Estimation of the Neutrality Index

Nina Stoletzki and Adam Eyre-Walker*
Centre for the Study of Evolution and School of Life Sciences
University of Sussex
Brighton
BN1 9QG
United Kingdom

* Correspondence: a.c.eyre-walker@sussex.ac.uk

Abstract
The McDonald-Kreitman test is a simple and widely used test of selection, in which the numbers of non-silent and silent substitutions ($D_n$ and $D_s$) are compared to the numbers of non-silent and silent polymorphisms ($P_n$ and $P_s$). The neutrality index ($NI = D_s P_n / D_n P_s$), 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 summarise 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 cut-off 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: $DoS = D_n/(D_n+D_s)-P_n/(P_n+P_s)$.

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 non-silent
(P\textsubscript{n}) and silent (P\textsubscript{s}) polymorphisms to the numbers of non-silent (D\textsubscript{n}) and silent (D\textsubscript{s}) substitutions per locus. When silent and non-silent substitutions are interspersed, and therefore share a common genealogy or genealogies and sampling scheme, one may perform a simple test of independence on the 2x2 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, \( P\textsubscript{n}/P\textsubscript{s} \) and \( D\textsubscript{n}/D\textsubscript{s} \), 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 \((P\textsubscript{n}/P\textsubscript{s})/(D\textsubscript{n}/D\textsubscript{s})\), however, the inverse of the NI has also been used (e.g. Tachida 2000, Presgraves 2005). Under the assumption that silent mutations are neutral, NI>1 indicates an excess of amino acid polymorphisms (as expected when there are slightly deleterious mutations), and NI<1 indicates an excess of non-silent 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 non-protein 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 \( D\textsubscript{n} \) or \( P\textsubscript{s} \) is zero. 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 for example Hughes et al. 2008, Meiklejohn et al. 2007, Bazin et al. 2006). 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 1/3rd 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 analysed $D_n$, $D_s$, $P_n$ and $P_s$ values from a number of studies: (i) 115 genes for which we have polymorphism data from Drosophila simulans and divergence between D. simulans and D. yakuba (Welch 2006); (ii) 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); (iii) 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 (iv) 11624 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, i.e. when there is no recombination, we randomly generated the total length of a genealogy by sampling from a set of exponential distributions; i.e. the time for $m$ branches to coalesce to $m-1$ branches is exponentially distributed with a mean of $\theta/(m(m-1))$, where $\theta$ is a constant. If we scale our trees so the average total length is one, then if the total length of a randomly generated tree is $\lambda$, the number of non-synonymous and synonymous polymorphisms are Poisson distribution with means of $\lambda E(P_n)$ and $\lambda E(P_s)$ 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, $NI_{True} = E(D_s) E(P_n) / (E(D_n) E(P_s))$, 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, $NI_{Simple}$ as
\[
E[NI_{\text{simple}}] = E[D_n] E[P_s] \sum_{x=0}^{\infty} Z(x,E[D_n]) \frac{1}{x} \sum_{x=1}^{\infty} Z(x,E[P_s]) \frac{1}{x} \sum_{x=1}^{\infty} Z(x,E[D_n]) \sum_{x=1}^{\infty} Z(x,E[P_s]) \]

where

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

is the Poisson distribution. Note the denominators in equation 1 can be simplified to \(1 - e^{-E[D_n]}\) and \(1 - e^{-E[P_s]}\) respectively.

We find that \(NI_{\text{Simple}}\) is substantially above or below \(NI_{\text{True}}\) 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 while 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 \(NI_{\text{Simple}}\) arises from two sources, first the exclusion of genes where NI is undefined, and secondly 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, \(NI_R\), estimated as \(E[D_n] E[P_s] / (E[D_n] E[P_s])\), when \(D_n > 0\) and \(P_s > 0\). \(NI_R\) 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 \(D_n/D_n\) or with undefined \(P_n/P_s\) excludes potentially large NI values. However, \(NI_{\text{Simple}}\) is greater than \(NI_R\) in each case, illustrating that even for genes for which NI is defined, \(NI_{\text{Simple}}\) 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 cut-off value $z$, including and below which values are excluded, is

