, if there are systematic correlations NI_{TG} will be biased; for example, NI_{TG} will be biased toward the NI of genes with many synonymous polymorphisms if there is a positive correlation between NI_{True} and $E[P_s]$. There is unfortunately no obvious solution to this problem but it is unlikely to be a major problem because NI_{True} and $E[P_s]$ are not generally correlated (T. Gossman and A. Eyre-Walker, unpublished results).

To investigate the bias of NI in real data, we took published data sets of D_n, D_s, P_n, and P_s from *Drosophila* (Presgraves 2005; Welch 2006), enteric bacteria (Charlesworth and Eyre-Walker 2006), and hominids (Bustamante et al. 2005). Because we do not know NI_{True}, we estimate it using NI_{TG}. However, NI_{TG} can be biased if there is insufficient data across all genes. To investigate whether there is sufficient data, we divided each data set into four groups, estimated NI_{TG} for each and then calculated the mean. We repeated this 1,000 times. We find that mean NI_{TG} for the subsamples are similar to NI_{TG} for the whole data set (*Drosophila*-Presgraves $NI_{TG} = 0.833$, mean $NI_{TG(subsample)} = 0.905$; *Drosophila*-Welch $NI_{TG} = 0.601$, mean $NI_{TG(subsample)} = 0.614$; *E. coli* $NI_{TG} = 0.826$, mean $NI_{TG(subsample)} = 0.827$; *S. enterica* $NI_{TG} = 1.609$, mean $NI_{TG(subsample)} = 1.615$; and humans $NI_{TG} = 1.594$, mean $NI_{TG(subsample)} = 1.596$). This suggests that the sample size is sufficient and NI_{TG} is likely to be an unbiased estimate ofNI_{True}.

As expected, the mean of NI across genes, excluding those with undefined values, NI_{simple}, is larger than NI_{TG} (table 2). This overestimate can be substantial and may be misleading (table 2). In both *E. coli* and *Drosophila*, NI_{TG}

Table 2. Estimates of Mean NI for Real Data Sets.

Polymorphism	Divergence	Number of Genes	Number of Genes with Defined NI	Mean D_n	Mean D_s	Mean P_n	Mean P_s	NI_{TG}	NI_{simple}	NI_{BEW}	NI_{FWW}	Reference
Drosophila simulans	*D. yakuba*	115	100	32	64	3.8	15	0.60	0.96	0.59	0.51	Welch (2006)
D. melanogaster	Ancestor[a]	98	62	4.7	9.2	3.1	8.7	0.83	2.03	0.97	0.71	Presgraves (2005)
Escherichia coli	*Salmonella enterica*	410	394	16	210	2.0	32	0.83	2.01	0.82	0.85	Charlesworth and Eyre-Walker (2006)
S. enterica	*E. coli*	410	386	16	210	2.0	17	1.59	4.11	1.62	1.63	Charlesworth and Eyre-Walker (2006)
Human	Chimpanzee	11,624	4,389	1.8	2.9	1.2	1.4	1.59	1.82	[b]	1.51	Bustamante et al. (2005)

[a] Ancestor of *D. melanogaster* and *D. simulans*.

[b] NI_{BEW} failed to converge for the hominid data set.

indicates adaptive evolution, as others have inferred before (Charlesworth and Eyre-Walker 2006; Welch 2006), while taking the simple average of NI, excluding undefined NI values, suggests slightly deleterious mutations dominate in *E. coli* and one of the *Drosophila* data sets; the other *Drosophila* data set shows a neutral pattern. In the *Salmonella* and hominid data sets, NI_{simple} is qualitatively similar to NI_{TG} but it is much larger in value.

