Improved Tests for Heterogeneity Across a Region of DNA Sequence in the Ratio of Polymorphism to Divergence

John H. McDonald
Department of Biology, University of Delaware

The neutral theory of molecular evolution predicts that the ratio of polymorphisms to fixed differences should be fairly uniform across a region of DNA sequence. Significant heterogeneity in this ratio can indicate the effects of balancing selection, selective sweeps, mildly deleterious mutations, or background selection. Comparing an observed heterogeneity statistic with simulations of the heterogeneity resulting from random phylogenetic and sampling variation provides a test of the statistical significance of the observed pattern. When simulated data sets containing heterogeneity in the polymorphism-to-divergence ratio are examined, different statistics are most powerful for detecting different patterns of heterogeneity. The number of runs is most powerful for detecting patterns containing several peaks of polymorphism; the Kolmogorov-Smirnov statistic is most powerful for detecting patterns in which one end of the gene has high polymorphism and the other end has low polymorphism; and a newly developed statistic, the mean sliding $G$ statistic, is most powerful for detecting patterns containing one or two peaks of polymorphism with reduced polymorphism on either side. Nine out of 27 genes from the Drosophila melanogaster subgroup exhibit heterogeneity that is significant under at least one of these three tests, with five of the nine remaining significant after a correction for multiple comparisons, suggesting that detectable evidence for the effects of some kind of selection is fairly common.

Introduction

The neutral theory of molecular evolution predicts that the amount of polymorphism in a region of DNA sequence will be correlated with the amount of DNA sequence divergence between species. Areas of low constraint will have both high polymorphism and high divergence, and areas of high constraint will have relatively little of each. The ratio of polymorphism to divergence should therefore be uniform across a region of DNA sequence. Due to recombination, some areas of sequence will have different phylogenies than others, which will cause some variation in the polymorphism-to-divergence ratio. There will also be sampling variation due to the finite number of sites. The expected phylogenetic and sampling variation can be estimated, and if there is greater heterogeneity in the polymorphism-to-divergence ratio than expected, it is evidence for selection (Hudson, Kreitman, and Aguade 1987).

There are a number of kinds of selection that can cause heterogeneity in the polymorphism-to-divergence ratio. The rapid fixation of an adaptive mutant can cause a selective sweep, reducing the neutral polymorphism near the adaptive site (Maynard Smith and Haigh 1974; Kaplan, Hudson, and Langley 1989). An old balanced polymorphism can accumulate neutral polymorphism near the balanced site (Strobeck 1983; Hudson, Kreitman, and Aguade 1987). Background selection due to selection against deleterious mutations may be able to reduce the polymorphism at linked neutral sites (Charlesworth, Morgan, and Charlesworth 1993). And mildly deleterious mutations, such as mutations from favored to unfavorable synonymous codons, may persist as polymorphisms but rarely go to fixation, causing the polymorphism-to-divergence ratio to be higher than for truly neutral mutations (Akashi 1995).

Because one of the goals of studies of DNA polymorphism and divergence is to detect evidence of natural selection, there is a need for a statistical test of heterogeneity in the polymorphism-to-divergence ratio. The traditional test is the HKA test (Hudson, Kreitman, and Aguade 1987). This test compares the polymorphism and divergence in discrete regions, and it is therefore ideal for comparing different genes. When used within a single gene, however, it requires that the gene be arbitrarily divided into two or more regions to be compared. There are many ways to subdivide a region of DNA sequence: two, three, four or more regions; regions equal in number of variable sites; regions equal in number of sites; or unequal regions based on exon/intron structure. Either the method of subdivision would have to be determined before collecting data (which would run the risk of missing interesting patterns), or all possible subdivisions would have to be tried (which would create problems of multiple comparisons). Choosing a method of subdivision after inspecting the data would, of course, be inappropriate. A test for heterogeneity which did not require any a priori decisions would therefore have advantages over the HKA test for single-gene studies.

I previously proposed a runs test for heterogeneity (McDonald 1996). In it, variable sites are classified as polymorphisms (sites that have two bases within a single species) or fixed differences (sites with one base in one species and a different base in the outgroup species). Heterogeneity is measured by the number of runs of polymorphic sites and fixed differences, where a “run” is a series of one or more sites of one kind preceded and followed by sites of the other kind. Greater heterogeneity leads to fewer runs. Phylogenetic and sampling variation under a neutral model are simulated using a coalescent algorithm, and the number of runs in the simulated data is compared with the observed number. The
statistical significance of the observed heterogeneity is the proportion of simulations with the number of runs equal to or less than that of the observed data. This does not require any a priori decision about the subdivision of the data. However, the runs statistic is generally considered to be statistically weak as a measure of heterogeneity.

