

Note

Effect of Misoriented Sites on Neutrality Tests With Outgroup

Emmanuelle Baudry and Frantz Depaulis¹

Laboratoire d'Ecologie, Centre National de la Recherche Scientifique UMR 7625-EPHE,
Université Pierre et Marie Curie, 75252 Paris Cedex 05, France

Manuscript received June 18, 2003
Accepted for publication July 31, 2003

ABSTRACT

Several neutrality tests use outgroups to infer the ancestral and derived states for polymorphism data. However, homoplasy can result in the incorrect inference of the derived variant. We show that empirically derived rates of misorientation strongly influence Fay and Wu's H -test, especially when the sample size is large.

INTRASPECIFIC polymorphism data are usually analyzed within the framework of the neutral Wright-Fisher model, which (besides the absence of selection) assumes a randomly mating population of constant size. Departures from this model are typically attributed to selective or demographic effects. A rarely considered alternative explanation is that the mutational process can cause or at least contribute to such departures (see, however, ROGERS 1992). Indeed, the neutral model can be applied to various mutational models. For sequence data, the expectations are usually derived from the infinite sites model (KIMURA 1969; WATTERSON 1975). This model assumes that each mutation occurs on a different site; *i.e.*, homoplasy is ignored. Numerous statistics have been developed to test whether an observed pattern of polymorphism is expected under the infinite sites neutral model (reviewed in WALL 1999; DEPAULIS *et al.* 2004). Among these statistics, several use an outgroup to infer the derived and ancestral states of polymorphic sites (FU and LI 1993; FAY and WU 2000). These statistics are potentially sensitive to the presence of multiple hits, due to the long branch leading to the outgroup where mutations can be superimposed. In this article, we estimate the level of homoplasy (including parallelisms and reversions) in empirical data sets and we use coalescent simulations to assess the influence of these homoplasy levels on the tests with outgroups.

Three neutrality tests with an outgroup have been proposed. Each one compares a pair of unbiased estimates of the mutational parameter of the population ($\theta = 4N_e\mu$ for an autosomal marker). FU and LI's (1993) D - and F -tests rely on standardized statistics:

$$D = \frac{\theta_w - \eta_e}{\hat{\sigma}(\theta_w - \eta_e)} \quad \text{and} \quad F = \frac{\pi - \eta_e}{\hat{\sigma}(\pi - \eta_e)}. \quad (1)$$

They compare WATTERSON'S (1975) θ_w estimator, which is based on the total number of polymorphic sites, and TAJIMA'S (1983) diversity π (respectively) to η_e , the number of derived unique mutations (mutations on external branches of the tree). They are, thus, highly sensitive to the relative proportion of the latter mutations. FAY and WU'S (2000) H -test,

$$H = \pi - \theta_H, \quad (2)$$

compares π to θ_H , an estimator weighted by the homozygosity of the derived variants. It is, thus, primarily sensitive to the relative proportion of high-frequency-derived variants. The H -test was designed to specifically detect positive selection in the presence of recombination (with recombination occurring between the region surveyed and the selected site during the selective stage). Since its introduction, an unexpectedly large number of significant H values have been reported in humans and *Drosophila* (PRZEWSKI 2002). PRZEWSKI (2002) suggested that this excess could be caused by population structure or by our incorrect modeling of the way positive selection operates. Alternatively, incorrect modeling of the mutational process could contribute to the observed departure from the neutral model.

In practice, an outgroup is used to identify the derived and ancestral variants of a polymorphic site. This inference can be incorrect if an undetected second mutation occurred at the same site on the outgroup branch, *i.e.*, if multiple hits are present. FAY and WU (2000) considered this possibility and computed the probability of misorientation, assuming a constant mutation rate and a finite site model. They proposed to incorporate this rate in the null hypothesis, but this is virtually never applied in practice. Furthermore, substitution rates usu-