Adaptive Evolution

It is possible to estimate the proportion of nonsynonymous substitutions that are adaptive, α, as 1-NI (Charlesworth 1993; Fay et al. 2001; Smith and Eyre-Walker 2002). Using NI_{TG} to estimate α has the advantage over the method of Fay et al. (2001) of avoiding Simpson's paradox and advantages over the maximum likelihood (ML) methods of Bierne and Eyre-Walker (2004) and Welch (2006) because it is much faster to compute; the rapid computation makes it practical for analyzing the large data sets now available. Confidence intervals and standard errors can be estimated using bootstrapping, which makes no assumptions about the underlying distributions of D_n, D_s, P_n, and P_s.

For comparison, we estimated NI, using the ML method of Bierne and Eyre-Walker (2004) to estimate α, NI_{BEW} (table 2), and using the method of Fay et al. (2001; NI_{FWW}) in which the values of D_n, D_s, P_n, and P_s are summed across genes. As expected the ML estimate of NI is very similar to NI_{TG}. Although contingency tables should generally not be summed because the sum of two tables with the same OR can yield a table with a different OR (Simpson 1951; see also Shapiro et al. 2007), the method of Fay et al. (2001) yields qualitatively similar answers for real data to NI_{TG} and NI_{BEW} (table 2).

Heterogeneity

An added advantage of the CMH framework is the ability to test for heterogeneity in NI using Woolf's test for homogeneity of log ORs (Woolf 1955; Selvin 2004). Although Woolf's test does make assumptions about the distributions of variables in the MK table. Interestingly, when we apply Woolf's test of heterogeneity (as in Sokal and Rohlf 1995 with a continuity correction) to the data sets that we have analyzed, we detect significant heterogeneity in the *D. simulans*–*D. yakuba* data set. For the other data sets used in this study, no significant heterogeneity can be detected; this is perhaps not surprising because tests of heterogeneity have little power, particularly when data are sparse, and the other data sets generally have fewer observations per gene (Jones et al. 1989; O'Gorman et al. 1990; Paul and Donner 1992). The presence of significant heterogeneity in the *D. simulans*–*D. yakuba* data set is surprising given that Welch (2006) has previously tested this same data set for heterogeneity in α, the proportion of adaptive nonsynonymous substitutions. Because $\alpha = 1 - \text{NI}$, this amounts to a test of heterogeneity in NI. The difference may be because Welch (2006) used a ML method to estimate α and constrained α to be positive (i.e., $\text{NI} < 1$) in testing for heterogeneity. Alternatively, the lack of power in the ML method may be due to the number of parameters that have to be estimated.

Point Estimators

NI_{TG} allows one to estimate an average NI estimate across genes; however, sometimes a point estimate of the NI may be needed; for example, if one is interested in correlates of NI, such as the rate of recombination. Clearly, NI_{simple} is not an appropriate statistic because it is biased when D_n and P_s are small. It may therefore be tempting to set a cutoff value and only consider those contingency tables with larger values. Yet, this does not remove the bias; the bias changes and shifts to higher expected values of the variables subject to the cutoff (fig. 1). This occurs for the following reason. Imagine that we have a gene for which the expected value of P_s is 5 (i.e., if we were to sample this gene repeatedly over a long period of time, the mean value of P_s would be 5). If we apply a cutoff of 5, many times the gene will have an observed value of P_s that is lower than 5 and hence be excluded from the analysis; only when the observed value is greater or equal to 5 will the gene be included, but this will lead to low NI values on average because we are only considering cases for this gene when P_s is relatively large.

Several modifications have been proposed to reduce the bias in the estimation of the OR either on the log or on the original scale; two examples are Jewell's estimator of the OR (Jewell 1986) and Haldane's estimator of the log OR (Haldane 1956). For NI, these become, respectively