There are other measures of heterogeneity in the distribution of binary states which may be more powerful than the runs statistic. The Kolmogorov-Smirnov statistic (Sokal and Rohlf 1981, pp. 719–720), like the runs statistic, is generally considered to be somewhat weak. Goss and Lewontin (1996) proposed several new statistics, of which the most powerful were the variance in interval length and a modified variance in interval length. These were designed to detect heterogeneity in the ratio of diverged to nondiverged sites for the identification of constrained regions in DNA or protein sequence data, a statistical problem similar to identifying heterogeneity in the polymorphism-to-divergence ratio. Here, I introduce two additional statistics. The maximum sliding G statistic, which is similar to the scan statistic with a variable window (Nagarwalla 1996), is found by sliding windows of all possible sizes across all possible positions in a series of polymorphic and diverged sites, calculating the log likelihood ratio statistic (G statistic; Sokal and Rohlf 1981, pp. 737–738) for the comparison of frequencies of polymorphisms and diverged sites inside and outside each window, and taking the maximum of these G statistics. I also introduce a similar statistic, the mean sliding G statistic, found by averaging the G statistics across all window positions for each window size and taking the largest of these mean G statistics. Here, I use simulated data sets containing simple patterns of heterogeneity to compare the power of these different statistics, and I apply each statistic to a number of published data sets from the *Drosophila melanogaster* subgroup.

**Materials and Methods**

A data set for the tests considered here consists of an ordered series of variable sites classified as either polymorphisms or fixed differences. Several DNA sequences from one species and one sequence from a closely related species are aligned, and variable sites are classified as polymorphisms and fixed differences as described in McDonald (1996). Areas aligned with a gap in any of the sequences are ignored. Polymorphisms are sites at which one species has two bases, while fixed differences are sites at which all the sequences from one species have a different base than the outgroup species. A site with three nucleotides in one species is treated as two adjacent polymorphisms. A site with two nucleotides in one species and a third nucleotide in the outgroup species is treated as one polymorphism and one fixed difference; to be conservative, these two sites are put in the order that maximizes the number of runs. This will usually be conservative for the other tests to be described, but if the results of a test are marginally significant, the effects of reversing the order of any such sites should be considered.

In this paper, heterogeneity in the distribution of polymorphic sites relative to fixed differences is measured using six different statistics.

**Runs statistic (K):** This is the number of runs of contiguous polymorphisms and contiguous fixed differences (Sokal and Rohlf 1981, p. 783; McDonald 1996).

**Kolmogorov-Smirnov statistic (D):** This statistic is based on the maximum absolute difference between the observed cumulative number of polymorphisms and the expected cumulative number of polymorphisms (Sokal and Rohlf 1981, pp. 719–720), modified for discrete data:

\[
D_K = \sup_{i=1,...,S} \left| \frac{S_{p(i)} - S_{p0}}{S} \right|
\]

where \(S_{p(i)}\) is the number of variable sites 1 through \(i\) that are polymorphisms; \(\hat{S}_{p(i)}\) is the expected number of variable sites 1 through \(i\) that are polymorphisms, \(\hat{S}_{p(i)} = i(S_p/S); S_p\) is the total number of polymorphisms; and \(S\) is the total number of variable sites.

**Interval length variance (V):** This is the variance in the length of intervals between members of one class of sites (Goss and Lewontin 1996, eq. 2). It is calculated for polymorphisms or fixed differences, whichever are less frequent. For example, if there are fewer polymorphisms than fixed differences, adjacent polymorphic sites have an interval length of one, polymorphisms separated by one fixed difference have an interval length of two, and so on.

**Modified interval length variance (Q):** This statistic, also proposed by Goss and Lewontin (1996, eq. 3), is similar to the interval length variance but includes a term for the covariance between consecutive interval lengths.

**Maximum sliding G statistic (G_max):** This statistic is based on the scan statistic with variable window (Nagarwalla 1996). A window of size \(n\) is placed at the beginning of the list of \(S\) variable sites, and the log likelihood ratio statistic (G statistic, Sokal and Rohlf 1981, pp. 737–738) is calculated for the \(2 \times 2\) table comparing the frequencies of polymorphisms and fixed differences inside and outside the window. This is repeated for every possible window position and every possible window size from \(n_{\text{min}}\) to \(n_{\text{max}}\). \(n_{\text{min}}\) is the smallest window size with an expected number of polymorphisms (and expected number of fixed differences) greater than or equal to five, an arbitrary number based on the usual rule of thumb for G-tests (Sokal and Rohlf 1981, p. 709); where the number of polymorphisms or fixed differences is less than 11, \(n_{\text{min}}\) is the smallest window with an expected number greater than or equal to three. \(n_{\text{max}}\) is \(S - n_{\text{max}}\). The largest G statistic from all the window sizes and positions is the maximum sliding G statistic.