¹Corresponding author: Université Pierre et Marie Curie, Laboratoire d'Ecologie, CNRS UMR 7625-EPHE, case 237, bat A, 7 quai St. Bernard, 75252 Paris Cedex 05, France. E-mail: fdepaulis@snv.jussieu.fr

ally vary among sites, suggesting a similar variation in neutral mutation rates (*e.g.*, NACHMAN and CROWELL 2000). Such heterogeneity of neutral mutation rates should substantially enhance the frequency of secondary mutation on variable sites (*e.g.*, the extreme case of mutation hot spots) and the corresponding rate of misorientations. This heterogeneity is, however, difficult to estimate precisely. In the absence of more information, a gamma distribution is generally assumed for substitution rates (*e.g.*, GU and ZHANG 1997). To avoid making assumptions about the distribution of mutations across sites, we used an empirically derived estimator of P_M , the average probability of misorientation per site. We first estimated P_D , the probability that a polymorphic site shows a detected second mutation on the outgroup branch, by determining the proportion of polymorphic sites for which the outgroup shows a third state. If all mutations were equally likely, then the rate of (undetected) misinferred sites P_M would just be half of P_D . However, in all genomes examined, transitions occur at a higher rate than transversions (YANG and YODER 1999), thereby increasing the rate of undetected misinferred sites. Let α be the rate of each possible transition and β that of transversions (Kimura's two-parameter model; KIMURA 1980). We computed the relative rates of undetected over detected multiple hits (two-state *vs.* three-state sites), given that a second mutation occurred on the outgroup branch. We neglect the possibility of more than two mutations.

If the first mutation, which produces the polymorphic site, is a transition (probability α), then the second mutation on the outgroup branch is undetected if it is also a transition (probability α) and it is detected if it is a transversion (probability 2β since there are twice as many possible transversions as transitions). If the first mutation is a transversion (probability 2β), the second mutation is undetected only if it is the same type of transversion (probability β) and it is detected if it involves the other type of transversion or a transition (probability $\alpha + \beta$). Hence, the ratio of undetected to detected multiple hits is

$$\frac{P_M}{P_D} = \frac{\alpha \times \alpha + 2\beta \times \beta}{\alpha \times 2\beta + 2\beta(\alpha + \beta)} \quad \text{or} \quad P_M = \frac{\alpha^2 + 2\beta^2}{2\beta(2\alpha + \beta)} P_D. \quad (3)$$

In practice, we simply estimated α and β for each data set by counting the proportions of sites with transitional and transversional differences between the sequences including all polymorphisms and fixed differences. Given the relatively low level of divergence between the sequences considered here (<9%, see below), this should provide a reasonably accurate approximation (YANG and YODER 1999). Refinements of (3) are possible to correct the corresponding estimator. However, simulations suggest that, for instance, the bias introduced by the ratio is negligible compared to the main source of imprecision lying in the generally low (or null) values of P_D (results not shown).

We then estimated the probability of misorientation

in data sets from three species that are frequently the focus of sequence polymorphism studies: *Homo sapiens*, *Drosophila simulans*, and *Arabidopsis thaliana*. Patterns of polymorphism in a population can be affected by several factors, including demographic history. To minimize this effect, we chose surveys of loci with comparable sampling schemes within a species. FRISSE *et al.* (2001) analyzed 10 noncoding regions in three human populations and used chimpanzee as an outgroup. BEGUN and WHITLEY (2000) collected data from 29 loci in a North American population of *D. simulans* (outgroup *D. melanogaster*). Finally, we used polymorphism data from 12 loci of *A. thaliana* (outgroup *A. lyrata* or *A. gemmifera*). This data set was collected by different researchers, but very similar samples composed of worldwide *A. thaliana* ecotypes were used in all studies. In the 10 human data sets, we did not observe any site with three states. Our estimate of P_M is therefore 0 for all loci (P_M is likely to be a poor estimator for low P_D values). In the *D. simulans* data sets, P_M varies between 0 and 10.6% (mean 3.3%). This is in agreement with INNAN and TAJIMA's (1997) suggestion that sites with more than one mutation are likely to be present in comparison between *D. melanogaster* and *D. simulans*. In the *A. thaliana* data set, P_M ranges from 0 to 19.1% (mean 5.1%). These differences in the average probability of misorientation probably reflect the difference in divergence between the studied species and the outgroup, which are, respectively, 1.1, 2.6, and 8.8% for human, *D. simulans*, and *A. thaliana* (average divergence considering all sites, with JUKES and CANTOR 1969 correction). The average estimates of probability of misorientation that we observed in *D. simulans* and *A. thaliana* are higher than that obtained following FAY and WU's (2000) procedure (3.3 *vs.* 0.9 and 4.7 *vs.* 2.9% in the two species, respectively). The difference probably reflects heterogeneity of the neutral mutation rates across sites. In contrast to Fay and Wu's approach, our model implicitly takes this heterogeneity into account. Such heterogeneity should explain the relatively high P_D rates observed. A significant negative correlation was observed between H and P_M in *A. thaliana* (Pearson's $R^2 = 0.33$, $P = 0.047$), suggesting that H is influenced by homoplasy in these data sets (Figure 1). We did not observe such a correlation in the *D. simulans* data sets, possibly because of the lower average value of P_M or because of small sample sizes of these data (average = 7.3) leading to reduced power. No correlation was observed in any data set between FU and LI's (1993) tests and P_M .