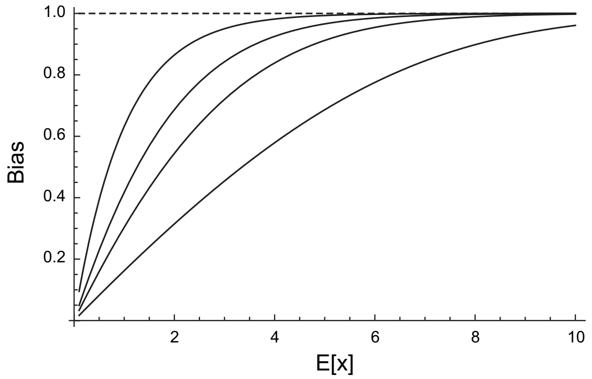

Fig. 2. The bias associated with NI_{Jewell}. Figure shows $E[1/(x+1)]/(1/E[x])$ as a function of $E[x]$ for different cutoff values, z, below which cases are excluded; from left to right $z = 0$, 1, 2, and 5. The straight line indicates the situation of no bias.

$$NI_{Jewell} = \frac{D_s P_n}{(D_n + 1)(P_s + 1)} \quad (4)$$

and

$$LNI_{Haldane} = \mathrm{Log}\left(\frac{(2D_s + 1)(2P_n + 1)}{(2D_n + 1)(2P_s + 1)}\right). \quad (5)$$

A variant of Haldane's estimator has also been employed in which one is added to each cell in the 2 × 2 contingency table: that is,

$$LNI_{Laplace} = \mathrm{Log}\left(\frac{(D_s + 1)(P_n + 1)}{(D_n + 1)(P_s + 1)}\right) \quad (6)$$

(Presgraves 2005; Li et al. 2008; Slotte et al. 2010); because the addition of one is a Laplace correction (see also Presgraves 2005), we refer to this as $LNI_{Laplace}$. All three statistics have the advantage that no gene will be excluded because of undefined values, and $LNI_{Haldane}$ and $LNI_{Laplace}$ have the added advantage that they are symmetrical in the sense that genes with NI = 0.5 and NI = 2 are equidistant from NI = 1 when logs are taken.

However, all these point estimators of the NI are expected to be biased when there is little data. If we assume free recombination and no epistasis, then the sampling errors of P_s, D_s, P_n, and D_n are independent and hence to understand the biases in NI_{Jewell}, we only need consider the bias associated with $1/(D_n + 1)$ and $1/(P_s + 1)$ and hence $1/(x + 1)$. The bias can be calculated using equation (2). The expected value of $1/(x + 1)$ is an underestimate of the true value, $1/E[x]$, until $x > 4$ after which it becomes essentially unbiased (fig. 2). However, just as with $1/x$ (fig. 1), excluding cases of x below a certain cutoff does not help the bias in $1/(x + 1)$; the bias simply shifts to higher expected values of x (fig. 2).

The biases in $LNI_{Haldane}$ and $LNI_{Laplace}$ are less easy to quantify because no single term can be singled out for analysis; the bias depends upon all the cells in the contingency table simultaneously. Therefore, to investigate the matter further, we simulated data under realistic parameter values using the mean values of D_n, D_s, P_n, and P_s from each of the data sets analyzed in table 2. For all genes, we assumed that the expected values of D_n and D_s were equal to the average values of D_n and D_s, but that the expected values of P_n and P_s were equal to average values of P_n and P_s multiplied by factor 2^x, where x varied between −3 and 3; thus, each simulated gene had the same expected value of Log(NI) but different levels of polymorphism, which varied between 8-fold lower and 8-fold higher than the mean to estimate Log(NI). For each combination of expected values of D_n, D_s, P_n, and P_s, we generated 1,000 simulated genes assuming D_n, D_s, P_n, and P_s were Poisson distributed and calculated Log(NI) according to equations (4 and 5).