**Mean sliding G statistic (G_mean):** This statistic is similar to the maximum sliding G statistic, except that the mean G statistic across all window positions is cal-
culated for each window size. The largest of these means is the mean sliding $G$ statistic.

To test the statistical significance of each statistic, repeated Monte Carlo simulations of a coalescent model incorporating recombination and both sampling and phylogenetic variation were conducted using parameters estimated from the data and an arbitrary value for the recombination parameter ($R$). A full description of the coalescent algorithm used to generate the simulated data sets, which is based on the work of Hudson (1983, 1990), has been presented earlier (McDonald 1996).

Each of the six test statistics was calculated for each simulated data set. The proportion of simulated data sets with a test statistic equal to or more extreme than the observed test statistic is the estimated probability of getting a test statistic that extreme under a model with that recombination parameter. To be conservative, different values of the recombination parameter are tried, starting with 1 and increasing until $P$ starts to decrease. The largest $P$ is used as the statistical significance of the test statistic.

A Macintosh program that will perform these tests, DNA Slider, is available by anonymous FTP from ftp.bio.indiana.edu in the directory /molbio/mac. It is also available from the author.

The power of the six tests was compared by applying them to simulated data sets containing varying amounts of heterogeneity. A very simple model of heterogeneity in the polymorphism-to-divergence ratio was used to generate these data sets, unlike the coalescent model used in the statistical tests themselves. Each simulated data set contained $S$ variable sites. The fixed differences were distributed randomly with uniform probability density across a sequence of 1,000 sites, while the polymorphisms were distributed randomly with two densities, $ZD$ in part of the sequence and $D$ in the rest. The sequence therefore had one or more blocks of high polymorphism and one or more blocks of low polymorphism (fig. 1). The fraction of the sequence occupied by blocks of elevated polymorphism was $W$, with $1 - W$ of the sequence having reduced polymorphism. For each value of $S$, the 5% critical value for each of the six test statistics was estimated based on 10,000 simulations with uniform densities of polymorphisms ($Z = 1$). Then, for each combination of peak pattern and value of $W$, 1,000 simulations were run for values of $Z$ of 2, 4, 6, $\ldots$, 16. For each test statistic, the number of simulations in which the test statistic was more extreme than its critical value (and therefore detected significant heterogeneity) was counted. The value of $Z$ at which a test would detect a significant deviation from homogeneity 50% of the time was estimated by linear interpolation, and this value was used as an estimate of the power of the test. A lower value of $Z$ means a test could detect a smaller deviation from homogeneity.

The six tests were also compared by applying them to published data on polymorphism and divergence in the *Drosophila melanogaster* subgroup (table 1). Only variable sites that were silent (synonymous or noncoding) were used, and only regions with at least seven polymorphisms and seven fixed differences were analyzed. When there were multiple sequences from both species, the outgroup sequence chosen by the original authors was used if there was one; otherwise, the outgroup sequence adjacent to the other species’ sequences in the published figure was used. For most data sets, the closely related species *D. melanogaster* and *D. simulans* were used as outgroups for each other; for G6pd in these two species, the more distantly related *D. yakuba* was also used in order to increase the number of diverged sites (Eanes et al. 1996). Simulations were run with $R = 1, 2, 4, 8, \ldots$ until the probability values for all tests had either gone above 0.2 or started to decline. The value of $R$ had a fairly small effect on $P$, and the maximum $P$ value was generally found with $R = 4$ to $R = 32$. For each value of $R$, 1,000 simulations were run; when one of the $P$ values was less than 0.10, an independent set of 10,000 simulations was run to obtain a more accurate estimate of the probability.

### Results

The power of the six tests was compared using simulated data sets with varying amounts and patterns of heterogeneity in the probability density of polymorphisms. Simulations were done with equal numbers of polymorphisms and fixed differences for 40, 80, and 160 total variable sites and for 160 total variable sites with 20, 40, and 60 polymorphisms. The results from simulations with unequal numbers of polymorphisms and fixed differences are quite similar to the results with equal numbers, so only the latter are shown. Total widths for the blocks of elevated polymorphism of $W = 0.1, 0.2, 0.3, 0.4, \text{ and } 0.5$ were simulated, but since the results for 0.2, 0.3, and 0.4 were intermediate to those with 0.1 or 0.5, only the results from these two values of $W$ are shown.