To study the effect of misorientation ranging from 0 to 20% (the observed range in the analyzed data sets) on tests with an outgroup, we used genealogies generated by a standard coalescent algorithm (HUDSON 1993). First, we generated the critical test values (5% significance level, two-sided for F and D , one-sided for H as in the original procedures) by simulations of the standard neutral model with a fixed number of segregating sites, no misorientation, and no recombination. Second, we

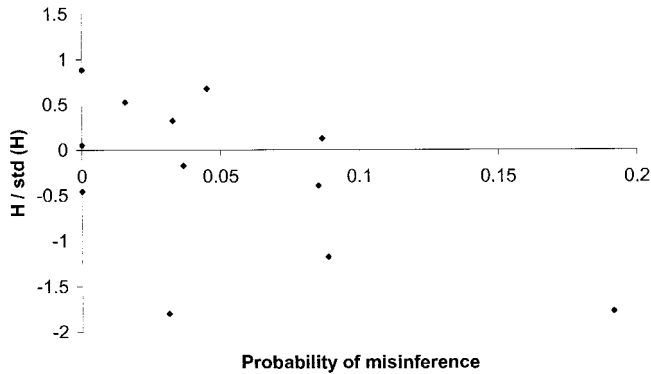


FIGURE 1.—Standardized Fay and Wu's H -statistics against probability of misorientation of the derived polymorphic state for 12 *A. thaliana* genes. The standardization allows comparisons between genes, as the variance of H increases with the number of segregating sites in a sample. Pearson's $R^2 = 0.33$, $P = 0.047$. Polymorphism data are from KAWABE and MIYASHITA (1999), KUITTINEN and AGUADE (2000), SAVOLAINEN *et al.* (2000), AGUADE (2001), MIYASHIYA (2001), and OLSEN *et al.* (2002).

simulated data sets in the presence of misorientation by exchanging the frequency of the derived and ancestral variant, with probability given by the misorientation parameter for each segregating site. This implicitly assumes that the rate of multiple hits is independent of the polymorphism frequency, as seems reasonable if such mutations principally occur on the branch connecting the root of the intraspecific tree to the outgroup. The proportion of simulations with a value of the statistics more extreme than the critical value(s) was recorded. Fifty thousand simulations were performed for each combination of parameter values.

The F - and D -statistics were found to be minimally affected by the presence of homoplasy (Figure 2). With levels of misorientation up to 20%, the null model is rejected <8% of the time for all sets of parameter values (sample sizes and numbers of segregating sites each ranging from 10 to 100; results not shown). On the contrary, H is very sensitive to the presence of homoplasy. For example, if $P_M = 0.15$ with a sample size of 50, the null model will be rejected $\sim 25\%$ of the time. The effect markedly increases with sample size (Figure 2), but is virtually insensitive to the number of segregating sites (results not shown). The difference of susceptibility of F - and D -tests *vs.* the H -test can be understood by considering the unfolded frequency spectrum of polymorphic sites. Unique variants are frequent under the neutral model, while high-frequency-derived variants are very scarce (Figure 3). Thus, in most cases, homoplasies transform a unique variant into a high-frequency one, which produces a large excess of such variants (Figure 3). Misorientation therefore strongly affects the H -test, which is designed precisely to detect an excess of high-frequency-derived variants (FAY and WU 2000). On the other hand, Fu and Li's tests are primarily sensitive to unique external mutations, which