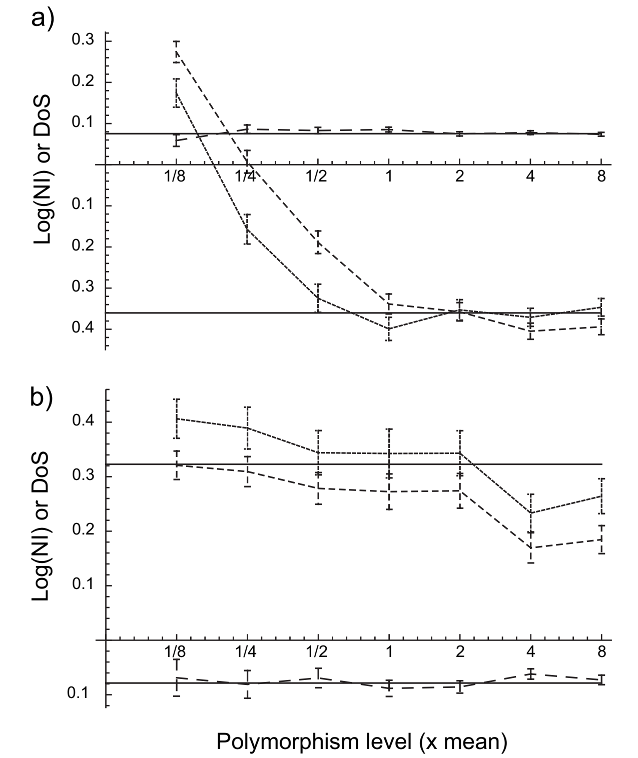

Fig. 3. The bias associated with $LNI_{Haldane}$, $LNI_{Laplace}$, and DoS based on the observed values of D_n, D_s, P_n, and P_s in (*a*) Drosophila using polymorphism data from *Drosophila melanogaster* and the divergence between *D. melanogaster* and the ancestor of that species and *D. simulans* and (*b*) hominids using the polymorphism data from humans and the divergence between humans and chimpanzees. Lines from short dashes to long are: $LNI_{Haldane}$, $LNI_{Laplace}$, and DoS. Also illustrated by solid lines are the true values of Log(NI) and DoS.

For all cases, we found that Log(NI), as estimated by $LNI_{Haldane}$ and $LNI_{Laplace}$, was a biased estimate of Log(NI) and that the bias depended on the level of polymorphism. For most data sets, except the hominids, Log(NI) was overestimated and this bias decreased as polymorphism levels increased; typically, polymorphisms needed to be between 4- and 8-fold higher than the mean levels to obtain an unbiased estimate (fig. 3*a*). In hominids Log(NI) was initially overestimated and then underestimated as the level of polymorphism increased (fig. 3*b*). In both cases,

Table 3. The Mean Value of DoS for Data Simulated Using the Expected Values Given in Table 1.

	True DoS	Mean DoS (standarad error)
1a	0	−0.001 (0.005)
1b	0	−0.004 (0.003)
1c	−0.1	−0.100 (0.005)
1d	−0.1	−0.092 (0.005)
1e	0.07	0.077 (0.007)
1f	0.07	0.067 (0.003)

the estimate of Log(NI) was negatively correlated to the level of polymorphism. Typically, $LNI_{Haldane}$ performed slightly better than $LNI_{Laplace}$.

The fact that point estimates of NI or Log(NI) are biased, with a bias that depends upon sample size, suggests that we may have to use other statistics to investigate whether patterns of selection vary between genes within a genome. A possible solution is to use a statistic, which we call the direction of selection (DoS):

$$\mathrm{DoS} = \frac{D_n}{D_n + D_s} - \frac{P_n}{P_n + P_s}. \qquad (7)$$

This is the difference between the proportion of substitutions and polymorphisms that are nonsynonymous. DoS measures the direction and extent of selection; DoS is positive when there is evidence of adaptive evolution, is zero if there is only neutral evolution, and is negative when there are slightly deleterious mutations segregating. It will be unbiased because it is the difference between two proportions. This lack of bias can be illustrated by calculating the mean DoS for the simulated data in table 1. In each case, DoS is unbiased, even though some genes are excluded because there are either no polymorphisms or substitutions (table 3). We also calculated the mean DoS for the simulated data in figure 3; again DoS appears to show no bias (fig. 3). However, it should be appreciated that DoS and NI are not equal and so two genes can have the same NI value but different DoS values and vice versa. DoS is calculated using the numbers of nonsilent and silent substitutions and polymorphisms per gene; the statistic is biased if the numbers are calculated per nonsynonymous (N) and synonymous (S) site, respectively. As N and S do not cancel out for DoS, as they do for NI, there might be some situations in which a correlation between DoS and some other variable is caused by variation in N/S across genes. However, whether variation in N/S matters depends on how N and S are calculated and the cause of the variation. In general, if the variation in N/S is caused by variation in the pattern of mutation, then it may be responsible for a correlation between DoS and some other variable.