A region divided into two areas, one with high polymorphism and one with low polymorphism (fig. 1), was used.
Three peaks for each set of parameters is shown in bold. Table 2 contains the power values for the statistics to detect patterns of heterogeneity using one or more of the test statistics (table 2). Three additional data sets are almost significant (0.05 < P < 0.07); because the maximum P value from trials with different values of the recombination parameter was conservatively used, these almost-significant results may be worthy of some attention. Five data sets have significant heterogeneity with the mean sliding G statistic and two more are almost significant, and for all seven of these, the mean sliding G statistic has a lower P value than the maximum sliding G statistic. This suggests that the relatively narrow peaks for which the maximum sliding G statistic is more powerful in simulations are not common in these data sets. Sliding-window graphs indicate that the five data sets with significant mean sliding G tests have one or two peaks of elevated polymorphism-to-divergence ratios (fig. 2c–e), which is the pattern that the simulations indicated this test would be the most powerful at detecting.

Two data sets exhibit their lowest significant P values with the runs statistic, and two more data sets are almost significant with the runs statistic but not close to significance with any other statistic (table 2). These data sets have multiple peaks of elevated polymorphism-to-divergence ratios (fig. 2f and g). Two data sets exhibit their lowest significant P value with the Kolmogorov-Smirnov statistic, and both have a difference between
the first and second half of the region in the polymorphism-to-divergence ratio (fig. 2h and i). These results suggest that the runs statistic and the Kolmogorov-Smirnov statistic detect patterns of heterogeneity that are missed by the sliding $G$ statistics. The interval length variance and modified interval length variance do not detect any significant results that are not detected with a lower $P$ value by one of the other statistics, which is consistent with their lack of power in the simulated data sets.

**Discussion**

Based on the results from both the simulated data sets and the real data sets, three measures of heterogeneity appear to be useful: the mean sliding $G$ statistic, the runs statistic, and the Kolmogorov-Smirnov statistic. Each of these three is the most powerful at detecting a different pattern of heterogeneity. The mean sliding $G$ statistic is best at detecting one or two peaks of elevated polymorphism, the runs statistic is best at detecting multiple peaks of polymorphism, and the Kolmogorov-Smirnov statistic is best at detecting a single change across a gene from low to high polymorphism. Although the simulations indicate that the maximum sliding $G$ statistic is slightly more powerful than the mean sliding $G$ statistic under some conditions, these conditions do not appear to be common in *Drosophila* data sets. The interval length variance and modified interval length variance do not appear to be useful for detecting heterogeneity in the polymorphism-to-divergence ratio; these statistics may still be useful for detecting heterogeneity in the ratio of diverged to uniform sites, their original purpose (Goss and Lewontin 1996).

The existence of three useful tests for heterogeneity raises the problem of multiple comparisons. When all three tests are used and the significance level of each test is set to 0.05, the probability of rejecting the null hypothesis when it is true with at least one of the three tests is considerably greater than 0.05. One solution would be to use only one test and ignore the results from the rest. The mean sliding $G$ statistic is the most powerful overall in both the simulated and real data sets, so it would be the likely choice if only one test was used. A disadvantage of this approach is that some patterns of heterogeneity that are quite significant with the runs or Kolmogorov-Smirnov tests, such as the multiple peaks in *Adh/Adhr* in *D. melanogaster* or the shift from low to high polymorphism in *G6pd* in *D. simulans,*
Fig. 2.—Sliding-window plots of the polymorphism-to-divergence ratio in data sets with significant heterogeneity. For each window of 15 variable sites (10 variable sites for *yp2*), the proportion of variable sites that were polymorphic was calculated. Each value is plotted at the nucleotide position midway between the endpoints of the window; each point therefore represents the polymorphism-to-divergence ratio for a fairly large region. Shaded bars indicate the positions of coding regions, lines indicate the positions of noncoding regions. Vertical lines above the gene diagram indicate the positions of amino acid polymorphisms; vertical lines below the gene diagram indicate the positions of fixed amino acid differences between species. For *Acp26Aa*, *G6pd*, and *Adh + Adhr*, sequences from *D. yakuba* or *D. teissieri* were used to determine which amino acid substitutions became fixed in the lineage of the species shown; only these substitutions are indicated. For the other genes, all fixed amino acid differences are shown. The nucleotide numbering systems shown are those used in the original publications.
Heterogeneity in Polymorphism-to-Divergence Ratio 383

would not be detected with the mean sliding \( G \) statistic. Another possibility would be to apply the Bonferroni correction to the significance level of each test, so that an individual \( P \) value would have to be less than 0.017 (0.05/3) in order to have an overall significance of 0.05. If this correction is applied to the data on Drosophila (table 2), only five of the nine data sets with individual \( P \) values less than 0.05 would be significant. A third possibility would be to consider each of the three test statistics to be an independent hypothesis test, since each one is best at detecting a different alternative hypothesis, and use this as a rationalization for using the individual significance values of 0.05. In the interest of exploring a hard-won data set to the maximum extent possible, this last approach seems reasonable; it should be acknowledged, however, that probabilities less than 0.05 but greater than 0.017 are statistically somewhat dubious rejections of the null hypothesis.