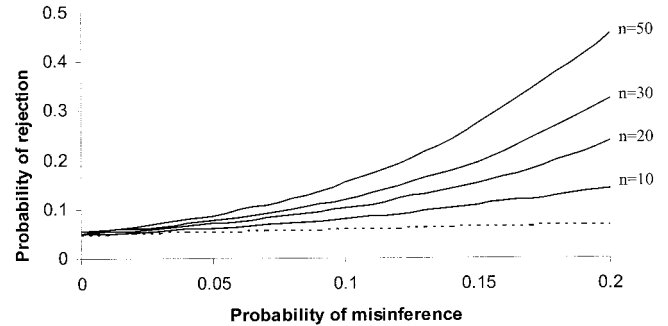


FIGURE 2.—Probability of rejection of the null model against rate of misorientation of the derived polymorphic state for Fay and Wu's H -test (solid lines). The number of segregating sites is $S = 50$, the sample sizes range from 10 to 50. Very similar results are obtained for values of S varying between 10 and 100 (results not shown). For each data point, 50,000 coalescent simulations were run. For comparison, the dashed curve indicates equivalent results for Fu and Li's D (all curves for Fu and Li's statistics were superimposed).

are diminished by only a fraction when homoplasy occurs at a reasonable rate.

Our results suggest that, in species where a relatively divergent outgroup is commonly used, like *D. simulans* or *A. thaliana*, misorientation of the derived state of variants can produce significant values of the H -test with appreciable frequency. This effect is likely to be underestimated in our study since more distant outgroups are sometimes used (*e.g.*, *D. yakuba* for *D. melanogaster* or gorilla for humans). On the contrary, the use of a very closely related outgroup could lead to yet another source of misorientation due to the occurrence of ancestral polymorphisms. Using several outgroups could potentially help. (To our knowledge, this is not done in practice when applying the test.) It does not, however, fully solve the problem since adding more outgroups, especially more distant ones, increases the probability of getting multiple hits. Our approximation that neglects higher-order mutations then becomes inappropriate. Using several outgroups can reduce the fraction of misoriented sites, but would correspondingly increase the number of sites that cannot be oriented. These sites would have to be removed from the analyses, thereby reducing the power of the test. The issue becomes a trade-off between power and robustness to homoplasy. Finally, two outgroups are typically far from being independent: a large part of the lineage linking an outgroup to the intraspecific tree is generally shared between two outgroups, thus providing little additional information. If several outgroups are to be used, our approach can still be applied by replacing the outgroups with an estimate of the ancestral sequence on the node that links the outgroups to the intraspecific tree, *e.g.*, with maximum-likelihood methods (YANG *et al.* 1995).

Any additional mutational bias (*e.g.*, base frequency heterogeneity) should increase the undetected over detected multiple-hit ratio. This is the case particularly for

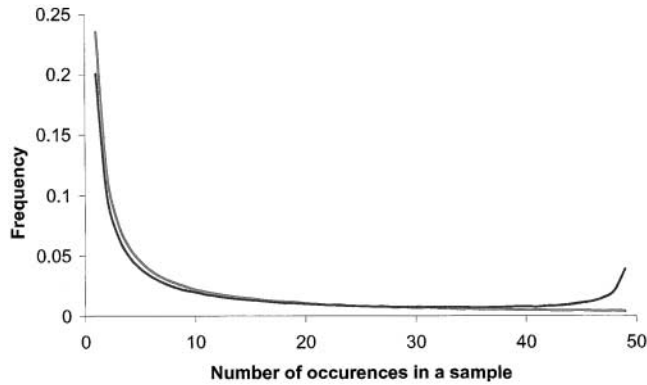


FIGURE 3.—Unfolded frequency spectrum expected under neutrality when misorientation is absent (gray line) or when the probability of misorientation is 15% (black line). The sample size is 50 and the number of segregating sites is 50.

data from coding regions where twofold degenerate sites tend to show high transition:transversion bias since only transitions are synonymous. This is taken into account in our average estimate of transition:transversion ratio, but we implicitly assume that this ratio is constant over sites, an assumption that may induce a slight bias. For data in coding regions, a rough correction can be performed by removing all twofold degenerate sites before estimating α , β , and P_D and deriving P_M with (3). The assumption of constant α over β ratio between replacement and silent polymorphisms becomes more realistic once those sites are removed. Then P_M should be corrected by the factor $L/(L - L_{2X})$, with L the total length of the sequence and L_{2X} the number of twofold degenerate sites. The rationale is that twofold degenerate sites cannot lead to detected multiple hits for synonymous polymorphisms.