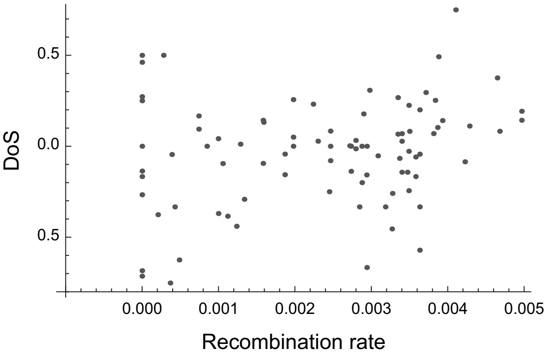

Fig. 4. The correlation between DoS and recombination rate in *Drosophila*.

To illustrate the use of DoS, let us reconsider the work of Presgraves (2005). He showed that the log of the fixation index, FI, where FI = 1/NI, was positively correlated to the rate of recombination when comparing the numbers of nonsynonymous and synonymous substitutions along the *D. melanogaster* lineage since *D .melanogaster* split from *D. simulans* to numbers of nonsynonymous and synonymous polymorphisms in *D. melanogaster* for 98 genes. Presgraves was well aware of the potential biases in using statistics such as NI, and used both Log(FI), and Log(FI) with one added as a continuity correction to each cell to reduce the biases. However, as figure 3 shows, Log(FI) is expected to be positively correlated to overall levels of polymorphism simply because of statistical bias. It is therefore possible that the positive correlation between Log(FI) and recombination is induced by a correlation between levels of diversity, and hence P_s, and the rate of recombination (Begun and Aquadro 1992; Begun et al. 2007), and a correlation between the bias in Log(FI) and the level of diversity in the genes. However, we find that DoS is positively correlated to recombination rate using the same data ($r = 0.23$, $P = 0.033$) (fig. 4); this correlation becomes much more significant if we weight the DoS estimates by a quantity related to the reciprocal of the variance of DoS, $1/(1/(D_n + D_s) + 1/(P_n + P_s))$, giving more weight to genes with more information ($r = 0.41$, $P < 0.0001$). This suggests that genes with high rates of recombination tend to have more adaptive evolution or a smaller proportion of slightly deleterious mutations.

It is sometimes of interest to determine whether the variation that is observed in NI is largely a consequence of variation in D_n/D_s or P_n/P_s (Presgraves 2005; Hughes et al. 2008). For example, Hughes et al. (2008) have shown that bacteria, which have many genes with NI < 1, tend to show an excess of genes with low P_n/P_s rather than high D_n/D_s.

Table 4. Spearman Correlations Between NI_{simple} and P_n/P_s and between DoS and $P_n/(P_n + P_s)$.

		NI_{simple} versus	$NI_{simple1}$ versus	$NI_{simple1}$ versus	DoS_1 versus	DoS_1 versus
Polymorphism	Divergence	P_n/P_s	P_{n2}/P_{s2}	P_{n1}/P_{s1}	$P_{n2}/(P_{n2} + P_{s2})$	$P_{n1}/(P_{n1} + P_{s1})$
Drosophila simulans	*D. yakuba*	0.75	0.36**	0.81	−0.01 NS	−0.38
D. melanogaster	Ancestor	0.56	0.12 NS	0.52	0.15 NS	−0.50
Escherichia coli	*Salmonella enterica*	0.91	0.21**	0.95	−0.04 NS	−0.81
S. enterica	*E. coli*	0.89	0.07 NS	0.95	−0.07 NS	−0.94
Human	Chimpanzee	0.73	0.20**	0.84	0.01 NS	−0.80

Note.—**$P < 0.01$; NS = not significant.