$$E[f(x)] = \frac{\sum_{x\leq z} Z(x, E[D_n]) f(x)}{\sum_{x\leq z} Z(x, E[D_n])}$$

from which the bias can be quantified as

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

In the case of 1/x this means setting $z=0$ 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 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 co-varies 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 for example Haldane 1956, Jewell 1984, 1986, Cochran 1954, Mantel and Haenszel 1956). 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 Shapiro et al. 2007 for an example). A popular method of obtaining the mean OR is to use the Mantel-Haenszel (1959) procedure, which is a particular instance of a general method suggested by Cochran (1954; henceforth the CMH procedure), in which a weighted average OR is computed. The classic CMH procedure has previously been applied to the summary of MK tables (see for example Bartolome 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. Since 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):

\[
N_{ITG} = \frac{\sum D_{si} P_{ni}}{P_{si} + D_{si}} \frac{(P_{si} + D_{si})}{\sum P_{si} D_{ni}} \frac{(P_{si} + D_{si})}{P_{si} + D_{si}}
\]

(3)

where the index refers to i\(^{th}\) gene. To illustrate the power of this estimator to overcome the biases inherent in simply averaging NI values, we generated 100 datasets 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. \(N_{ITG}\) is almost exactly equal to, and not significantly different to \(N_{ITrue}\). 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 towards 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 since $NI_{True}$ and $E[P_s]$ are not generally correlated (Gossman and 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). Since 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 dataset into 4 groups, estimated $NI_{TG}$ for each and then calculated the mean. We repeated this 1000 times. We find that mean $NI_{TG}$ for the sub-samples are similar to $NI_{TG}$ for the whole dataset (Drosophila-Presgraves $NI_{TG}$ = 0.833; mean $NI_{TG(subsample)}$ = 0.905; Drosophila-Welch $NI_{TG}$ = 0.601; mean $NI_{TG(subsample)}$ = 0.614; Escherichia coli $NI_{TG}$ = 0.826, mean $NI_{TG(subsample)}$ = 0.827; and Salmonella enterica $NI_{TG}$ = 1.609, mean $NI_{TG(subsample)}$ = 1.615, 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 of $NI_{True}$.
As expected, the mean of NI across genes, excluding those with undefined values, \( N_{\text{Simple}} \) is larger than \( N_{\text{TG}} \) (table 2). This overestimate can be substantial and may be misleading (table 2). In both \textit{E. coli} and \textit{Drosophila}, \( N_{\text{TG}} \) 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 \textit{E. coli} and one of the \textit{Drosophila} datasets; the other \textit{Drosophila} dataset shows a neutral pattern. In the \textit{Salmonella} and hominin datasets, \( N_{\text{Simple}} \) is qualitatively similar to \( N_{\text{TG}} \) but it is much larger in value.

\textit{Adaptive evolution}

It is possible to estimate the proportion of non-synonymous substitutions that are adaptive, \( \alpha \), as 1-NI (Charlesworth 1993; Fay et al. 2001; Smith and Eyre-Walker 2002). Using \( N_{\text{TG}} \) to estimate \( \alpha \) has the advantage over the method of Fay et al. (2001) of avoiding Simpson’s paradox, and advantages over the 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 datasets 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 \( \alpha \), \( N_{\text{BEW}} \) (table 2), and using the method of Fay et al. (2001, \( N_{\text{FWW}} \)) 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 \( N_{\text{TG}} \). 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 \( N_{\text{TG}} \) and \( N_{\text{BEW}} \) (table 2).