In the above discussion we considered only misorientations caused by homoplasy effects but misorientations could also result from other type of biases such as poor alignment with the outgroup sequences when indels are frequent. Finally, in the presence of recombination, the use of critical values for the case of no recombination is overly conservative at the expense of a drastic reduction in the power of the tests (WALL 1999). The homoplasy effect is likely to be much stronger when the tests are applied with recombination and thus have much tighter confidence intervals. More generally, we conclude that when a significant H -test is observed, misorientation effects should be considered if the test is to be applied in the standard way. Alternatively, a misorientation rate should be incorporated in the null hypothesis with estimators more conservative than those proposed in the original procedure (FAY and WU 2000).

We thank A. Di Rienzo and L. Frisse for providing data sets and D. Begun, D. Carlini, E. Heyer, H. Innan, C. Müller-Graf, and anonymous reviewers for comments on the manuscript. E.B. is supported by a grant from the Ecole Pratique des Hautes Etudes and F.D. by a grant from the Centre National de la Recherche Scientifique.

LITERATURE CITED

- AGUADE, M., 2001 Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes, in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **18**: 1–9.
- BEGUN, D., and P. WHITLEY, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. *Proc. Natl. Acad. Sci. USA* **97**: 5960–5965.
- DEPAULIS, F., S. MOUSSET and M. VEUILLE, 2004 Powers of neutrality tests against bottleneck and hitchhiking. *J. Mol. Evol.* (in press).
- FAY, J. C., and C.-I. WU, 2000 Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.
- FRISSE, L., R. R. HUDSON, A. BARTOSZEWICZ, J. D. WALL, J. DONFACK *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**: 831–843.
- FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- GU, X., and J. ZHANG, 1997 A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* **14**: 1106–1113.
- HUDSON, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanism of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Japan Scientific Societies Press/Sinauer Associates, Sunderland, MA.
- INNAN, H., and F. TAJIMA, 1997 The amount of nucleotide variation within and between allelic classes and the reconstruction of the common ancestral sequence in a population. *Genetics* **147**: 1431–1444.
- JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.
- KAWABE, A., and N. T. MIYASHITA, 1999 DNA variation in the basic chitinase locus (ChiB) region of the wild plant *Arabidopsis thaliana*. *Genetics* **153**: 1445–1453.
- KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. *Genetics* **61**: 893–903.
- KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**: 111–120.
- KUITTINEN, H., and M. AGUADE, 2000 Nucleotide variation at the CHALCONE ISOMERASE locus in *Arabidopsis thaliana*. *Genetics* **155**: 863–872.
- MIYASHIYA, N., 2001 DNA variation in the 5' upstream region of the Adh locus of the wild plants *Arabidopsis thaliana* and *Arabidopsis gemmifera*. *Mol. Biol. Evol.* **18**: 164–171.
- NACHMAN, M. W., and S. L. CROWELL, 2000 Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**: 297–304.
- OLSEN, K. M., A. WOMACK, A. R. GARRETT, J. I. SUDDITH and M. D. PURUGGANAN, 2002 Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. *Genetics* **160**: 1641–1650.
- PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. *Genetics* **160**: 1179–1189.
- ROGERS, A., 1992 Error introduced by the infinite-site model. *Mol. Biol. Evol.* **9**: 1181–1184.
- SAVOLAINEN, O., C. H. LANGLEY, B. P. LAZZARO and H. FREVILLE, 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Mol. Biol. Evol.* **17**: 645–655.
- TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.
- WALL, J. D., 1999 Recombination and the power of statistical tests of neutrality. *Genet. Res.* **74**: 65–79.
- WATTERSON, G. A., 1975 On the number of segregation sites. *Theor. Popul. Biol.* **7**: 256–276.
- YANG, Z., and A. D. YODER, 1999 Estimation of the transition/transversion rate bias and species sampling. *J. Mol. Evol.* **48**: 274–283.
- YANG, Z., S. KUMAR and M. NEI, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**: 1641–1650.

Communicating editor: D. BEGUN