They therefore argue that NI < 1 is not symptomatic of adaptive evolution and that the MK framework for detecting adaptive evolution is misleading. However, there is nonindependence between NI and P_n/P_s and D_n/D_s, which is likely to set up correlations of the sort reported by Hughes et al. (2008) just through sampling error; that is, error in the estimate of P_n/P_s will induce a positive correlation between NI and P_n/P_s. This is easily demonstrated using the data from table 2: $\mathrm{NI}_{\mathrm{simple}}$ is strongly correlated to P_n/P_s in all data sets (table 4). The correlations would be judged to be highly significant under the null hypothesis that the two variables are independent, which they are not. However, we can remove the nonindependence by splitting P_n and P_s into two independent halves by randomly resampling them from a hypergeometric distribution (unfortunately the data set of Hughes et al. is not available in a form that allows this to be performed): That is, we randomly distribute P_s between two bins without replacement to yield P_{s1} and $P_{s2} = P_s - P_{s1}$. Note that this is similar to the method suggested by Piganeau and Eyre-Walker (2009), who suggested sampling from a binomial distribution; the binomial closely approximates the hypergeometric in most instances, but the hypergeometric is the correct distribution to use. We then calculate $\mathrm{NI}_{\mathrm{simple}}$ with one pair of P_n and P_s values and P_n/P_s with the other: for example, $\mathrm{NI}_{\mathrm{simple1}} = D_s\,P_{n1}/(D_n\,P_{s1})$ versus P_{n2}/P_{s2}. If we do this, then the correlation between $\mathrm{NI}_{\mathrm{simple}}$ and P_n/P_s becomes much weaker for all data sets (table 4); furthermore, we can show that this is not due to the decrease in the sample size because there is still a strong correlation between $\mathrm{NI}_{\mathrm{simple1}}$ and P_{n1}/P_{s1} (table 4). However, there is still a substantial correlation between $\mathrm{NI}_{\mathrm{simple1}}$ and P_{n2}/P_{s2} for some species. This might be due to a statistical bias because genes with little polymorphism will tend to show bias in both $\mathrm{NI}_{\mathrm{simple}}$ and P_n/P_s in the same direction or a genuine correlation between these two variables. To investigate this, we tested whether DoS_1 was correlated to $P_{n2}/(P_{n2} + P_{s2})$ and $P_{n1}/(P_{n1} + P_{s1})$. There is no significant correlation to the former, suggesting that any correlations between $\mathrm{NI}_{\mathrm{simple1}}$ and P_{n2}/P_{s2} are likely to be due to statistical bias. It is likewise possible to induce a negative correlation between D_n/D_s and NI.

Conclusion

Averaging NI values across genes and excluding genes from the analysis can result in biased overall estimates and may thus lead to incorrect inferences about the nature of selection. The solution is to use the weighted summary statistic $\mathrm{NI}_{\mathrm{TG}}$ that performs well whether NI is homogeneous or heterogeneous across genes. There are occasions when a point estimator of NI is required; for example, to test whether NI is correlated with some other quantity. If the numbers of substitutions and polymorphisms are sufficient, we recommend $\mathrm{LNI}_{\mathrm{Haldane}}$, when data are relatively plentiful because of its relatively symmetrical distribution. However, when data are sparse (i.e., any cell < 5), then we suggest that patterns of selection be investigated using DoS. We have implemented $\mathrm{NI}_{\mathrm{TG}}$, Woolf's test of homogeneity and DoS, within the Distribution of Fitness Effects (DoFE) package available at http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/.

Acknowledgments

We are very grateful to anonymous reviewers for constructive comments. N.S. and A.E.W. were supported by the Biotechnology and Biological Sciences Research Council.