\textit{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 odds ratios (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 datasets that we have analyzed, we detect significant heterogeneity in the *D. simulans* – *D. yakuba* dataset. For the other data sets used in this study, no significant heterogeneity can be detected; this is perhaps not surprising since tests of heterogeneity have little power, particularly when data is sparse, and the other datasets 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*- *yakuba* dataset is surprising given that Welch (2006) has previously tested this same dataset for heterogeneity in $\alpha$, the proportion of adaptive non-synonymous substitutions. Since $\alpha=1-\text{NI}$, this amounts to a test of heterogeneity in NI. The difference may be because Welch (2006) used a maximum likelihood (ML) method to estimate $\alpha$, and constrained $\alpha$ to be positive (i.e. 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*

$\text{NI}_{TG}$ 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 $\text{NI}_{\text{simple}}$ is not an appropriate statistic since 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 cut-off (figure 1). This occurs for the following reason. Imagine that we have a gene for which the expected value of $P_s$ is five (i.e. if we were to sample this gene repeatedly over a long period of time the mean value of $P_s$ would be five). If we apply a cut-off of five, many times the gene will have an observed value of $P_s$ that is lower than five and hence be excluded from
the analysis; only when the observed value is greater or equal to five 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

\[
NI_{Jewell} = \frac{D_s P_n}{(D_n + 1)(P_s + 1)} \tag{4}
\]

and

\[
LNI_{Haldane} = \log \left( \frac{(2D_s + 1)(2P_n + 1)}{(2D_n + 1)(2P_s + 1)} \right) \tag{5}
\]

A variant of Haldane’s estimator has also been employed, in which one is added to each cell in the 2x2 contingency table: i.e.

\[
LNI_{Laplace} = \log \left( \frac{(D_s + 1)(P_n + 1)}{(D_n + 1)(P_s + 1)} \right) \tag{6}
\]

(Presgraves 2005; Li et al. 2008; Slotte et al. 2010); since 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 of 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 \( P_s \), \( D_s \), \( P_n \), \( P_n \), 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 (figure 2). However, just as with $1/x$ (figure 1), excluding cases of $x$ below a certain cut-off does not help the bias in $1/(x+1)$; the bias simply shifts to higher expected values of $x$ (figure 2).

The biases in $LNI_{Haldane}$ and $LNI_{Laplace}$ are less easy to quantify since 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 datasets analysed 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 1000 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.

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 datasets, 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 (figure 3a). In hominids Log(NI) was initially overestimated and then underestimated as the level of polymorphism increased (figure 3b). In both cases 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:

\[ \text{DoS} = \frac{D_n}{D_n + D_s} - \frac{P_n}{P_n + P_s} \]  

(7)

This is the difference between the proportion of substitutions and polymorphisms that are non-synonymous. 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 since 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 (figure 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 non-silent and silent substitutions and polymorphisms per gene; the statistic is biased if the numbers are calculated per non-synonymous (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.

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 non-synonymous and synonymous substitutions along the *D.melanogaster* lineage, since *D.melanogaster* split from *D.simulans*, to numbers of non-
synonymous 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 ($r=0.23$, $p=0.033$)(figure 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 $D_n/D_s$. 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 non-independence 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; i.e. 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: NI$_{simple}$ is strongly correlated to $P_n/P_s$ in all datasets (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 non-independence by splitting $P_n$ and $P_s$ into two independent halves by randomly resampling them from a hypergeometric distribution (unfortunately the dataset
of Hughes et al. is not available in a form that allows this to be performed): i.e. we randomly distribute \( P_s \) between two bins without replacement to yield \( P_{s1} \) and \( P_{s2} = P_s - P_{s1} \). Note 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 \( N_{I_{simple}} \) with one pair of \( P_n \) and \( P_s \) values, and \( P_n \) and \( P_s \) with the other: e.g. \( N_{I_{simple1}} = D_s \frac{P_{n1}}{(D_n P_{s1})} \) versus \( P_{n2}/P_{s2} \). If we do this then the correlation between \( N_{I_{simple}} \) and \( P_n/P_s \) becomes much weaker for all datasets (table 4); furthermore we can show that this is not due to the decrease in the sample size since there is still a strong correlation between \( N_{I_{simple1}} \) and \( P_{n1}/P_{s1} \) (table 4). However, there is still a substantial correlation between \( N_{I_{simple1}} \) and \( P_{n2}/P_{s2} \) for some species. This might be due to a statistical bias, since genes with little polymorphism will tend to show bias in both \( N_{I_{simple}} \) 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 \( N_{I_{simple1}} \) 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 \( NI_{TG} \) 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 \( LNI_{Haldane} \), 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 \( NI_{TG} \), Woolf’s test of homogeneity and DoS, within the DoFE package available at www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/.
Acknowledgements
We are very grateful to anonymous reviewers for constructive comments. NS and AEW were supported by the BBSRC.