The model of sequence evolution underlying these tests assumes a uniform recombination rate across the region, random union of gametes (no geographic structure), and constant population size. Violation of these assumptions may be one cause of heterogeneity in the polymorphism-to-divergence ratio, although the conditions under which this could occur do not seem likely (McDonald 1996). Perhaps the most likely nonselective explanation of significant heterogeneity would be nonpoint mutations. When comparing different species, it is not uncommon to see what appears to be the replacement of two or more adjacent bases by a different set of bases. If this is counted as several variable sites when it is actually the result of a single mutation, the heterogeneity could be artificially inflated. One solution for data sets where this appears to be a problem would be to count variable sites that are immediately adjacent to each other as if they were a single variable site.

Once significant heterogeneity has been detected, the next step is to examine the data to see which kinds of selection might be responsible. Balancing selection on an old balanced polymorphism can result in a region of elevated silent polymorphism around the balanced polymorphism. The AdhIAdh region in D. melanogaster has a well-known peak of polymorphism around the F/S amino acid polymorphism (fig. 2f), which extensive evidence indicates is affected by selection (Berry and Kreitman 1993; Stam and Laurie 1996). G6pd in D. melanogaster has elevated silent polymorphism around the A/B amino acid polymorphism (fig. 2h), which might suggest that it also is an old balanced polymorphism.

Selective sweeps are suggested when areas of low polymorphism are centered on amino acid substitutions that are fixed between species. This is especially suggestive if data are available for a third species that is an outgroup to the two being compared; in that case, it can be determined which lineage the amino acid substitutions occurred in. The best example of this pattern is G6pd in D. simulans, where several amino acid substitutions have occurred in the D. simulans lineage near the S' end of the gene, an area of low silent polymorphism (fig. 2i). Many of the amino acid differences between D. simulans and D. melanogaster are adaptive (Eanes, Kirchner, and Yoon 1993), so they are possible causes of a selective sweep. However, it should be noted that D. melanogaster and D. simulans have similar patterns of low polymorphism near the 5' end and higher polymorphism near the 3' end of the gene. While balancing selection on the A/B polymorphism in D. melanogaster and selective sweeps caused by the many amino acid fixations in D. simulans are plausible explanations for the patterns in each species, the similarity between the patterns suggests that there might be some other cause, one that is the same in both species.

There is increasing evidence that weak selection occurs against unfavored synonymous codons in Drosophila (Akashi 1995, 1996). Because mildly deleterious mutations can persist as polymorphisms but are unlikely to become fixed, coding regions with weak selection on synonymous sites could have a higher polymorphism-to-divergence ratio than noncoding regions with truly neutral variation. This pattern is not obvious in any of the data sets examined here (fig. 2). However, there are tests that are specifically designed to look for evidence of mild selection on synonymous codons (Akashi 1995, 1996); these tests, or an HKA test comparing all coding regions combined versus all noncoding regions, would be more powerful than simply examining sliding-window plots for visual evidence of a difference between coding and noncoding regions.

Background selection is the reduction in silent polymorphism caused by the selective elimination of deleterious mutations at linked sites (Charlesworth, Morgan, and Charlesworth 1993). It is not clear whether background selection could cause heterogeneity on the small scale of a single gene. Since most deleterious mutations are likely to be in coding regions, a pattern of higher polymorphism-to-divergence ratios in noncoding regions than in coding regions could suggest background selection as a possible cause. As with mildly deleterious selection due to codon bias, an HKA test comparing coding versus noncoding regions would be an informative follow-up to the detection of significant heterogeneity.

Acknowledgments

I thank B. C. Verrelli for many fruitful discussions and S. A. Sawyer for suggesting the Kolmogorov-Smirnov statistic. This research is supported by NSF grant number DEB-9317949.

LITERATURE CITED


STANLEY A. SAWYER, reviewing editor

Accepted December 3, 1997

Downloaded from https://academic.oup.com/mbe/article-abstract/15/4/377/1021388 by guest on 16 December 2018

384 McDonald