References

Bartolomé C, Maside X, Yi SJ, Grant AL, Charlesworth B. 2005. Patterns of selection on synonymous and nonsynonymous variants in *Drosophila miranda*. *Genetics* 169:1495–1507.

Bazin E, Glémin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. *Science* 312:570–572.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519–520.

Begun DJ, Holloway AK, Stevens K, et al. (10 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol* 5:e310.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in Drosophila. *Mol Biol Evol*. 21:1350–1360.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (11 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Charlesworth B. 1993. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*. 63:213–227.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in bacteria. *Mol Biol Evol*. 23:1348–1356.

Cochran WG. 1954. Some methods for strengthening the common χ^2 tests. *Biometrics* 10:417–451.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.

Greenland S. 1982. Interpretation and estimation of summary ratios under heterogeneity. *Stat Med*. 1:217–227.

Haldane JBS. 1956. The estimation and significance of the logarithm of a ratio of frequencies. *Ann Hum Genet*. 20:309–311.

Hughes AL, Friedman R, Rivailler P, French JO. 2008. Synonymous and non-synoymous plymorphism versus divergence in bacterial genomes. *Mol Biol Evol*. 25:2199–2209.

Jenkins DL, Ortori CA, Brookfield JF. 1995. A test for adaptive change in DNA sequences controlling transcription. *Proc R Soc Lond Ser B Biol Sci*. 261:203–207.

Jewell NP. 1984. Small-sample bias of point estimators of the odds ratio from matched sets. *Biometrics* 40:421–435.

Jewell NP. 1986. On the bias of commonly used measures of association for 2 × 2 tables. *Biometrics* 42:351–358.

Jones MP, O'Gorman TW, Lemke JH, Woolson RF. 1989. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics* 45:171–181.

Li YF, Costello JC, Holloway AK, Hahn MW. 2008. "Reverse Ecology" and the power of population genomics. *Evolution* 62:2984–2994.

Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 22:719–748.

Maside X, Charlesworth B. 2007. Patterns of molecular variation and evolution in *Drosophila americana* and its relatives. *Genetics* 176:2293–2305.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.

Meiklejohn CD, Montooth KL, Rand DM. 2007. Positive and negative selection on the mitochondrial genome. *Trends Genet.* 23:259–263.

O'Gorman TW, Woolson RF, Jones MP, Lemke JH. 1990. Statistical analysis of 2 × 2 tables: a comparative study of estimators/test statistics for association and homogeneity. *Environ Health Perspect.* 87:102–107.

Paul SR, Donner A. 1992. Small sample performance of tests of homogeneity of odds ratios in k 2 × 2 tables. *Stat Med.* 11:159–165.

Piganeau G, Eyre-Walker A. 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS One.* 4:e4396.

Presgraves D. 2005. Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr Biol.* 15:1651–1656.

Rand DM, Kann A. 1996. Polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol Biol Evol.* 13:735–748.

Selvin S. 2004. Statistical analysis of epidemiologic data, 3rd ed. New York: Oxford University Press.

Shapiro JA, Huang W, Zhang C, et al. (8 co-authors). 2007. Adaptive genic evolution in the *Drosophila* genomes. *Proc Natl Acad Sci U S A.* 104:2271–2276.

Simpson EH. 1951. The interpretation of interaction in contingency tables. *J R Stat Soc Ser B.* 13:238–241.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with large effective population size. *Mol Biol Evol.* 27:1813–1821.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415:1022–1024.

Sokal RR, Rohlf SI. 1995. Biometry, 3rd ed. New York: WH Freeman and Company.

Tachida H. 2000. DNA evolution under weak selection. *Gene* 261:3–9.

Tarone RE. 1981. On summary estimators of relative risk. *J Chron Dis.* 34:463–468.

Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in Drosophila. *Genetics* 173:821–837.

Woolf B. 1955. On estimating the relation between blood group and disease. *Ann Hum Genet.* 1919:251–253.