References


Charlesworth B. 1993. The effect of background selection against deleterious


Table 1. Example MK tables. Tables contain the expected numbers of $D_n$, $D_s$, $P_n$ and $P_s$. The mean of $N_{ITG}$ is given with the standard error; these were obtained by generating 100 datasets of 100 contingency tables using the expected numbers. The $N_{ITG}$ value in parentheses is for simulations in which the NI is lognormally distributed with a variance parameter of one.

<table>
<thead>
<tr>
<th>NI&lt;sub&gt;True&lt;/sub&gt; = 1</th>
<th>Poly.</th>
<th>Div.</th>
<th>NI&lt;sub&gt;Simple&lt;/sub&gt;</th>
<th>NI&lt;sub&gt;TG&lt;/sub&gt;</th>
<th>NI&lt;sub&gt;True&lt;/sub&gt; = 1.5</th>
<th>Poly.</th>
<th>Div.</th>
<th>NI&lt;sub&gt;Simple&lt;/sub&gt;</th>
<th>NI&lt;sub&gt;TG&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>silent</td>
<td>2</td>
<td>2</td>
<td>1.00±0.01</td>
<td>silent</td>
<td>4</td>
<td>4</td>
<td>0.98±0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>non</td>
<td>1</td>
<td>1</td>
<td>(1.06±0.03)</td>
<td>non</td>
<td>2</td>
<td>2</td>
<td>(1.03±0.02)</td>
<td></td>
</tr>
<tr>
<td>NI&lt;sub&gt;True&lt;/sub&gt; = 1.5</td>
<td>Poly.</td>
<td>Div.</td>
<td>NI&lt;sub&gt;Simple&lt;/sub&gt;</td>
<td>NI&lt;sub&gt;TG&lt;/sub&gt;</td>
<td>NI&lt;sub&gt;True&lt;/sub&gt; = 2.49</td>
<td>Poly.</td>
<td>Div.</td>
<td>NI&lt;sub&gt;Simple&lt;/sub&gt;</td>
<td>NI&lt;sub&gt;TG&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>silent</td>
<td>1</td>
<td>1.5</td>
<td>1.53±0.04</td>
<td>silent</td>
<td>5</td>
<td>7.5</td>
<td>1.51±0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>non</td>
<td>1</td>
<td>1</td>
<td>(1.48±0.04)</td>
<td>non</td>
<td>5</td>
<td>5</td>
<td>(1.51±0.02)</td>
<td></td>
</tr>
<tr>
<td>NI&lt;sub&gt;True&lt;/sub&gt; = 0.75</td>
<td>Poly.</td>
<td>Div.</td>
<td>NI&lt;sub&gt;Simple&lt;/sub&gt;</td>
<td>NI&lt;sub&gt;TG&lt;/sub&gt;</td>
<td>NI&lt;sub&gt;True&lt;/sub&gt; = 1.30</td>
<td>Poly.</td>
<td>Div.</td>
<td>NI&lt;sub&gt;Simple&lt;/sub&gt;</td>
<td>NI&lt;sub&gt;TG&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>silent</td>
<td>1</td>
<td>0.75</td>
<td>0.77±0.02</td>
<td>silent</td>
<td>4</td>
<td>3</td>
<td>0.73±0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>non</td>
<td>1</td>
<td>1</td>
<td>(0.79±0.03)</td>
<td>non</td>
<td>4</td>
<td>4</td>
<td>(0.78±0.01)</td>
<td></td>
</tr>
</tbody>
</table>